Manoj Kumar Gupta
Lambodar Behera *Editors*

Bioinformatics in Rice Research

Theories and Techniques

Springer

Bioinformatics in Rice Research

Manoj Kumar Gupta • Lambodar Behera
Editors

Bioinformatics in Rice Research

Theories and Techniques

Springer

Editors
Manoj Kumar Gupta
Crop Improvement Division
ICAR-National Rice Research Institute
Cuttack, Odisha, India

Lambodar Behera
Crop Improvement Division
ICAR-National Rice Research Institute
Cuttack, Odisha, India

ISBN 978-981-16-3992-0 ISBN 978-981-16-3993-7 (eBook)
https://doi.org/10.1007/978-981-16-3993-7

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Dr. Manoj Kumar Gupta and Dr. Lambodar Behera dedicate this book to every member of Gupta, Behera, and Benzenberg families for their endless love, inspiration, and motivation.

Preface

In the genomics era, bioinformatics, a discipline that uses data from molecular sequence (nucleotide and protein) research to make observations regarding biological functions and entails the application of information technology to biology, has grown exponentially. It is now an integral part of biological research, with applications in a wide range of fields. Bioinformatics is expanding in tandem with rapid technical and scientific advancements in a variety of fields. The science of “big data” is the product of collaboration between ultrahigh-throughput technological growth and high-performance computers. In recent years, genetics, genomics, proteomics, metabolomics, and metagenomics have had a greater impact on biology. The Human Genome Project resulted in next-generation sequencing technologies, which allow for whole-exome and whole-genome sequencing (NGS) in as little as 24–56 h. This changed the way of unusual disorder genetic testing all over the world. Almost every nation has scientists who can analyse NGS data for diagnostic purposes. The world would profit enormously from the decoding of plant and animal genomes. Bioinformatics methods would be useful for looking and explaining the roles of genes that are useful to the agricultural population inside certain genomes. This basic genetic information would be useful for the development of crops that are more drought, disease, and insect tolerant, as well as increase livestock production by making them better, disease-resistant, and more profitable.

Bioinformatics in Rice Research: Theories and Techniques provides an up-to-date review of various classical and advanced bioinformatics and molecular biology approaches that are used in various biological fields. This book comprises 26 chapters divided into 3 parts. The first part describes in brief the importance of bioinformatics and statistics in biological research. This part also gives a brief idea about various rice varieties and biological databases that have been developed for a better understanding of the rice genome complexity. In the second part, various tools and techniques for bioinformatics analysis of genomic and proteomic sequences have been discussed in detail. In the third part, principles and techniques of various high-throughput technologies have been described. All chapters in this book also aim at scientists/researchers who are interested in developing bioinformatics tools. These books will be highly useful for both beginners and experienced bioinformaticians interested in solving biological questions.

Cuttack, Odisha, India

Manoj Kumar Gupta
Lambodar Behera

Acknowledgements

At first, we, the editors, would like to express our gratitude to God, Almighty, who gave us wisdom, health, knowledge, strength, and energy to succeed in our dream of writing a book. We are thankful to the Director and Head of the Division of Crop Improvement of ICAR-National Rice Research Institute (formerly CRRI), Cuttack, Odisha, India, for providing inspiration and motivation. We also express our sincere thanks to Dr. Ramakrishna Vadde, Associate Professor, Department of Biotechnology and Bioinformatics, Yogi Vemana University, for his inspiration, valuable suggestions, and guidance on some chapters. We also thank Mr. Jan Benzenberg, Incident Manager, Carrier Radio Access Network (Germany), for his valuable guidance, inspiration, encouragement, and indefatigable attitude that helped us throughout the journey of writing and preparing this manuscript.

This book would not have been possible without the authors and co-authors who supported in writing the chapters. We feel immense pleasure to express our gratitude to all authors and co-authors for writing and successfully completing it within time. We heartily appreciate the team members, especially, Dr. Gayatri Gouda, Dr. Ravindra Donde and Mr. S. Sabarinathan, for arranging the manuscripts.

We want to acknowledge Springer Nature (Publisher) for accepting our book proposal and showing interest in publishing it.

Contents

Part I Introduction

Part II Bioinformatics Analysis of Genomic and Proteomic Sequences

Part III High Throughput Technologies

About the Editors

Manoj Kumar Gupta, Ph.D. has completed his B.Sc. (Botany (Honours)) from Calcutta University, Kolkata, in the year 2007; MCA (distance course) from ICFAI University Tripura in the year 2010; M.Sc. (Biotechnology) from Berhampur University, Odisha, in the year 2011; and Ph.D. (Biotechnology and Bioinformatics), Yogi Vemana University, Andhra Pradesh, in the year 2019. During his Ph.D., he did an internship at Max Planck Institute of Evolutionary Biology, Germany. He is working as a Research Associate in the Crop Improvement Division under Dr. Lambodar Behera, Principal Scientist, ICAR-National Rice Research Institute, India. The central theme of his research is the application and development of computational tools and techniques to understand how specific biological functions emerge from the dynamics of biological networks. He has expertise in molecular biology, bioinformatics, structural biology, and evolutionary medicine. He has successfully guided/co-guided three postgraduate students for M.Sc. (Biotechnology and Bioinformatics) degrees. He has also published over 30 research articles in peer-reviewed international journals and authored or co-authored many book chapters.

Lambodar Behera, Ph.D. has completed his B.Sc. (Agriculture) from the Orissa University of Agriculture and Technology, Bhubaneswar, in the year 1989; M.Sc. (Biotechnology) from Tamil Nadu Agricultural University, Coimbatore, in the year 1992; and Ph.D. (Genetics) from Osmania University, Hyderabad, India, in the year 2000. He is presently working as Principal Scientist (Biotechnology) in ICAR-National Rice Research Institute, Cuttack, Odisha, India. He has significantly contributed to the field of rice molecular breeding, molecular biology, and genomics. He has significantly contributed to the development of 14 rice varieties suitable for different ecosystems of India. He has worked on the identification of QTLs/genes for resistance to BPH, gall midge, root-knot nematode and sheath blight disease, yield and related traits, tolerance to cold and drought; pyramiding of genes for BLB, submergence and drought resistance/tolerance, yield-related traits in the background of HYVs; assessment of the diversity of rice germplasm, insects, nematodes; molecular mechanism of low light tolerance; GWAS for identification of QTLs for drought tolerance, grain yield and related traits, and whole-genome sequencing of elite rice cultivars and donors, computational approach towards understanding the structural and functional role of different genes related to yield, and abiotic and biotic stress tolerance/resistance. He has developed a marker-assisted selection kit for gall midge resistance gene, *Gm4*, which is presently used in the incorporation of gall midge resistance gene, *Gm4*, into high yielding cultivars, Kavya, Lalat, Tapaswini, Swarna, and Samba Mahsuri. He has published 85 research articles in reputed national and international journals. He has written 11 book chapters, 2 popular articles, 11 technical bulletins, 2 practical manuals, and handled over 12 research projects financed by DBT, DST, ICAR, BIRAC, PPV and FRA of India, and BBSRC, UK. He has successfully guided/co-guided 8 research students for Ph.D. and 35 for M.Sc. (Biotechnology and Bioinformatics) degrees. He has evaluated thesis for Ph.D., M.Phil., and M.Sc. (Biotechnology and Bioinformatics) of OUAT, Bhubaneswar; Sambalpur, University, Odisha; BAU, Ranchi, Jharkhand; BCKV, Mohanpur, West Bengal; JNTU, Osmania; and PJTS Agricultural University, Hyderabad, India. Also, he has set questions and evaluated answer papers for B.Sc. (Ag), M.Sc. (Ag), M.Sc. (Biotechnology), Ph.D. students of OUAT, BBSR, Odisha, India, in the Biotechnology and Bioinformatics courses. He has received **Fellow of the ISGPG-2018** from the Indian Society of Genetics and Plant Breeding, New Delhi, **Rashtriya Gaurav Award-2019** from India International Friendship Society (IIFS), New Delhi, and **Leading Educationalist of India in 2020** by Friendship Forum, New Delhi.

Part I

Introduction

Introduction to Bioinformatics

1

Manoj Kumar Gupta and Lambodar Behera

Abstract

Recent advancements in technologies have led to the accumulation of a vast number of biological sequences and structures. However, storing and analyzing them are a grand challenge to biologists. Recently developed bioinformatic approaches serve as a key solution to solve these problems. Today, bioinformatics is used in various applications ranging from biofuel production to the drug discovery process. However, as most of the biological data and databases have distinct information and file formats, analyzing them requires knowledge from diverse fields, like molecular biology, mathematics, and computer sciences. Hence, the majority of the biologists end up using only basic bioinformatic tools. For solving complex biological problems, they continuously have to depend on expertise. Thus, there is an urgent requirement to teach biologists the necessary computer and mathematics skills during undergraduate and graduate studies. Even extensive initiatives also need to start to standardize the format of biological databases and ontology, which, in the near future, will help biologists to work independently and will make bioinformatics a more coherent discipline of biology.

Keywords

Bioinformatics · History · Genome · High-speed digital computers · Proteins · Sequences

M. K. Gupta · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_1

Abbreviations

GWAS	Genome-Wide Association Study
DNA	Deoxyribonucleic acid
RNA	Ribonucleic acid

1.1 Introduction

Recent advancements in sequencing technologies have unmasked the components and molecular mechanisms associated with various biological processes more comprehensively. Discovery of the structure and function of deoxyribonucleic acid (DNA) by Watson and Crick in 1953 was a notable milestone in molecular biology [1]. The researchers, who once believed that biological systems could be explained only by employing the physiochemical aspect, started analyzing their dataset using various molecular genetic techniques [2]. Thereafter, the most important discovery was the "central dogma of life," which describes how "DNA makes ribonucleic acid (RNA) and RNA make protein" [3]. This hypothesis gave us a few important belief, like genomic information is persistent throughout the organism's life and between cell types and individuals [4–6], modification within somatic cells are noninheritable [7], and important information for cellular function is contained within the gene sequence.

Though this information was essential to understand the simple biological process, they failed to explain biological systems' complexity. For instance, in 1978, Gilbert proposed the concept of "alternative splicing" [8] opposing the "one gene, one enzyme" hypothesis [9]. The "Alternative splicing" mechanism helps us to understand how ~20,000 human protein-coding genes encode more than 90,000 different proteins [8]. Constitutive splicing involves removing introns and ligation of exons in the order they appear in a gene. However, during alternative splicing, certain exons are skipped, which, in turn, generates numerous forms of mature mRNA [10]. Shorter exon length, weaker splicing signals nearby alternative splice sites, and higher sequence conservation nearby orthologs alternative exon determine which exon will be included in the matured mRNA [10]. This complete process is modulated via the spliceosome, which works in an antistatic manner and a synergistic manner [11, 12]. Until date, numerous studies have been performed to understand the function of alternative splicing across different biological systems, including testis, immune system, and brain [13], and reported that >95 percent of human genes experience tissue-specific, developmental, or signal-transduction-dependent splicing [14].

Another major problem associated with molecular biology data is incompleteness, noise, and redundancy. As of October 2020, the "Nucleic Acids Research" (NAR) online database reports the existence of ~1700 biological databases (http://www.oxfordjournals.org/nar/database/a/). The few most popular biological databases are UCSC Genome Browser (https://genome.ucsc.edu/), NCBI Entrez (https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html), EBI Ensemble (https://www.ensembl.org/index.html), and KEGG (https://www.genome.jp/kegg/). Though the initiative has been taken for data standardization, for instance, the Gene Ontology Consortium (http://geneontology.org/), most of these databases have their own data format. In order to use them effectively, users have to learn about the structure of each database distinctly [15]. Nevertheless, since the 1990s, with the advancement of genome sequencing technology, information from statistics, mathematic, physics, and chemistry also started getting incorporated into biology data. Thus, no single individual could analyze any dataset alone. Additionally, due to the continuous generation of biological data, storing them has also become a major problem. Recently developed bioinformatic approaches that involve information from diverse fields, like computer, mathematics, and molecular biology, serve as a key solution to solve these problems. Today, bioinformatics is used from biofuel production to drug discovery process. In this chapter, the authors attempted to understand what makes bioinformatics a key driving force in biological research that it is today.

1.2 Bioinformatics—Terminology

For the first time, in 1978, the term "bioinformatics" was coined by Ben Hesper and Paulien Hogewen and defined as "the study of informatics processes in biotic systems" [16, 17]. Initially, it was mainly used in studying genetics and genomics, specifically for analyzing large-scale DNA sequencing data. However, due to the rapid development of technologies, the amount of biological information also started accumulating. Hence, researchers started integrating information and theories from physics, computer science, mathematics, and biology for managing and analyzing heterogeneous biological data [16, 17]; what distinct bioinformatics from other approaches is its target to develop and apply computational techniques to achieve its aims [18].

1.3 History of Bioinformatics

The computer started serving as an important tool much before the development of DNA sequencing. However, the pioneers of bioinformatics did not use the term "bioinformatics" for describing their work. Nevertheless, they had a clear vision of combining information from various disciplines, like computer science, molecular biology, and mathematics, which may help us answer fundamental questions in the life sciences more significantly [19]. Three important factors that led to the emergence of bioinformatics in the early 1960s were (a) generation of protein sequence, (b) hypothesis that macromolecules are a carrier of information, and (c) production of high-speed digital computers, which were initially developed for weapons research programs at the time of World War II and eventually became accessible to academic biologists [19].

1.3.1 Generation of Protein Sequence

For the first time, during World War II, Emil Smith proposed that proteins' information is stored in linear sequences of amino acids [20]. Subsequently, the first protein sequence of bovine's insulin was published in the early 1950s by the English biochemist Frederick Sanger and his colleague [21, 22]. On the contrary, to earlier belief that proteins were somewhat amorphous [19], Sanger firmly established the protein structure polypeptide theory. For the same, he also got the Nobel Prize in 1958 [19]. This was a breakthrough in the world of protein biochemistry because when proposed for the first time in 1902, this theory was opposed by various alternative theories [19]. This finding serves as an encouragement to other researchers for developing approaches that can generate protein sequences more significantly. The Edman degradation method [23] emerged as one of the best methods that synthesize one amino acid at a time starting from the N-terminus. In combination with automation, this method helped us to sequence more than 15 different protein families during the next 10 years. However, Edman sequencing's main problems were its incapability to synthesize larger protein sequences and incomplete yield. Theoretically, the protein sequence from a single Edman reaction will be composed of ~50–60 amino acids. Thus, larger sequences were generally cleaved into smaller fragments and sequenced separately [24]. Another problem that arose with the increase in sequences number was comparing several sequences manually. To overcome this, Margaret Oakley Dayhoff compiled one of the first biological sequence databases, namely "*Atlas of Protein Sequence and Structure." She also* established approaches of sequence alignment and molecular evolution [25, 26]. The first version of the database comprised 65 interspecific protein sequences and served as an ideal dataset for researchers who proposed that protein sequences can be used to trace the evolutionary history of any species [24].

1.3.2 Macromolecules as a Carrier of Information

After the well establishment of the polypeptide theory and availability of approaches for sequencing proteins, the hypothesis that "proteins are information-carrying macromolecules" also became widespread. This general theory emerged in three widely related areas: genetic code, a protein's three-dimensional structure relative to its function, and protein evolution. The genetic code is not the nucleotide sequence of the genome. Rather, it is a group of laws that describes how any gene encodes proteins and how nucleotides are translated into amino acids. Interestingly, most of the rules are applicable to almost all organisms on earth, hence once called the "universal genetic code" [27]. Nevertheless, major challenges associated with them are to understand the origin of each code and how protein encoded via different genes modulates the function of the various organisms or organs or organelles [28]. In the late 1950s, Christian Anfinsen and the team reported that after denaturing, ribonuclease instinctively gets refolded and retains its original enzymatic activity [29]. This served as evidence that amino acid sequences determine the three-dimensional structure of any protein. However, only sequence information is not sufficient for predicting the secondary and tertiary structures of any protein. Combining biochemical techniques (e.g., Edman sequencing) along with biophysical techniques (e.g., X-ray crystallography) together may explain how molecular information in amino acids allows a protein to assemble into a particular, sometimes highly complicated, three-dimensional structure [29–31], which, in turn, modulates the function of the various organisms or organs or organelles [28].

Before 1960, most of the research was conducted only to understand the working mechanism of enzymes, antibodies, hormones, and respiratory pigments. Little attention was given to understand how the information within these macromolecules is preserved throughout evolution. During the 1960s, molecular biologists and biochemists started unmasking these questions [19]. Zuckerkandl and Pauling (1965) denoted nucleic acids and proteins as “semantides,” whose subunit sequences may be used for tracking evolutionary past. They hypothesized that “paleogenetic” study that combines molecular biology and biochemistry techniques might help us to answer the evolutionary question more precisely [32]. Subsequently, using paleogenetic approaches, researchers identified that myoglobin and hemoglobin originated through gene duplication. A comparison of homologous protein sequences was also conducted to trace phylogenetic relationships among protein themselves and the species that carried them [19].

Later, researchers have also developed the “Molecular Clock” hypothesis to predict evolutionary occurrences. However, several researchers opposed it, and several conflicts arose between traditional naturalists and molecular evolutionists [33–35]. The sequence analysis also had to contend with very well-established molecular biology techniques, including the immunological estimation pioneered by Morris Goodman and the team for unraveling phylogenetic relations [34]. The subsequent development of more advanced sequencing techniques and the incorporation of information from various other fields, like mathematics, molecular biology, and computer science, started filling the gap between traditional naturalists and molecular evolutionists slowly, which gradually laid the evolutionary basis for today’s bioinformatics [35–37].

1.3.3 The Emergence of High-Speed Digital Computers

Initially, scientists emphasized the importance of instrumentation in protein biochemistry research [38, 39]. However, when John Kendrew employed computers to determine the three-dimensional hemoglobin structure, researchers started understanding the importance of computers in protein biochemistry [40, 41]. However, during World War II, computers were mostly used for military purposes. Computers of the “second generation” became widely accessible for academic uses in the early

1960s. In 1957, the International Business Machine (IBM) Company developed FORTRAN (formula translation), the first high-quality programming language. FORTRAN was well-suited for scientific applications and was easy to use in comparison with the earlier machine languages. For the first, it was possible to write a program without knowing the computer architecture. This fortified the growth of bioinformatics [19]. These computational advancements helped researchers like Margaret Oakley Dayhoff and Stein and Moore in sequence and phylogenetic analysis of various proteins like cytochrome c, in a shorter duration [42–44]. Using the computational approaches, biophysicist Cyrus Levinthal and his team also constructed three-dimensional models of cytochrome *c*. For the first time, they could project the molecules on an oscilloscope screen and control the model's turning via a hand-operated device. They could even manipulate the molecules employing either a light pen or a keyboard [45]. In due course of time, all this information and approaches encourage the development of various more powerful computational tools and techniques used during bioinformatic analysis today.

1.4 Application of Bioinformatics

Researchers are continuously employing bioinformatic tools and techniques to unmask various biological questions ranging from health care to agriculture. The majority of bioinformatic approaches include genome assembly, sequence alignment, gene recognition, drug design, genome-wide correlation studies, prediction of the protein structure, protein–protein interaction, prognosis of genetic expression, protein structure alignment, and evolutionary modeling [18].

1.4.1 Genome Sequence

The analysis of continuously growing genomic sequences and the human genome project is a milestone accomplishment for bioinformatics [46]. In 1995, the whole genome of *Haemophilus influenzae* was sequenced using the "shot-gun" technique. This was the first whole genome of any free-living organism sequenced [47]. Subsequently, the whole-genome sequence of *Yersinia pestis* [48], *Mycobacterium tuberculosis* [49], and *Mycoplasma genitalium* [50] was done. The complete genome sequence of the first eukaryotic organism was from *Saccharomyces cerevisiae* followed by other eukaryotic species such as *Arabidopsis thaliana* [51], *Drosophila melanogaster* [52], and *Caenorhabditis elegans* [53]. Genome sequences help map gene location in various organisms, including humans, and enable us to unmask the molecular function associated with each gene [54, 55].

1.4.2 Gene Expression and Variation

Bioinformatics also helps us detect the pattern of gene expression and variation associated with any disease or trait [46, 56]. For instance, the Cancer Genome Atlas project was launched in order to promote genetic cancer research by systematic sequencing approaches by the National Cancer Institute and the National Human Genome Discovery Institute. The science world has evidence accessible. Detailed ovarian cancer evidence was published suggesting that *BRCA1, TP53,* and *BRCA2* mutation was strongly correlated with ovarian cancer [57]. The 1000 Genome Project has begun to raise awareness of human genome variations (SNPs, haplotypes, and structural variants) through population-scale sequencing by international cooperation. The project recorded its pilot phase in 2010, and new data were published every month [6]. The 1000 Genome Project's main objective is to identify, genotype, and provide correct haplotype details on all aspects of the polymorphism in DNA of various human groups. In particular, the aim is to characterize more than 95% of the variants in each of the main population groups (populations in or from ancestries from Southern Africa, Europe, western Africa, East Asia, and the Americas) [6]. The Human Intestinal Tract project aims at connecting human health with the intestinal microbiota. The partial gut metagenomes of 124 people in Europe were identified in major research. Results revealed that the bowel environment can be utilized for diagnosing individuals' well-being [58].

In another study, the author employed bioinformatic approaches to detect ischemic stroke-associated 10 key genes, namely *IL1α, ICAM1, IL1β, IL6, CCL4, CXCL1, IL8, CXCL2, CXCL20,* and *PTGS2*. Functional enrichment analysis reveals that these genes were mainly involved in biological immune response and apoptotic processes, including *NOD* and *TNF*-like receptor signaling pathways [59]. In 2019, Guo and the team identified a total of 782 differentially expressed genes (DEGs), including 392 upregulated and 390 downregulated DEGs. Hierarchical clustering shows that the DEGs are able to separate intracranial aneurysm specifically from the superficial temporal artery [60]. The GO enrichment analysis reveals that upregulated DEGs are mainly associated with inflammatory reaction and extracellular matrix regulation. The downregulated DEGs mainly engage in the mechanism of vascular fluid musculoskeletal contraction. These genes are primarily associated with "Leishmaniasis," "Toll-like receptor pathway," and "vascular smooth muscle contraction" [60]. Earlier, we have also employed bioinformatic approaches and detected four key genes, namely *CCL2*, *ELMO1, TCF7L2*, and *VEGFA*, along with *FOS,* which plays a key role in causing type 2 diabetes and related diseases (such as neuropathy, rheumatoid arthritis, and nephropathy) and cancer through p53 or Wnt signaling pathways [61]. In another study, we identified that during Japanese encephalitis virus infection, the *STAT1* gene gets downregulated. The gene *STAT1* was shown to interact with the family members of tyrosine–protein kinase and had good interaction with the genes *JAK1* and *JAK2* [62].

1.4.3 Structure and Function of the Protein

One of the major problems that biology and protein engineering researchers experience is the development of manufacturing processes, where the tertiary structure of amino acid proteins is determined for building new protein and new medicinal products. Many of the protein structures identified to date are achieved by NMR spectroscopy experiments, X-ray crystallography, and cryo-EM, but these methods are more time-consuming and cost-ineffective [63]. Modeling by bioinformatic programs has succeeded in predicting many proteins' atomic structure from their amino acid sequence's relative. Additionally, these processes are quicker and more economical and provide better resolution result [64]. Earlier, several bioinformatic studies have been conducted for determining the structure and function of several protein [65, 66] and gene regulation networks [61, 62], which in turn play a key role during the drug discovery process.

In 1964, Feynman presciently stated, "Certainly no subject or field is making more progress on so many fronts at the present moment than biology, and if we were to name the most powerful assumption of all, which leads one on and on in an attempt to understand life, it is that all things are made of atoms and that everything that living things do can be understood in terms of the jigglings and wigglings of atoms" [67]. Molecular dynamics (MD) is an important computational method in understanding the structure's physical foundation, its dynamic growth, and its operation. A first MD simulation for BPTI (bovine pancreatic trypsin inhibitor) was released after 15 years later of its original structure identification. While BPTI had a reasonably clear X-ray structure at that time, its physiological role remained unclear [68]. One of the most valuable MD simulation tools is the "Structural Bioinformatics Research Collaboratory" (RCSB, www.rcsb.org) that allows experimentally specified biological macromolecular structural data accessible. The RCSB Protein Data Bank (PDB) also serves a global archive for macromolecular 3D structure data processing and sharing and a vital platform for biomolecular modeling [69].

With the increase in the structure number in the PDB database, structure of large amount of target protein can be predicted by homology modeling [70]. However, if no structure with clear target protein sequence similarity in PDB is detected, proteins with similar structures can still be associated with the target protein. The technique of recognizing template structures from the PDB is known as folding identification or threading. This is based on an algorithm that fits the target sequence, and the structure is distant homologous. The underlying assumption for threading is that the protein structure grows extremely conservatively and that the amount of specific structural folds is limited in nature. Template-dependent structure methods may be called threading methods (on the basis of fold recognition) and homology (on the basis sequence comparison). In contrast to threading methods and homology modeling, the ab initio approach attempts to construct a structure on the basis of the physics' first principles that do not depend on systems already resolved. The ab initio approach is often established by discovering the second genetic code.

However, effective ab initio approaches are very unusual, and many difficulties and obstacles continue to be solved [70].

To date, several bioinformatic tools and techniques have been developed for characterizing the structure and function of the protein [71]. For instance, the modeler generates a protein model by matching the target sequences to the template; it detects nonhydrogen atoms for creating a model and also employed for loop simulation and protein optimization [72]. “Protein Interactive Modeling” (PRIMO) is a protein monomer homology system. It provides functionality that aids users for modeling ligands and ions within a protein target complex [73]. I-TASSER is a hierarchical technique that models protein structure based on secondary structure-enhanced thread profile alignment and iterative threading assembly optimizing software execution [74]. SWISS-MODEL workspace can build and validate protein models [75]. PROCHECK is a designed protein validation software that produces a Ramachandran plot and examines atomic distances, torsion angles, bond length, and surface area [76]. ERRAT (https://servicesn.mbi.ucla.edu/ERRAT/) is a protein structure validation technique that evaluates the structure and refinement of the crystallographic model. Thus, advances in biocomputing have developed more accurate and simpler methods for modeling proteins, which in turn reduces analytical time and expense. However, research is still required to ensure that the hypothesis is accurate, in addition to enhancing the techniques’ reliability [77].

1.4.4 Evolution of Gene and Protein

Sequencing the genome of the various organisms has helped explain the array of all its genes, elucidate the functions and properties they convey, and how they regulate pathways in different metabolic processes through countless protein–protein interaction. For determining their probable functions, genetic and biochemical analyses are performed via classical approaches. Nevertheless, sequencing has helped us understand the amino acids’ composition and homology analysis between amino acids may help us to infer the functions of these proteins [78]. Bioinformatics is a challenge for protein analysis because, in recent years, phylogenetic profiling has become important to compare homologous proteins by aligning their sequences, in which many share extremely conserved domains and associated structures [79]. Phylogeny analyses the variations throughout the genomes and categorizes them in the consequence diagram termed as the phylogenetic tree. All sequences corresponding to the same family can be categorized into clade and subfamily, and provide descriptions of their origin including functional diversity [80].

Few studies have also reported that during their development, eukaryotic cells acquired microbes that formed chloroplasts, mitochondria, and other organelles, whose genomes were transported to the nuclear genome, which in turn facilitated the transmission of encoded proteins throughout the nucleus. These studies have also suggested that several proteins in eukaryotes are closely associated with prokaryotic proteins [81]. For instance, the amino acid composition, sequence, length, and conserved regions of mitochondrial and chloroplast’s proteins are very similar to those of prokaryotes. One of the drawbacks in the study of proteins among different organisms is that genomes must be complete in order to establish whether or not certain species have similar genes [82].

When any species adjust to specific environmental circumstances, they undergo mutational shifts in genome sequences, inducing amino acid substitutes in enzymes, increasing their performance and precision, and retaining their catalytic role. Not all genes that code for proteins are vulnerable to mutation due to the involvement of important amino acids in structure, stabilization, and folding, causing a limitation. Most of the mutations are typically spontaneous and attributable to environmental strain in certain proteins in which these modifications were found. If the protein plays a significant role in the organism's activities and the mutation enhances their performance, the genetic alteration is preserved and refined via positive selection, i.e., preferred by the selective selection. Otherwise, the protein function is not appropriate, and mutation is removed via negative selection [83–85].

The analysis of ancestral enzymes indicated that these posed a high thermostability owing to the thermophilic Precambrian age. The alignment of ancestral protein with contemporary protein indicates slow structural growth, although not in amino acids [86]. Enzymes are thus the result of decades of development, through which improvements have been produced to attain a certain function and a stronger affinity with the substrate and/or function on multisubstrate. Genetic diversity has indeed created homologous genes (e.g., orthologs that have originated from a common ancestor), which encode adapted proteins for performing their catalysis under extreme environment. However, paralog genes encode proteins with various tasks [87]. Basic features, e.g., attachment of a receptor or response mechanisms, are sometimes retained, but they carry out specialized functions on a different substrate. On the contrary, orthological proteins have a similar function and their sequences are also strongly conserved [88]. Bioinformatics can also be employed in tracing the evolutionary history and ages of any genes or species [83, 84, 89]. Additionally, it can also be used in tracing the difference between the numbers, locations, and biochemical functions of various genes present within different organisms [26, 90, 91].

1.4.5 Agriculture

For decades, plant breeders are also employing bioinformatic approaches to detect the resistance and yield-related polymorphisms in various crop plants, like tomato [92, 93], rice [94, 95], maize [96, 97], and soybean [98]. Recently, our laboratory also employed both wet-lab and bioinformatic approaches to detect three high grain specific single nucleotide polymorphisms in rice plants [66]. In another study, we employed computational approaches to understand the impact of biotic [99] and abiotic stress in rice plants [100]. Bioinformatics has also been proved useful in deciphering the complex interaction among various organisms present within the soil [101] and plant rhizosphere [102]. It also enabled us to understand how the microbial composition within the rhizosphere, as well as in soil, changes in response to different environmental conditions [103–105]. Bioinformatic approaches also help us understand how various microorganisms modulate plant nutrition [106, 107] or elements [108] in soil. It can also be used to detect novel genes, develop bioproducts (Esposito et al. 2016), and modulate biodefense countermeasures [109–111]. The utilization of long-read sequence and long-ranging mapping techniques and conformation capture of chromosomes have helped us generate the various contiguous crop genome assembly, including nonmodel crop [112].

Because of the recent development of third-generation sequencing technologies, vast amounts of data are also accessible to researchers regarding gene-to-population crop traits [113]. While important sequence collection resources, like European Molecular Biological Laboratory (EMBL) [114], GenBank [115], PlantGDB [116], and Phytozome [117], are important, their primary goal is to manage and store genomic data without combining variant or phenotype information from different sources. This information is focused primarily on genomics. This makes it more difficult for plant biologists and breeders to relate genotype to phenotype and takes details regarding genomics, epigenomics, phenotypes, and conditions. Though crop databases with this information are accessible, e.g., marker and expressions are embedded into GrainGenes [118], further more databases with are still needed for resolving this void [119]. Bioinformatics also play an important role in designing optimal RNA-guided structure for effective and precise CRISPR/Cas gene editing. While genome editing provides a useful approach for rapidly integrating advantageous mutations into elite cultivars, genomic selection improvises selection efficiency without needing awareness of underlying genetic drivers. Machine learning algorithms may also employ high-throughput phenotyping and genomic data to simplify aspects of the gene discovery process like genome annotation [120].

Improvements in bioinformatic techniques have provided an extra possibility for genome-wide association search in plant and agricultural research. For instance, PLINK is a commonly used bioinformatic approach for genome-wide association studies. It employs traditional regression analysis for correlating genotypes with phenotypes [121]. Nevertheless, standard regression cannot have reasonable specificity for GWAS for rare variants [122]. TASSEL is another popular method for GWAS that utilizes a mixed linear model and involves population and family structure. Additionally, TASSEL can also control population results, unlike PLINK [122]. Other bioinformatic tools, like GAPIT, effectively manage a broad data collection of more than one million SNPs employing compressed linear models and model-based prediction and selection methods in 10,000 individuals.

1.5 Conclusion and Future Perspective

In conclusion, bioinformatics is continuously helping researchers to decipher biological questions, and most of the analysis involves knowledge from diverse fields, like molecular biology, mathematics, chemistry, biochemistry, and computer science. Hence, most biologists end up using only basic bioinformatic tools

[123]. For solving complex problems, they continuously depend on expertise. Thus, there is an urgent requirement to teach basic computer skills along with molecular biology techniques during undergraduate and graduate studies. This will help researchers to bridge the gap between theoretical biology and experimental biology. Additionally, the result obtained from molecular biology approaches will help bioinformaticians to support their finding more strongly [123].

Another major problem with the bioinformatic analysis is the heterogeneity of how data are annotated, displayed, and analyzed, and absence of compatibility among usable data [123]. This problem is mainly because of the young age of bioinformatics and lack of specific format in conventions and discipline related to an established scientific community. To overcome this problem, recently, researchers have set up a scholarly association for information curators (www.biocurator.org) and projects that connect various databases of model organisms (www.gmod.org). Even the initiation has also started to maintain a standard gene ontology format (http://geneontology.org/). These efforts will make biologists more independent researchers and also bioinformatics a more coherent discipline of biology.

Considering the benefits and lacuna of the bioinformatics, in the present book, we tried to understand the importance of bioinformatics in rice research. Efforts have been made to include chapter ranging from topic of basic bioinformatics, like sequence alignment to protein modeling to its application in rice research, like the importance of metabolomics and computational epigenetics in rice research. We have also included few molecular biology topics, like high-throughput sequencing and genome-wide association study, that may be useful for the beginner and experienced researcher. Toward the end, advanced topic, like intellectual property right, has also been discussed. Thus, the content of the book has been designed in such a way that almost every individual working in the bioinformatic-related field will be benefitted.

Conflicts of Interest None.

References

1. Moore P. Francis Crick dies. Genome Biol. 2004 Jul 30;5(1) spotlight-20040730-01.
2. Ayyildiz D, Piazza S. Introduction to Bioinformatics. In: Bolón-Canedo V, Alonso-Betanzos-A, editors. Microarray Bioinformatics [Internet]. New York: Springer; 2019. p. 1–15. [cited 2020 Aug 25]. (Methods in Molecular Biology). Available from:. https://doi.org/10.1007/978-1-4939-9442-7_1.
3. Crick F. Central dogma of molecular biology. Nature. 1970 Aug;227(5258):561–3.
4. Roach JC, Glusman G, Smit AFA, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010 Apr 30;328 (5978):636–9.
5. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, Smith JP, et al. The characterization of twenty sequenced human genomes. PLoS Genet. 2010 Sep 9;6(9):e1001111.

6. Durbin RM, Altshuler D, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, et al. A map of human genome variation from population-scale sequencing. Nature. 2010 Oct;467 (7319):1061–73.
7. McDaniell R, Lee B-K, Song L, Liu Z, Boyle AP, Erdos MR, et al. Heritable individual-specific and allele-specific chromatin signatures in humans. Science. 2010 Apr 9;328 (5975):235–9.
8. Wang Y, Liu J, Huang B, Xu Y-M, Li J, Huang L-F, et al. Mechanism of alternative splicing and its regulation. Biomed Rep. 2015 Mar;3(2):152–8.
9. Beadle GW, Tatum EL. Genetic control of biochemical reactions in Neurospora. Proc Natl Acad Sci U S A. 1941 Nov 15;27(11):499–506.
10. Zheng CL, Fu X-D, Gribskov M. Characteristics and regulatory elements defining constitutive splicing and different modes of alternative splicing in human and mouse. RNA. 2005 Dec;11 (12):1777–87.
11. Effenberger KA, Perriman RJ, Bray WM, Lokey RS, Ares M, Jurica MS. A high-throughput splicing assay identifies new classes of inhibitors of human and yeast spliceosomes. J Biomol Screen. 2013 Oct;18(9):1110–20.
12. Wahl MC, Will CL, Lührmann R. The spliceosome: design principles of a dynamic RNP machine. Cell. 2009 Feb 20;136(4):701–18.
13. Sebestyén E, Zawisza M, Eyras E. Detection of recurrent alternative splicing switches in tumor samples reveals novel signatures of cancer. Nucleic Acids Res. 2015 Feb 18;43(3):1345–56.
14. Nilsen TW, Graveley BR. Expansion of the eukaryotic proteome by alternative splicing. Nature. 2010 Jan 28;463(7280):457–63.
15. Can T. Introduction to Bioinformatics. In: Yousef M, Allmer J, editors. miRNomics: MicroRNA Biology and Computational Analysis [Internet]. Totowa: Humana Press; 2014. p. 51–71. [cited 2020 Aug 31]. (Methods in Molecular Biology). https://doi.org/10.1007/978-1-62703-748-8_4.
16. Hogeweg P. The roots of bioinformatics in theoretical biology. PLoS Comput Biol. 2011 Mar 31;7(3):e1002021.
17. Hogeweg P. Simulating the growth of cellular forms. Simulation [Internet]. 2016 Aug. 18 [cited 2020 Sep 4]; https://doi.org/10.1177/003754977803100305.
18. Rao VS, Das SK, Rao VJ, Srinubabu G. Recent developments in life sciences research: Role of bioinformatics. Afr J Biotechnol [Internet]. 2008;7(5) [cited 2020 Sep 4]. Available from: https://www.ajol.info/index.php/ajb/article/view/58463.
19. Hagen JB. The origins of bioinformatics. Nat Rev Genet. 2000 Dec;1(3):231–6.
20. Srinivasan PR, Fruton JS, Edsall JR, editors. The origins of modern biochemistry: a retrospect on proteins. New York: New York Academy of Sciences; 1993. 375 p.
21. Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. I. the identification of lower peptides from partial hydrolysates. Biochem J. 1953 Feb;53(3):353–66.
22. Sanger F, Thompson EOP. The amino-acid sequence in the glycyl chain of insulin. 2. The investigation of peptides from enzymic hydrolysates. Biochem J. 1953 Feb 1;53(3):366–74.
23. Edman P. A method for the determination of amino acid sequence in peptides. Arch Biochem. 1949 Jul;22(3):475.
24. Gauthier J, Vincent AT, Charette SJ, Derome N. A brief history of bioinformatics. Brief Bioinform. 2019 Nov 27;20(6):1981–96.
25. Eck RV, Dayhoff MO. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science. 1966 Apr 15;152(3720):363–6.
26. Moody G. Digital code of life: how bioinformatics is revolutionizing science, medicine, and business. Hoboken: John Wiley & Sons; 2004. 408 p.
27. Hinegardner RT, Engelberg J. Rationale for a universal genetic code. Science. 1963 Nov 22;142(3595):1083–5.
28. Keeling PJ. Genomics: evolution of the genetic code. Curr Biol. 2016 Sep 26;26(18):R851–3.
29. Anfinsen CB. Principles that govern the folding of protein chains. Science. 1973;181 (4096):223–30.

30. Stein WH, Moore S. The chemical structure of proteins. Sci Am. 1961;204(2):81–95.
31. Moore S, Stein WH. Chemical structures of pancreatic ribonuclease and Deoxyribonuclease. Science. 1973;180(4085):458–64.
32. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol. 1965 Mar 1;8(2):357–66.
33. Zuckerkandl E. On the molecular evolutionary clock. J Mol Evol. 1987 Nov 1;26(1):34–46.
34. Dietrich MR. Paradox and persuasion: negotiating the place of molecular evolution within evolutionary biology. J Hist Biol. 1998;31(1):85–111.
35. Hagen JB. Naturalists, molecular biologists, and the challenges of molecular evolution. J Hist Biol. 1999 Sep 1;32(2):321–41.
36. Jungck JR, Friedman RM. Mathematical tools for molecular genetics data: an annotated bibliography. Bltn Math Biol. 1984 Jul 1;46(4):699–744.
37. Perutz M. [1]Early days of protein crystallography. In: Methods in Enzymology [Internet]. Academic Press; 1985. p. 3–18. [cited 2020 Sep 9]. (Diffraction Methods for Biological Macromolecules Part A; vol. 114). Available from: http://www.sciencedirect.com/science/article/pii/0076687985140036
38. Kay LE. The molecular vision of life: Caltech, the Rockefeller Foundation, and the rise of the new biology. Oxford University Press; 1992. p. 318.
39. Brock WH. Proteins, enzymes, genes: the interplay of chemistry and biology. Med Hist. 2000 Jul;44(3):409–10.
40. Kendrew JC, Bodo G, Dintzis HM, Parrish RG, Wyckoff H, Phillips DC. A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. Nature. 1958 Mar 8;181 (4610):662–6.
41. Kendrew JC, Dickerson RE, Strandberg BE, Hart RG, Davies DR, Phillips DC, et al. Structure of myoglobin: a three-dimensional Fourier synthesis at 2 a. resolution. Nature. 1960 Feb 13;185(4711):422–7.
42. Fitch WM. An improved method of testing for evolutionary homology. J Mol Biol. 1966 Mar 1;16(1):9–16.
43. Hersh RT. Atlas of protein sequence and structure, 1966. Syst Biol. 1967 Sep 1;16(3):262–3.
44. Dayhoff MO. Computer analysis of protein evolution. Sci Am. 1969;221(1):86–95.
45. Levinthal C. Molecular model-building by computer. Sci Am. 1966;214(6):42–53.
46. Bayat A. Bioinformatics BMJ. 2002 Apr 27;324(7344):1018–22.
47. Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science. 1995 Jul 28;269(5223):496–512.
48. Parkhill J, Wren BW, Thomson NR, Titball RW, Holden MT, Prentice MB, et al. Genome sequence of Yersinia pestis, the causative agent of plague. Nature. 2001 Oct 4;413 (6855):523–7.
49. Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, et al. Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature. 1998 Jun 11;393 (6685):537–44.
50. Fraser CM, Gocayne JD, White O, Adams MD, Clayton RA, Fleischmann RD, et al. The minimal gene complement of mycoplasma genitalium. Science. 1995 Oct 20;270 (5235):397–403.
51. Initiative AG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature. 2000 Dec 14;408(6814):796–815.
52. Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, et al. The genome sequence of Drosophila melanogaster. Science. 2000 Mar 24;287(5461):2185–95.
53. C. elegans Sequencing Consortium. Genome sequence of the nematode C. elegans: a platform for investigating biology. Science. 1998 Dec 11;282(5396):2012–8.
54. Stein L. Genome annotation: from sequence to biology. Nat Rev Genet. 2001 Jul;2 (7):493–503.

55. Subramanian G, Adams MD, Venter JC, Broder S. Implications of the human genome for understanding human biology and medicine. JAMA. 2001 Nov 14;286(18):2296–307.
56. Gupta MK, Donde R, Gouda G, Vadde R, Behera L. De novo assembly and characterization of transcriptome towards understanding molecular mechanism associated with MYMIV-resistance in Vigna mungo - A computational study. bioRxiv. 2019 Nov;16:844639.
57. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. Nature. 2011 Jun 29;474(7353):609–15.
58. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, et al. A human gut microbial gene catalogue established by metagenomic sequencing. Nature. 2010 Mar 4;464 (7285):59–65.
59. He Y, Liu J, Zhao Z, Zhao H. Bioinformatics analysis of gene expression profiles of esophageal squamous cell carcinoma. Dis Esophagus. 2017 May 1;30(5):1–8.
60. Guo T, Hou D, Yu D. Bioinformatics analysis of gene expression profile data to screen key genes involved in intracranial aneurysms. Mol Med Rep. 2019 Nov 1;20(5):4415–24.
61. Gupta MK, Vadde R. Identification and Characterization of Differentially Expressed Genes in Type 2 Diabetes using in silico Approach. Comput Biol Chem [Internet]. 2019 Jan 24; [cited 2019 Jan 27]; Available from: http://www.sciencedirect.com/science/article/pii/S1476927118302238.
62. Gupta MK, Behera SK, Dehury B, Mahapatra N. Identification and characterization of differentially expressed genes from human microglial cell samples infected with Japanese encephalitis virus. J Vector Borne Dis. 2017 Jun;54(2):131–8.
63. Cheung NJ, Yu W. De novo protein structure prediction using ultra-fast molecular dynamics simulation. PLoS One. 2018;13(11):e0205819.
64. Bonneau R, Baker D. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct. 2001;30:173–89.
65. Gupta MK, Vadde R. A computational structural biology study to understand the impact of mutation on structure–function relationship of inward-rectifier potassium ion channel Kir6.2 in human. J Biomol Struct Dyn. 2020 Feb 23;39(4):1–14.
66. Gouda G, Gupta MK, Donde R, Kumar J, Parida M, Mohapatra T, et al. Characterization of haplotypes and single nucleotide polymorphisms associated with Gn1a for high grain number formation in rice plant. Genomics. 2020 May 1;112(3):2647–57.
67. Feynman RP, Leighton RB, Sands M. The Feynman Lectures on Physics. In: The New Millennium Edition: Mainly Mechanics, Radiation, and Heat, vol. I. New York: Basic Books; 2011. 562 p.
68. McCammon JA, Gelin BR, Karplus M. Dynamics of folded proteins. Nature. 1977 Jun 16;267 (5612):585–90.
69. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res. 2000 Jan 1;28(1):235–42.
70. Al-Lazikani B, Jung J, Xiang Z, Honig B. Protein structure prediction. Curr Opin Chem Biol. 2001 Feb 1;5(1):51–6.
71. Skariyachan S, Garka S. Chapter 1 - Exploring the binding potential of carbon nanotubes and fullerene towards major drug targets of multidrug resistant bacterial pathogens and their utility as novel therapeutic agents. In: Grumezescu AM, editor. Fullerens, Graphenes and Nanotubes [Internet]. William Andrew Publishing; 2018. p. 1–29. [cited 2020 Oct 4] Available from: http://www.sciencedirect.com/science/article/pii/B9780128136911000014.
72. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Meth Enzymol. 2003;374:461–91.
73. Hatherley R, Brown DK, Glenister M, Tastan Bishop Ö. PRIMO: An Interactive Homology Modeling Pipeline. PLoS One [Internet]. 2016 Nov 17;11(11) [cited 2020 Oct 4]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5113968/.
74. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. 2015 Jul 1;43((Web Server issue)):W174–81.

75. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 2003 Jul 1;31(13):3381–5.
76. Laskowski RA, MacArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr. 1993;26(2):283–91.
77. Xu D, Xu Y, Uberbacher EC. Computational tools for protein modeling. Curr Protein Pept Sci. 2000 Jul;1(1):1–21.
78. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D. Detecting protein function and protein-protein interactions from genome sequences. Science. 1999 Jul 30;285 (5428):751–3.
79. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. PNAS. 1999 Apr 13;96(8):4285–8.
80. Vries R, de Tsang A, Grigoriev IV, editors. Fungal genomics: methods and protocols [Internet], Methods in molecular biology. 2nd ed. Totowa: Humana Press; 2018. [cited 2020 Oct 4]; Available from: https://www.springer.com/gp/book/9781493978038.
81. Margulis L. Origin of eukaryotic cells: evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth. New Haven: Yale University Press; 1970. 349 p.
82. Valencia A, Pazos F. Computational methods for the prediction of protein interactions. Curr Opin Struct Biol. 2002 Jun;12(3):368–73.
83. Gupta MK, Vadde R. Genetic basis of adaptation and maladaptation via balancing selection. Zoology. 2019 Jul;10:125693.
84. Gupta MK, Vadde R. Divergent evolution and purifying selection of the Type 2 diabetes gene sequences in Drosophila: a phylogenomic study Genetica [Internet]. 2020 Aug 17 [cited 2020 Aug 29]; https://doi.org/10.1007/s10709-020-00101-7
85. Hernández-Domínguez EM, Castillo-Ortega LS, García-Esquivel Y, Mandujano-González V, Díaz-Godínez G, Álvarez-Cervantes J. Bioinformatics as a Tool for the Structural and Evolutionary Analysis of Proteins. Comput Biol Chem [Internet]. 2019 Oct 22; cited 2020 Oct 5]; Available from: https://www.intechopen.com/online-first/bioinformatics-as-a-tool-for-the-structural-and-evolutionary-analysis-of-proteins.
86. Merkl R, Sterner R. Ancestral protein reconstruction: techniques and applications. Biol Chem. 2016 Jan;397(1):1–21.
87. Tyzack JD, Furnham N, Sillitoe I, Orengo CM, Thornton JM. Understanding enzyme function evolution from a computational perspective. Curr Opin Struct Biol. 2017;47:131–9.
88. Kaminska KH, Milanowska K, Bujnicki JM. The basics of protein sequence analysis. Prediction of protein structures, functions, and interactions. Wiley Online Library; 2009. p. 1–38.
89. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 2003 Oct;13(10):2213–9.
90. Ames RM, Money D, Ghatge VP, Whelan S, Lovell SC. Determining the evolutionary history of gene families. Bioinformatics. 2012 Jan 1;28(1):48–55.
91. Gupta MK, Vadde R, Gouda G, Donde R, Kumar J, Nayak S, et al. The impact of natural selection on gene associated with panicle number formation in Oryza sativa. Canad J Biotechnol Longueuil. 2017 Oct;1(Special):198.
92. Aflitos S, Schijlen E, de Jong H, de Ridder D, Smit S, Finkers R, et al. Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. Plant J. 2014;80(1):136–48.
93. Aflitos SA, Sanchez-Perez G, de Ridder D, Fransz P, Schranz ME, de Jong H, et al. Introgression browser: high-throughput whole-genome SNP visualization. Plant J. 2015;82(1):174–82.
94. Subbaiyan GK, Waters DLE, Katiyar SK, Sadananda AR, Vaddadi S, Henry RJ. Genome-wide DNA polymorphisms in elite indica rice inbreds discovered by whole-genome sequencing. Plant Biotechnol J. 2012;10(6):623–34.

95. Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol. 2012 Jan;30(1):105–11.
96. Lai J, Li R, Xu X, Jin W, Xu M, Zhao H, et al. Genome-wide patterns of genetic variation among elite maize inbred lines. Nat Genet. 2010 Nov;42(11):1027–30.
97. Hufford MB, Xu X, van Heerwaarden J, Pyhäjärvi T, Chia J-M, Cartwright RA, et al. Comparative population genomics of maize domestication and improvement. Nat Genet. 2012 Jul;44(7):808–11.
98. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010 Dec;42(12):1053–9.
99. Gupta MK, Vadde R, Donde R, Gouda G, Kumar J, Nayak S, et al. Insights into the structure–function relationship of brown plant hopper resistance protein, Bph14 of rice plant: a computational structural biology approach. J Biomol Struct Dyn. 2018 Apr 10;37(7):1–17.
100. Donde R, Gupta MK, Gouda G, Kumar J, Vadde R, Sahoo KK, et al. Computational characterization of structural and functional roles of DREB1A, DREB1B and DREB1C in enhancing cold tolerance in rice plant. Amino Acids. 2019 May 1;51(5):839–53.
101. Mendes LW, Kuramae EE, Navarrete AA, van Veen JA, Tsai SM. Taxonomical and functional microbial community selection in soybean rhizosphere. ISME J. 2014 Aug;8 (8):1577–87.
102. Carbonetto B, Rascovan N, Álvarez R, Mentaberry A, Vázquez MP. Structure, composition and metagenomic profile of soil microbiomes associated to agricultural land use and tillage Systems in Argentine Pampas. PLoS One. 2014 Jun 12;9(6):e99949.
103. Bevivino A, Paganin P, Bacci G, Florio A, Pellicer MS, Papaleo MC, et al. Soil bacterial community response to differences in agricultural management along with seasonal changes in a Mediterranean region. PLoS One. 2014;9(8):e105515.
104. Pan Y, Cassman N, de Hollander M, Mendes LW, Korevaar H, Geerts RHEM, et al. Impact of long-term N, P, K, and NPK fertilization on the composition and potential functions of the bacterial community in grassland soil. FEMS Microbiol Ecol. 2014 Oct 1;90(1):195–205.
105. Souza RC, Hungria M, Cantão ME, Vasconcelos ATR, Nogueira MA, Vicente VA. Metagenomic analysis reveals microbial functional redundancies and specificities in a soil under different tillage and crop-management regimes. Appl Soil Ecol. 2015 Feb 1;86:106–12.
106. Lavecchia A, Curci M, Jangid K, Whitman WB, Ricciuti P, Pascazio S, et al. Microbial 16S gene-based composition of a sorghum cropped rhizosphere soil under different fertilization managements. Biol Fertil Soils. 2015 Aug 1;51(6):661–72.
107. Pii Y, Borruso L, Brusetti L, Crecchio C, Cesco S, Mimmo T. The interaction between iron nutrition, plant species and soil type shapes the rhizosphere microbiome. Plant Physiol Biochem. 2016 Feb 1;99:39–48.
108. Stempfhuber B, Richter-Heitmann T, Regan KM, Kölbl A, Wüst PK, Marhan S, et al. Spatial Interaction of Archaeal Ammonia-Oxidizers and Nitrite-Oxidizing Bacteria in an Unfertilized Grassland Soil. Front Microbiol [Internet]. 2016 Jan 22; [cited 2020 Sep 11];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4722141/.
109. Valdivia-Granda WA. Bioinformatics for biodefense: challenges and opportunities. Biosecurity and Bioterrorism: Biodefense Strategy, Practice, and Science. 2010 Mar 1;8 (1):69–77.
110. Tegos GP. Biodefense Virulence. 2013 Nov 15;4(8):740–4.
111. Khan NT. The emerging role of bioinformatics in biotechnology. JBBS. 2018 Aug 7;1(3):13.
112. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36:64–70.
113. Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. J Appl Genet. 2011 Nov;52(4):413–35.

114. Kanz C, Aldebert P, Althorpe N, Baker W, Baldwin A, Bates K, et al. The EMBL nucleotide sequence database. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D29–33.
115. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL. GenBank. Nucleic Acids Res. 2008 Jan;36(Database issue):D25–30.
116. Duvick J, Fu A, Muppirala U, Sabharwal M, Wilkerson MD, Lawrence CJ, et al. PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res. 2008 Jan;36(Database issue): D959–65.
117. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40(Database issue):D1178–86.
118. Matthews DE, Carollo VL, Lazo GR, Anderson OD. GrainGenes, the genome database for small-grain crops. Nucleic Acids Res. 2003 Jan 1;31(1):183–6.
119. Lai K, Lorenc MT, Edwards D. Genomic databases for crop improvement. Agronomy. 2012 Mar;2(1):62–73.
120. Hu H, Scheben A, Edwards D. Advances in integrating genomics and bioinformatics in the plant breeding pipeline. Agriculture. 2018 Jun;8(6):75.
121. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007 Sep;81(3):559–75.
122. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics (Oxford, England). 2007 Oct 1;23(19):2633–5.
123. Rhee SY. Bioinformatics. Current limitations and insights for the future. Plant Physiol. 2005 Jun 1;138(2):569–70.

Statistics for Bioinformatics

2

Manoj Kumar Gupta, Gayatri Gouda, Ravindra Donde, S. Sabarinathan, Goutam Kumar Dash, Menaka Ponnana, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

High-throughput methods are rapidly becoming prevalent in biological sciences and clinical studies. Thus, more rigorous statistical techniques are required to accurately predict the resulting big datasets. In this chapter, the authors attempted to understand how statisticians develop and employ various strategies to investigate and analyze these big datasets. Results obtained revealed that, to date, several statistical methods have been developed for analyzing large-scale biological data, like multiple testing, unsupervised learning and data

Goutam Kumar Dash and Menaka Ponnana contributed equally with all other contributors.

M. K. Gupta (✉) · G. Gouda · R. Donde · G. K. Dash · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_2

visualization, clustering, and bootstrapping. However, a larger dataset analysis often faces various challenges, like multiple comparisons, high dimensionality, small n and large p problem, noise, and heterogeneous information. Additionally, while statisticians play an essential role in numerous bioinformatics studies, most of them are only interested in obtained and preprocessed data. Few researchers have proposed that statistician participation in the initial data collection and preprocessing phase will minimize errors and contribute to more critical scientific conclusions.

Keywords

Bioinformatics · Clustering · Experimental Design · Genomics · Statistics · Statistical Modeling

Abbreviation

ANNs	Artificial Neural Networks
CI	Confidence Interval
FDR	False Discovery Rate
HMM	Hidden Markov Model
MDS	Multidimensional Scaling
PCA	Principal Component Analysis
PHATE	Potential of Heat-Diffusion for Affinity-Based Transition Embedding
PPI	Protein-Protein Interaction
UMAP	Uniform Manifold Approximation and Projection for Dimension Reduction

2.1 Introduction

Bioinformatics refers to an interdisciplinary field that requires knowledge about various domains like molecular biology, computer science, and statistics, to analyze the larger biological dataset. However, analyzing larger biological dataset is extremely challenging for several reasons, like multiple comparison, small *n* and large *p* problems, and high dimensionality associated with biological datasets [1]. While molecular biologists and computer scientists play a crucial role in generating data, statisticians have unique skills to understand variability and uncertainty associated with the larger dataset. These skills are required to develop methods for exploring and extracting information from a larger dataset [2]. Statisticians are basically "data scientists" who comprehend the profound impact of error propagation generated from multistep processing algorithms, the possible loss of information from overly reductionistic feature extraction approaches, and sampling design decisions on downstream analysis. They also have expertise in inferential reasoning that equips them to recognize the significance of multiple testing adjustments for avoiding false results as discoveries. They can also design algorithms properly for searching high-dimensional spaces and building predictive models while estimating precise measures of their predictive accuracy [2]. Statistics play a crucial role in various bioinformatics studies, ranging from detecting drug discovery to identifying variants associated with multiple diseases or traits [3–7]. Though statisticians play a crucial role in various bioinformatics studies, most of them are only interested in data that have already been collected and preprocessed. Few researchers have suggested that statisticians' involvement during the initial phase of data collection and preprocessing may reduce error and derive scientific conclusions more significantly [2]. Thus, in this chapter, the author attempts to understand the challenges associated with analyzing larger biological datasets and how statisticians develop and employ various techniques for investigating and examining these larger datasets [1].

2.2 Challenges

Major challenges associated with larger datasets are multiple comparisons, high dimensionality, small n and large p problem, noise, and heterogeneous information.

2.2.1 Multiple Comparisons Issue

When a dataset is subjected to statistical testing several times, either at several time points or several endpoints or through several subgroups, it is called multiple comparisons, multiple testing problems, and multiplicity [8]. The result of each test is generally inferred either through *P* value or confidence interval. The "*P*" value denotes that the probability of obtaining the result by chance and is usually inferred in term of "alpha" or "Type 1 error." When a comparative study is performed between more than one group, Type 1 error estimates the probability of difference between groups that occurs by chance. This is contrary to "Type 2 error," which fails to identify any real difference between groups. Conventionally, Type 1 error is set at 5% or less. By doing so, we are at least 95% confident that if any difference occurs between groups, it is true and not by chance. For a 5% Type 1 error, the "*P*" value less than 0.05 is considered statistically significant. One common misconception is that a group with lower *P* values is more effective than a group with higher *P* values. This is only applicable to a given sample size and is inappropriate for comparing experiments of different sizes. For a given sample size, a *P* value of 0.05 and 0.01 merely denotes that the results obtained due to chance are 5% and 1%, respectively [9].

However, the *P* value often fails to give information about the uncertainty of the result obtained. Hence, confidence interval (CI) and *P* value often serve as better parameters to infer any statistical result. CI provides the range of plausible values for the estimated variable that might be obtained if the study was repeated on multiple samples of the same size drawn from the population. The defined probability is known as the "confidence level," and the CI's termination points are known as the "confidence limits" [10]. Traditionally, the CI is generally set at 95%. This is demarcated as "a range of values for a variable of interest constructed so that this range has a 95% probability of including the true value of the variable." Hence, we can be 95% confident that the actual values lie somewhere between 95% CIs [10]. For instance, if the mean, as well as the 95% CIs of the systolic blood pressure during an investigation, is 130 mmHg (95% CI, 115–142), we are 95% sure that the actual mean of systolic blood pressure in the population lies between 115 mmHg and 142 mmHg [9]. Thus, the 95% CI is equivalent to hypothesis testing with *P value* <0.05 [10].

The main advantage of using CI is that we can also determine significance from it. For instance, if CI determines values indicating "no effect," it denotes a nonsignificant difference. If it excludes value indicating "no effect," it denotes a significant difference. Thus, along with "statistical significance" (*P value*<0.05), intervals representing the smallest as well as the largest effects that are likely to happen can also be estimated via CI. Another benefit of using CI in comparison to *P* value is that CI provides additional information. For example, the lower and upper limits of the CI suggest about the nature of the effect. The narrower is the CI, the better fit it is [10]. Thus, as recommended by Akobeng AK, CI provides more information than *P* value, and hence, reporting CI by the researchers is highly recommended during any study [11]. However, other studies suggested that these two statistical parameters are not contradictory but complementary because if we know the sample size as well as the dispersion, we can easily determine the P value or vice versa [12].

Another important challenge authors face during multiple testing is when researchers attempt to save a negative study. If the main endpoint does not demonstrate statistical significance, considering several other less important comparisons often generates a "positive" result, specifically when there are numerous such comparisons. Hence, researchers may analyze several endpoints, among several subsets of samples, employing multiple statistical tests, and so forth, so the possibility for multiplicity can be significant [13, 14]. One valid example is studying the impact of treatment among a patient's subgroup based on their prognostic characteristics, such as sex, tumor location, age, stage, grade, and histology. If three conditions are considered, e.g., analysis at several endpoints, between several samples' subset, using different statistical tests, only eight ($=2^3$) subset can be formed. However, among these eight, only one out of three (33% likelihood) is likely to have a significant impact (*P* value <0.05) [8].

These problems with *P* value can be overcome by employing two techniques: the family-wise error rate and the false discovery rate (FDR). While the former tries to control the overall false-positive rate for all comparisons (e.g., Bonferroni, Tukey, Holm's step-down, and Hochberg approaches), the later attempts to control the portion of "false significant results" among the significant results only (e.g., Benjamini and Hochberg approach). Hence, to avoid such issues, the researcher must understand the importance of various statistical testing before starting their work. They must precisely select the most appropriate test required for their study and also describe them in detail within the research protocol, along with suitable adjustments for multiple testing [8].

2.2.2 High Dimensionality

Due to the recent development in sequencing technologies, there is an accumulation of huge biological data. Since humans are visual learners, it is highly required that these datasets are instinctively introduced to investigators for comprehending both the general shape and the fine granular structure of the information. This is particularly noteworthy in biological frameworks, where structure occurs at a wide range of scales, and a reliable observation can initiate hypothesis generation [15]. To date, numerous dimensionality reduction techniques for visualization have been developed [16–18], of which principal component analysis (PCA) [19] and t-SNE [20–22] are most commonly used. Regardless, these methods are suboptimal approaches for examining high-dimensional regular data because they are sensitive to noise, and methods, like Isomap and PCA [17], are often incapable of removing this noise visualization, which in turn make it difficult to recognize fine-grained local structures. Additionally, nonlinear visualization methods, e.g., t-SNE, frequently complicate the overall data framework. Several dimensionality reduction techniques, like diffusion maps and PCA, are not primarily designed for visualization; hence, they often fail to optimize during two-dimensional visualization [15].

Moreover, dimensionality reduction techniques often are also deprived of computational scalability. The rate at which biological data are generating every year is outpacing Moore's Law, which states that "the density of transistors on the die was doubling every 18 months" [23]. Approaches like multidimensional scaling (MDS) and t-SNE also often fail to scale datasets because of memory constraints or speed. Though few heuristic improvements of t-SNE [22, 24] and MDS [15] are reported to solve these problems, they work only on smaller datasets, which in turn severely limits their application in the medium to long term. Few strategies also attempt to reduce visualization challenges by straightforwardly imposing an inherent structure or fixed geometry on the data. Nevertheless, strategies that impose a structure on the data often fail to report the user about the correct structural assumption. For instance, any data can be modified to fit cluster or tree via t-SNE and Monocle2, respectively [15]. While these techniques help the dataset that fit their earlier presumption, they may produce ambiguous results and, hence, are often unfit for data exploration or hypothesis production [15].

To overcome these problems, recently, Moon and the team developed a novel dimensional reduction technique for data visualization, namely "potential of heat diffusion for affinity-based transition embedding (PHATE)." PHATE produces a low-dimensional embedding, particular for visualization, which represents an accurate and denoised portrayal of both global and the local structures of a dataset in the desired dimensions number without stating any robust assumptions on the data structure and is exceptionally scalable both in runtime and memory [15].

2.2.3 Large P and Small N Problem

Analysis of high-throughput genomic data often faces a “large p and small n” or “short-fat data” problem, which mainly arises when the dimension of the covariate, i.e., the number of genes, (p) significantly exceeds sample sizes (n). This problem becomes worse when independent variables are in multiple correlations [25]. For instance, owing to discovering a large number of single nucleotide polymorphisms in a single experiment, “large p and small n” is one of the major problems that researchers face while analyzing human molecular genetic data. Researchers also face this problem while analyzing microarray datasets, where the expression level of a thousand genes is generated from fewer subjects [26]. Additionally, due to cryptic relationships between any trait and molecular markers, researchers often face problems while deducing their relation using traditional statistical models [27]. This problem can be overcome by reducing the covariate’s number through variable selection [28] or projecting them toward lesser dimensions using principal components or similar approaches [29]. For instance, the association between any trait and molecular markers is generally investigated employing a naïve single regression model for every molecular marker and linear regression models employing the Bayesian framework and few machine learning approaches that generally ignore nonlinearity as interactions [30]. Like artificial neural networks (ANNs) that require less formal statistical training, few computational approaches also serve as an essential approach for extracting the vital information from larger data [31, 32]. Recently, numerous penalized techniques, like ridge [33] and LASSO [34], have also been developed. Earlier studies have claimed that penalized approaches generally provided more accurate results and simple interpretations than nonpenalized approaches, specifically when the variables number is larger than the sample number. Few penalized approaches automatically select appropriate variables by assigning the coefficients of irrelevant variables to zero. Additionally, penalized methods improvise the prediction accuracy by reducing the nonzero elements’ coefficients using the data-adaptive adjusting variable [35]. Thus, there is an urgent requirement to develop approaches like ANNs and penalized approaches to solve this problem more effectively.

2.2.4 Noise

Since the last decades, several high-throughput sequencing technologies have been employed to study various biological systems and processes for measuring readouts like protein–DNA interactions using ChIP-Seq [36] and mRNA concentration employing RNA-seq [37]. However, while retrieving information from the high-throughput data, researchers often experience noise, which needs to be characterized and control before concluding any result [38]. The noise level in high-throughput data is influenced by several factors like the robustness and the quality of the assay itself and the robotic platform’s quality [39]. For instance, Illumina sequencing technologies generate hundreds of “short” reads, having a mean substitution error rate of $<1\%$, and INDEL rates orders of lower magnitude [40]. Additionally, these errors are also associated with the read position, which causes position-dependent noise characteristics and affects downstream analysis, mainly during variant calling [41]. Since these variants serve as a key biomarker during drug design, extreme care needs to be taken while identifying a disease or trait-associated variants [41]. For overcoming this problem, to date, numerous tools and techniques have been developed. These denoisers primarily try to repair sequencing errors via modifying every base in the read while maintaining the original quality scores [41]. For instance, approaches like maximally informative models [42] and precise physical models [43] may help us to detect sequence–function relationships even when the noise characteristics are unknown [38]. Recently, Fischer-Hwang and the team developed a novel denoising tool, namely SAMDUDE, which takes benefit of alignment information present within the SAM file for both denoising reads and updating quality scores [41]. Thus, there is an urgent requirement for the development of effective and potent denoising software tools and techniques like SAMDUDE, which will be highly useful for researchers and clinicians during the treatment of any disease or drug discovery process.

2.2.5 Heterogeneous Information

Recent research activity within the biological domain is continuously producing huge amounts of extremely heterogeneous data, primarily comprised of results from a high-throughput experiment, clinical records, and publication collections (Fig. 2.1). These big data impose enormous challenges during data mining, data integration, and knowledge discovery due to their complex, heterogeneous, uncertainty, dynamics, and high-dimensional nature. Hence, analyzing these big data by a single person is impossible. A group of bioinformatician scientists with diverse expertise, like database management, performing expression analysis, and protein–protein interaction (PPI), may play a key role in addressing these issues more effectively [44].

Noise causes include error in estimation and uncertainty in sampling. However, real heterogeneity is the product of true impact difference due to (1) dynamic biological existence that involves feedback loops as well as temporal associations; and (2) multifactorial complexity. Growing the sample sizes is the best way to circumvent noise and obtain accurate impact sizes, but only during research and standardization and calibration that restrict the generalizability of the conclusions may be modified if necessary (Adapted from [45].

2.3 Application of Statistics in Bioinformatics

Various statistic techniques, like probability and Bayes's theory, hypothesis testing and significance, clustering and classification, multidimensional analysis and visualization, statistical models, experimental designs, statistical resampling techniques,

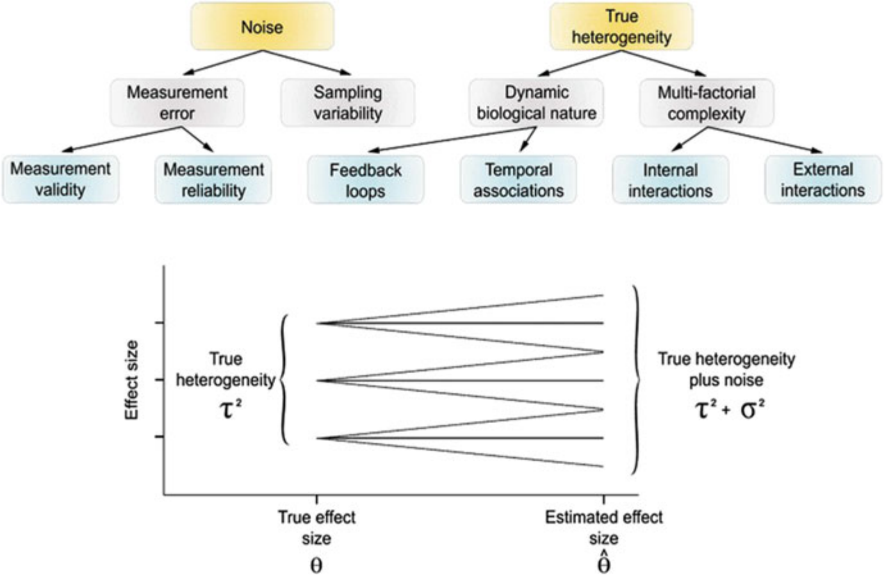


Fig. 2.1 Noise and actual heterogeneity in complicated structures

and statistical network analysis, are employed in a wide range of bioinformatics analysis [1] (Table 2.1).

2.3.1 Probability and Bayes's Theory

Probabilistic theory plays a crucial role in most research fields ranging from molecular biology to sociology. It estimates the likelihood of an event that will occur during a random phenomenon [57, 58]. For decades, probability theory has been widely employed for answering numerous biological questions, like what will be the probability of occurrence of each base pair at a specific position in a given set of sequence, what will be the probability of forming a gene mutation that causes cancer in a population, and what will be the probability that a gene will be unregulated in a given microarray dataset. The probability of occurrence of any event is estimated as P(A) = n (E) / n (S), where P(A), n(E), and n(S) denote the probability, number of a favorable outcome, and total number of outcomes, respectively. The value of P (A) ranges between 0 and 1, where 0 and 1 designate impossibility and certainty, respectively. The higher the probability, the higher the likelihood that the event will occur [57, 58]. Though this approach will work effectively for estimating the probability of a single event at a time (e.g., for estimating the probability of occurrence of each base pair at a specific position in a given set of sequence), to estimate the probability of multiple events, we have to employ joint probability and

Table 2.1 Statistic techniques and their application in bioinformatics research

Statistical techniques	Application	Software and packages	References
Probability and Bayes's theory	Estimate the likelihood of an event that will occur during a random phenomenon	IBM SPSS	https://www.ibm.com/products/spss-statistics
		R packages	[46]
		MATLAB	https://www.mathworks.com
		XLSTAT	https://www.xlstat.com/en/
		BayesiaLab	https://www.bayesia.com/
Hypothesis testing and significance	Determining the probability that a given hypothesis is true	IBM SPSS	https://www.ibm.com/products/spss-statistics
		R packages	[46]
		MATLAB	https://www.mathworks.com
		XLSTAT	https://www.xlstat.com/en/
		BayesiaLab	https://www.bayesia.com/
Clustering and classification	Resolve several complex biological issues relating to gene regulation, specific drug design, gene co-expression, functional animal and animal system research, protein–protein interactions, and organism–environment interactions	R packages	[46]
		MATLAB	https://www.mathworks.com
		Cytoscape	[47]
		Gephi	[48]
Multidimensional analysis and visualization	Unmasking patterns and producing significant information from scientific data	PCA	[49]
		UMAP (uniform manifold approximation and projection for dimension reduction)	[50]
		2.1.1 t-SNE (t-distributed stochastic neighboring embedding),	[20]
		2.1.2 focusedMDS.	[49]

(continued)

Table 2.1 (continued)

Statistical techniques	Application	Software and packages	References
Statistical models	Understanding of biological and structural functions	Neural network model	[51]
		Stoichiometric models	[51]
		Comprehensive kinetic models	[51]
		Kinetic modeling	[52]
Experimental designs	Reuse of experimental results	Factor-based experimental design	[53]
		Practical experimental design approach	[54]
		ProtocolNavigator	[55]
Resampling techniques	Replicating sampling from a provided sample/population or to approximate statistical accuracy	Bootstrapping	[56]
		Normal resampling	
		Permutation resampling	

Bayes's theorem, e.g., while estimating the probability of co-occurrence of the most common allele at two polymorphic sites within a chromosome.

Bayesian methods are used in various domains of bioinformatics. Biological sequence analysis was one of the first fields to benefit from applying Bayesian methods. It is already acknowledged that Bayesian methods are highly helpful while dealing with probabilistic models [59]. Although it is possible to calculate parameters employing traditional statistical methods (e.g., maximum likelihood through the EM algorithm) [60, 61], there are numerous fascinating problems where a traditional method would be cumbersome or unsatisfactory [62]. Concurrent multiple sequence alignment [63], prediction of transcription factor binding site and motif discovery [64, 65], and prediction of protein secondary structure [66] are the best examples. One of the main advantages of the Bayesian method is that it helps complexity to be accurately propagated across various modeling stages. Thus, traditional approaches, which estimate phylogeny based on pre-estimated multiple alignments, often fail to propagate any ambiguity from the phylogeny's alignment. The reverse is also true; alignment models depend implicitly on an expected phylogeny, so ambiguity (if any) in phylogeny will not induce ambiguity in alignment. Simultaneous prediction is also feasible using the Bayesian method [67]. The Bayesian approach makes the sufficient inclusion of prior information even during the HMM-based ab initio DNA sequence cleavage, which in turn allows us to retrieve more relevant information regarding the model variable [68]. In addition, since model structure uncertainty is handled consistently with parameter uncertainty, variable dimension algorithms, for instance, reversible jump Markov chain Monte Carlo [69] can be utilized, along with all other aspects of the model, for calculating the segments number as well as the base's order dependence [70].

Bayesian methods are also often used to unmask protein informatics. For instance, Bayesian strategies for site matching and alignment are especially useful at the structural level [71, 72]. Since mass spectrometry data experience large amounts of heterogeneity, Bayesian statistics may also play an essential role during unmasking the structure of a sample's peptide or protein [73]. Several Bayesian approaches for predicting protein–protein interactions from genomic data have also been proposed [74–76]. In 2004, Friedman and the team developed a novel framework that employs Bayesian networks to explore associations among genes based on gene expression. These approaches are particularly interesting because of their ability to capture information from noisy observations and complex stochastic processes [77].

Since bioinformatics is still an emerging field, there is a scope for developing a more effective Bayesian method. The main drawback associated with Bayesian methods developed to date is the computational demands during analysis. This, in turn, limits their usage during bioinformatics research, especially during whole-genome annotation [59]. Though probabilistic regression models, like HMM, are becoming the accepted framework for analysis [59], earlier studies have suggested that complete Bayesian methods are only the best tool for solving the dynamic statistical inference problems. Hence, increased usage of computer hardware, parallel computer clusters, and the development of Bayesian machine-based algorithms may help us answer more important complicated inferential issues in a timely manner [78].

2.3.2 Hypothesis Testing and Significance

Hypothesis testing uses various statistical tests to determine the probability that a given hypothesis is true. Usually, the hypothesis process comprises of four steps, that is, the formulation of null hypothesis (H_0) (proposes that there is no difference among certain population characteristics) and alternative hypothesis (H_1) (proposes that there is a difference among certain population characteristics), determining a statistical test to measure the truth of the null hypothesis, estimation of *P* value, and comparing the *P* value with an appropriate value (sometimes referred to as an α value). If *P* value is statistically significant, the null hypothesis gets rejected, and the alternative hypothesis is valid (https://mathworld.wolfram.com/).

Perhaps, the first thing, a bioinformatics scientist will do after identification of a biological sequence of interest is to scan a bio-sequence database. During sequence alignment, the test statistic is usually the alignment score, and the null hypothesis is that the paired sequences are not biologically related. The alternative hypothesis is that they are homologs and have a shared ancestor. Here, the *P* value of a score is the likelihood that this or higher score is a false-positive estimate from a biological relationship [79]. Hypothesis testing has also been used in exon prediction [80], biological network [81], gene ontology [82], gene set enrichment analysis [83], phylogenies [84], microarray analysis [85], genome-wide association studies [86], RNA-seq [87], and single-cell RNA sequencing [88]. We have recently employed hypothesis testing and detected that the Type 2 diabetes gene sequences in Drosophila are evolving under divergent evolution and purifying selection [6].

The emergence of high-throughput sequencing technologies has contributed to developing a huge number of hypothesis tests for different biological studies and a parallel elevation of the appropriate thresholds. However, to get more significant results, recently, several studies have effectively employed filtering approaches, using unbiased knowledge to remove the least promising tests, which in turn minimize multiple testing. Nevertheless, there are a few issues that need to be addressed before implement filtering, for instance, when does filtering work, and when is it harmful? To filter efficiently, how strong is the independent knowledge necessary? What is the filter cutoff that distinguishes samples from studies that bypass the filter? [89]. Recently, Kim and Schliekelman studied the impact of filter information quality, filter cutoff, and other filter performance variables. Result obtained revealed that if the filter is extremely likely (e.g., 70%) to rank true positive features (e.g., top 10%), then filtering will lead to a drastic improvement (e.g., tenfold) in discovery likelihood when high redundancies exist. If the redundancy between hypothesis tests is diminished, filtering becomes less effective and, hence, it is an advantage also declines rapidly [89]. Additionally, the result depends primarily on the preference of filter cutoff. Hence, if the cutoff were selected without referring to the data under consideration, it might cause significant discovery probability loss. Nevertheless, as true effect size and optimum estimation schemes remain uncertain, several authors demand the development of more robust tools and algorithms for fetching more significant information [89].

2.3.3 Clustering and Classification

Application of clustering, an unsupervised learning method, in biology has a history dating back to Aristotle’s attempt to classify various life forms [90]. Nowadays, cluster analysis marks out as one of the best solutions to cope with high-dimensional data generated from various high-throughput technologies, including microarray gene expression and RNA sequencing [91]. Clustering analysis helps us to resolve several complex biological issues relating to gene regulation, specific drug design, gene co-expression, functional animal and animal system research, protein–protein interactions, and organism–environment interactions [91]. During gene expression analysis, samples that display similar expression across all genes are clustered together, and genes that display similar expression across conditions are clustered together [92]. During gene clustering, genes and samples are considered as objects and features, respectively. However, in sample-based clustering, samples can be divided into similar groups, and samples and genes are viewed as objects and features, respectively [93]. Thus, the essential elements of gene-based and sample-based clustering techniques depend on distinct characteristics of clustering tasks for gene expression data [94].

Usually, algorithms employed during gene expression clustering can be broadly divided into two classes, namely hierarchical and partitional (Fig. 2.2). One of the

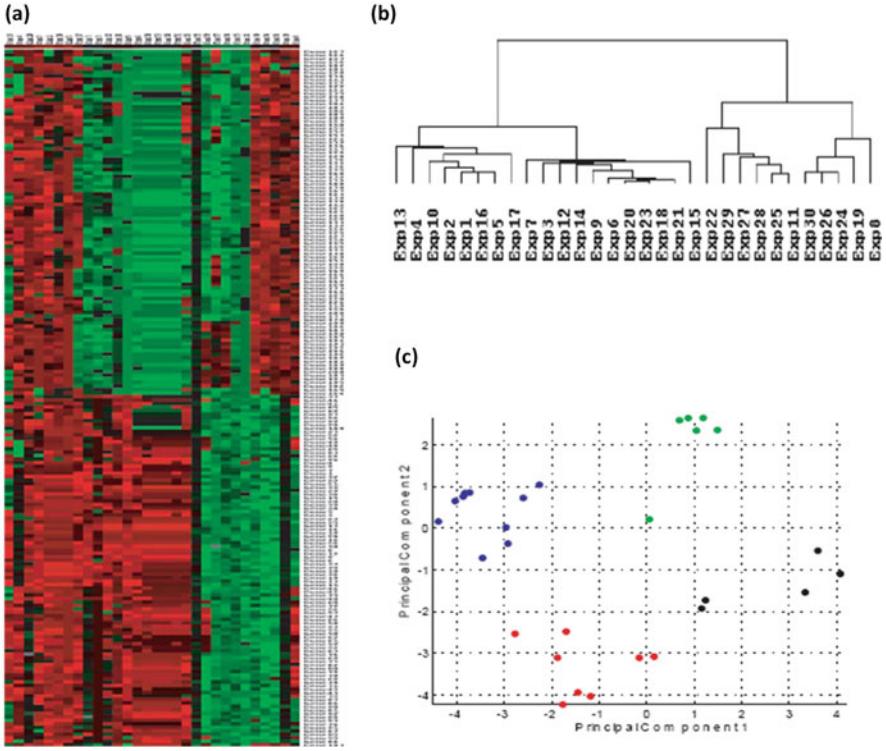


Fig. 2.2 Hierarchical cluster. (A) Heatmap interactive display, hierarchical clustering, and principal component analysis. (**a**) The illustration of a heat map of 30 virtual profiles enables the researcher to imagine four study classes along the x-axis with distinct expression models for 300 genes. The heat map helps to classify altered genes and sample clusters, but does not express any spatial correlation between clustered samples. (**b**): A dendrogram illustration generated by hierarchical clustering of the simulation data in Fig. 2.2A. A dendrogram is a diagram made up of several U-shaped lines that link artifacts to represent hierarchic clusters. Four sample clusters are developed based on distinct expression signatures in this dendrogram. (**c**): A two-dimensional schematic visualization of the study of the key components (PCA) based on the details seen in Fig. 2.2a

most primitive clustering algorithms widely used for clustering genes is a hierarchical algorithm (HC). HC consists of a group of nested clusters arranged as a tree. However, the algorithm's output is noise sensitive. It is also not sensitive to the missing data and does not provide essential information, e.g., the number of necessary clusters and confidence measures for individual clusters [95, 96]. Earlier studies have also reported that HC faces some difficulties while clustering larger datasets [92]. HC has robustness and inversion problems, which complicates the hierarchy analysis [94, 97, 98]. HC algorithms also face high computational complexity [94]. Partitional clustering splits a dataset into different disjoint clusters [99]. Because a dataset is comprised of N points, this algorithm produces K (N $\geq$ K) number database partitions, and each partition forms a cluster

[99]. However, the key disadvantage of these cluster algorithms is that the when one point is close to the middle of another cluster, or there is overlap among data points, it creates a bad result [92].

Cluster algorithms can also be grouped based on the relationship among clusters, data representation, data distribution, and other characteristics. For instance, clustering approaches can be either complete or partial. While a complete clustering assigns every gene to a cluster, partial clustering does not. The fact that gene expression data generally contain few genes or samples that are not related suggests that partial clusters are more adaptable to gene expression. Moreover, it helps us ignore various insignificant contributions by restricting the incorporation of a few genes into well-established clusters. Thus, partial clustering helps avoid situations that hold an interesting subgroup in a cluster by not adding unknown genes [100]. Clustering can also be overlapping or hard [100]. Overlapping clusters assign membership to each gene in multiple clusters. During its activity and production, hard clusters assign each gene to one cluster. An overlapping clustering can be changed into a hard clustering by allocating each gene to the cluster with a dominant membership [92]. Irrespective of the development of various clustering algorithms and techniques, Pirim and the team reported that no clustering algorithm can provide the best results for all clustering problems [91]. Hence, there is an urgent requirement for the development of sophisticated algorithms that can solve these problems.

2.3.4 Multidimensional Analysis and Visualization

Visualization is required for unmasking patterns and producing significant information from scientific data. Nevertheless, high-dimensional data often pose challenges because structures or patterns may not precisely visualize two or three dimensions except in more than three dimensions. One best example is analyzing data from comparative high-performance sequencing research, in which a key step is to explore the samples' characteristics and to detect if duplicate samples are identical and to distinguish outliers. Samples are drawn as points on a 2D plane to reflect the relationships between them. PCA that monitors data components illustrating the most uncertainty or multidimensional scaling (MDS) that try to capture and represent the correlation between points in 2D space across all measures is common to construct this form of visualization [49].

Similarly, during the single-cell RNA sequencing (RNA-seq) data analysis, one also desires to decrease the high-dimensional expression data into a 2D plot, so that related transcriptomes cells appear near together. Hence, in addition to MDS and PCA, researchers are also employing UMAP (uniform manifold approximation and projection for dimension reduction) [50] and t-distributed stochastic neighboring embedding (t-SNE) [20] approaches for data visualization. t-SNE is an optimization algorithm that employs probability distribution in both low and high dimensions for generating 2D or 3D representations, while UMAP is a manifold training approach for generating 2D or 3D representations [49]. UMAP is developed from a Riemannian geometry and algebraic topology theoretical context. The outcome is a realistic efficient algorithm relevant to data from the real world. The UMAP algorithm competes with t-SNE in efficiency and retains more of the global framework with better runtime output. UMAP avoids program limitations on the embedding dimension and is thus feasible for machine learning as a general-purpose dimension reduction technique. In the near future, developing techniques like UMAP is highly required for unmasking significant information from various biological data types, including single-cell RNA sequencing, mass cytometry, and the gut microbiome [50]. Recently, Urpa and Anders developed two tools: distnet and focusedMDS that help to determine the validity of a dimension reduction plot and analyze in-depth interactive relations between objects. The distnet investigates differences between the points' location within a two-dimensional visualization and the actual similarity between feature space points. The focusedMDS approach is an instinctual, immersive multidimensional scale technique that investigates links between data points, which might be beneficial in personalized medicine [49].

The 3D spatial arrangement of chromosomes is an exciting emerging field in genomics. However, a precise, atomic-scalable genome model is a major challenge, which can be accomplished someday by modern experiment techniques, like Hi-C; 35, that can recognize spatial chromatin contacts between genomic pairs [101]. However, these techniques have poor resolution and higher false-positive rates and, hence, to date, reliable 3D representations for chromosomes are still not determined [102]. It still continues to be difficult to view the high-dimensional datasets generated from RNA-seq studies. Gene expression values that are considered relevant are generally viewed as heat maps clustered following meticulous experimental design and statistical analysis [63], a methodology that has dominated since the first experiments on microarray [103]. Nevertheless, the magnitude of individual values or folding deviations among pairs of values can be difficult to quantify with optical illusions within these visualizations. With the increase in the rows and columns number and decrease in cell size, these effects worsen, and all-important findings are not seen as colossal heat maps. Additional problems can occur because clustered heat map's rows and columns are typically organized to group related genes and conditions, which in turn may highlight regulatory effects [104]. Nevertheless, values are also put next to each other for condition and genes without substantial correlations, which may worsen perceptional issues. Separating unrelated columns and rows can help, but these problems cannot be significantly solved [105], especially for poorly clustered genes. There may be inadequate evidence to address these genes' regulatory networks as a 1D order in such cases. So, it is critical that the support and degree for cluster-inferred relationships are also demonstrated. However, the implementation of tree diagrams further limits the heat map scale that can be shown without issue. So, it is recommended that only the most informative subset of genes and conditions can be shown in this manner [106].

2.3.5 Statistical Models

The application of concepts from information theory and stochastic modeling of symbolic processes has helped to understand the biological and structural functions precisely. Specifically, previous analyses of CG dinucleotide location along the genome have highlighted its epigenetic role in DNA methylation, showing a distinct distribution tail relative to other dinucleotides [107]. Recently, Merlotti and the team performed an analysis of the entire distribution of the CG distance over a chosen collection of higher-order species. They have also employed the best-fitting probability density function for a large number of species ($>$ 4400) that have various complexities (from bacteria to mammals) and characterize some new global characteristics [107]. The result revealed that compared with other distributions, the gamma distribution is ideal for a selected subset. When introduced to a wider range of species, this distribution's parameter permits some biologically relaxing characteristics that can be used for categorizing purposes [107].

Recently another group of researcher proposed a neural network model for classifying various cancer types [108]. They employed different machine learning techniques strategies, namely term frequency (tf)-inverse document frequency (idf), term frequency (tf)-relevance frequency (rf), and Best Match 25(BM25), for retrieving knowledge for weighting genes based on mutation information. Result obtained reveals that compared to the other representations, the BM25-tf-rf strategies lead to improved classification accuracy. Interestingly, in earlier studies, a subgroup of the resulting genes was also indicated as the target or casual genes, thereby suggesting that this approach can also be used for detecting candidate genes [108].

Incorporating and studying experimental datasets for measuring metabolic fluxes and modeling metabolic networks are also an important method in metabolic engineering. To date, several computational approaches have been proposed for modeling as well as stimulating metabolism processes, both quantitatively and qualitatively. This varies from the topological study of network models (which explores metabolite interconnections) to the stoichiometric models (where restrictions can be imposed for describing the metabolic potential flux state-space or analyze utilizing petri networks) to the comprehensive kinetic models (which adjusts the concentration of metabolites over time) [51]. A network model is the most fundamental framework for understanding a cellular biological mechanism [109]. A network can be defined as a graph in which biological entities, e.g., metabolites, transcripts, genomes, and proteins, symbolize nodes, and edges are represented via interactions between nodes, like protein–protein interaction, and biochemical transformation and co-expression [110]. In plant biology, the main networks are protein–protein exchange, gene-to-metabolite, metabolic networks, gene regulation, and transcriptional control [111]. Network analysis refers to using algorithms for classifying structurally essential elements or network components and graph theoretic models that employ statistical techniques to classify and infer complicated functional relations between them [110]. However, this topological methodology does not accommodate the system's complex actions, which in turn demand the development of other approaches [109]. Stoichiometric models are often used for analyzing larger metabolic networks (nearly1000 reactions) by applying limitations to determine the spectrum of possible metabolic flux states. Imposing those restrictions enables estimating the feasible flux distributions [112].

The increased volume of the annotated genome sequence has allowed a few metabolic models to be scaled into genome-scale metabolic model systems [113]. Almost gene-to-protein-to-reaction interaction associated with every metabolic response is reflected in the ideal genome-size metabolism model [114]. The dimensions and sophistication of genome-scale metabolic models suggest that the possible behavior, particularly flux balance analysis, can only be studied with restricted methods. While the metabolic model of genome scale has become the common method for modeling over the past couple of years, it should not be used as an individual mathematical model [109].

Metabolism flow analysis focuses primarily on experimentation evidence that provides isotopically labeled precursors (usually 13C) to the target, and after the system has reached a stable state, the distribution of metabolic intermediates and end products is analyzed [115]. A model of the target metabolic network is used to predict the labels' redistribution in a steady-state scenario. The experimentally measured pattern is contrasted to the anticipated labeling pattern, and modifications were made within model flows until the relationship between the measures projected and measured is as closely as practicable. This method is repeated several times and leads inevitably to hundreds of flux maps from which the model that provides the best prediction for the system can then be deduced [115].

Kinetic modeling describes the enzyme's dynamic activity, and therefore, it is the most predictive and detailed mathematical description. Kinetic modeling considers enzyme's dynamic activity and is usually applicable to limited portions of the metabolic network (~10 to 50 reactions) [52]. The dynamic model measures the system's time-based behavior with reference to the concentration of both metabolite and flux. However, detailed kinetic formulas for several enzymes are not usable; hence, heterologous systems or literature should be taken as hypotheses [116]. Relevant criteria to achieve a unified solution is another problem [117]. Additionally, extra care is needed while designing a statistical model for plant systems because compared to the prokaryotic cell, plant cells have a large number of the compartment and complex metabolic pathways. Hence, continuous development of more user-friendly applications, languages, statistical models, and databases that integrate and process complicated knowledge would be vital for handling the complex biological data analysis [51].

2.3.6 Experimental Designs

Experimental reproducibility is central to science's advancement [55]. Nonreproductive research limits the productivity of fundamental biological science as well as drug development, which in turn impedes the reuse of experimental results. A major factor that mainly contributes to reproductivity is uncertainty in decoding complicated laboratory techniques and prototypes from a literary form and in determining differences between various experiments. Present bioinformatics projects rely on data processing or management laboratory's information systems [55]. Big data bioinformatics approaches basically help us to unmask significant biological findings from detailed and large biological databases. However, an additional benefit from processing big data is only feasible if supported by appropriate metadata annotation. Intelligent methods are particularly required in high-performance research, for monitoring the experimental design including the research conditions and details that may be of relevance to the analysis of failures or potential research. In addition to handling this knowledge, researchers are desperately looking for an integrated architecture and interfaces for structured data annotation [53].

Recently, Friedrich and the team proposed a factor-based experimental design method allowing scientists to effectively construct large-scale studies employing a Web-based framework [53]. They introduce a novel Web-based framework implementation enabling arbitrary metadata processing. To share and modify details, they provide a human-readable, spreadsheet-based format. Subsequently, sample sheets and metadetails for data generation facilities may be generated. Data files generated after sample calculation can be transferred to a datastore, where they are automatically connected to the previously created experimental model [53].

Another group of researchers implements a practical experimental design approach to pick dynamic models from results. The process, inspired by biological uses, facilitates critical experiment design: it specifies a highly insightful set of measurement readings and time points. Based on previous results, they show systematic guarantees of design efficiency. By reducing analysis to the concept of graphical models, they demonstrate that the proposed method discovers an almost optimum option of designs with a polynomial number of assessments. Furthermore, the procedure provides the strongest constant approximation factor for polynomial complexity, unless P = NP. In comparison with proven alternatives, including ensemble noncentrality, the author assesses the system's efficiency based on example models of various heterogeneities. Effective architecture promotes the loop among modeling and experimentation: It also permits an inference of complicated processes, including those which regulate centralized metabolism to be deducted [54].

Recently, Khan and the team used the wx package to construct the graphical user interface in Python 2.7, namely ProtocolNavigator [55]. It primarily discusses the largely ignored understanding and implementation issues associated with various biological domains. It offers a scientist-friendly, emulation-based open-source platform to plan, record, and replicate biological experiments [55]. ProtocolNavigator comprises of three functional and display screens. Through the "inventory panel," the user can construct an instance "inventory" with comprehensive explanations of objects like reagents, instrumentation, and components. Interestingly, this inventory can be reuse, modify, and distribute. The previously created inventory instances are added to separate experimental objects with the bench panel (equivalent to a laboratory workbench). ProtocolNavigator's time-integrated action-based recording method is innovative in allowing researchers to record real-life experimental experience (e.g., temporal behavior variation). The map panel dynamically illustrates the experimental map or design, with spatially temporarily connected divisions and operation icons to capture action information as well as experimentally generated data. Inherently, comparing laboratory experience with real experimental evidence on a jargon-independent map provides a strong basis for communicating and distinguishing experimental design and the underlying difference in operation—an important prerequisite for reproducibility [55].

Most importantly, the completely accessible map can indeed be distributed with peers, thereby adding a remarkable capacity for refining and organizing experimental design collective and synchronized computationally, which in turn significantly minimizes project iterations and associated costs. The map could be transformed and copied into a time-stamped, systematic summary of actions at the physical benches or for dissemination. The map's datasheet can be conveniently reformatted and parsed [55]. However, authors have also suggested that there is still scope for the addition of new tools or materials that may be given higher priority by prospective consumers [55]. For instance, a distinctive but important question still remains open: How to accurately classify candidate models' parameters? This question also needs special consideration as architecture and modeling are members of the very same hypothetico-deductive method [54].

2.3.7 Resampling Techniques

Resampling techniques are a series of methods employed for either replicating sampling from a provided sample/population or approximate statistical accuracy [56, 118]. For instance, if we are doing a concurrent likelihood ratio test but we do not reach a conclusion, we will resample and repeat the test [56, 118]. Resampling techniques, like permutation procedures, also provide an appealing alternative to the traditional inferential approaches; they are flexible and need fewer hypotheses. Traditionally, the key drawbacks of using these approaches are that they were computationally expensive and sometimes demanded custom-written machine codes. However, these drawbacks are no longer a major issue, even for exceptionally large amounts of data, due to the availability of personal computers and open-source analytical software programs for the application of resampling-based approaches [119].

The key approaches for resampling techniques are bootstrapping and normal resampling (i.e., normal distribution sampling), permutation resampling (also known as rerandomization or rearrangements), and cross-validation. During bootstrapping, several smaller samples of the same scale, with substitution, are repeatedly taken from one original sample. Normal resampling is somewhat similar to bootstrapping because it is a special case of the normal shift model, which is one of the bootstrapping principles [56]. Both bootstrapping and normal resampling presume that samples are obtained from the actual (either a real or a theoretical) population. Additionally, both methods permit substitution. However, insufficient resources can preclude optimal statistics. Contrary to bootstrapping, permutation resampling required no "population." Here, resampling relies on assigning units to treatment classes. This is the reason we treat particular samples instead of populations and why it is often referred to as the traditional gold bootstrapping strategy. Another significant distinction is that permutation resampling is a sampling process without substitution. Cross-validation will evaluate a predictive model. During cross-validation, subsets of data will be discarded to be used as a validation package; other data will be used to construct a training set to predict the validation set [56, 118].

Molinaro et al. (2005) performed extensive comparisons of resampling methods with simulated (high signal-to-noise ratio) microarray (intermediate signal-to-noise ratio) and proteomic data (low signal-to-noise ratio) and estimated prediction error, including growing sample sizes with a broad number of features. The effect of the collection of features on the efficiency of the different cross-validation methods was illustrated. The findings set out the "right" resampling strategies for future studies involving high-dimensional data for preventing excessively positive estimation of the model's performance [120]. However, only a few implementations of the bootstrap approach can be found in the literature [121]. For instance, the bootstrap was employed by Zhang and Zhao (2000) in hierarchical cluster analysis. In a consensus tree, they summarize individual dendrograms [122]. Their approach includes estimation of gene expression inaccuracy. Bhattacharjee et al. (2001) also employed bootstrapping to determine cluster stability and validate the result obtained through hierarchical clustering [123]. In another study, Kerr and Churchill (2001) employed the bootstrap for determine the stability of the outcomes of cluster tests. It uses an ANOVA model to predict gene expression and consider the microarray data variance from other sources. The proportion of genes within bootstrap clusters is a stability indicator [124]. Dudoit and Fridlyand employ bagging (bootstrapping and aggregation) to increase the cluster partitioning process's precision. The separate partitions are merged into one final partition, or a new dissimilarity matrix is created and acts as the basis for final classification [125].

Since bootstrapping is a substitution drawing and the bootstrap sample size is the same as the initial data size, certain findings are excluded during analysis. The estimated points' proportion in the initial study from the bootstrap sample is estimated as $(1–1/n)^n$, which converges to 1/e for $n \rightarrow \infty$ or ~ 36.8% [121]. Recently, Gana and the team proposed the usage of continuous weights instead of bootstrap. Continuous weight excludes zero elements and makes noninteger weights instead. Thus, the entire dimensionality of space contributing to retention is reflected in the sampling sample for every element of the original dataset [121]. Comparative analysis of continuous weights with bootstrapping utilizing real datasets as well as simulation studies shows the benefit of continuous weights, particularly when there are few observations in the dataset, few differentially expressed genes, and low folding of differentially expressed genes. They even advocated using continuous weights in both small and large datasets since continuous weights yield at least the same effects as traditional bootstrapping and often exceed it [121].

2.3.8 Statistical Network Analysis

Euler first developed the concept of graph theory, a subfield of mathematics, and employed this method in 1736 to prove the famous “Seven Bridges of Konigsberg” problem. The graph theory’s significant developments are mainly due to Paul Erdos and Alfred Renyi, who have researched random graphs’ characteristics. They just wondered, what will be the outcome of tossing a couple buttons onto a table and then tried to link them arbitrarily? Also known for proving the “map color” problem, Erdos and Renyi address the query, what is the minimum amount of colors required to clearly color a map [126]. Watts and Strogatz‘s pioneering publications in 1998 [127] and Barabási and Albert in 1999 [128] popularized the idea that complex structures can be interpreted as networks where components within a complex system can be described as nodes and connected by their interactions, called edges [126].

As earlier stated above, networks provide graphic representations of relationships (edges) among variables (nodes) (Fig. 2.1). Edges depict detailed information regarding the path as well as the power of node partnership. The edge may be positive (e.g., a positive covariation/correlation among variables) or negative (e.g., a negative covariation/correlation among variables). The polarity among the relationships is graphically represented through multiple colored lines that reflect the ends. The edges may or may not be weighted. A weighted edge represents the intensity of the interaction between nodes by the shift in the edge’s thickness and color linking the nodes. Alternatively, the edge can be unweighted and reflect merely the existence versus lack of a connection; the absence of connection in such a network results in nodes with no linking edges.

There are two types of edges in the network: (1) a direct edge—the nodes are joined, and the edge head has a single directional arrowhead; or (2) an undirected edge—the nodes have a connecting line that shows a connection with one another, but no arrowheads to guide the direction of impact (Fig. 2.3). Networks may be defined as guided (i.e., all edges are directed) or undirected (i.e., there are no edges). A guided network can either be acyclic (i.e., we could not originate from a node and

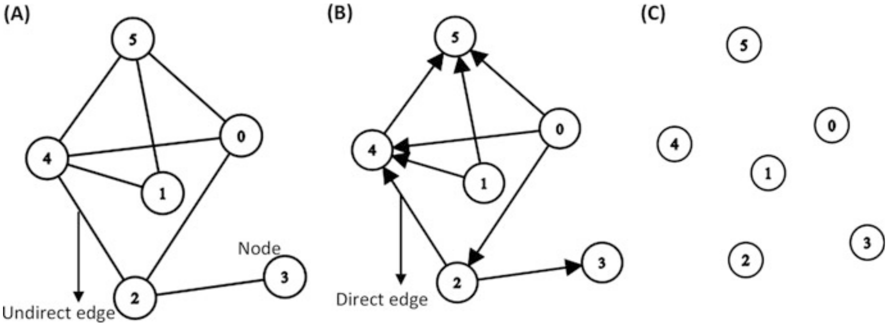


Fig. 2.3 (**a**) Guided graph with an undirected edge, (**b**) guided graph with directed edge, and (**c**) unguided graph

return to the node by observing the directional borders) or cyclic (i.e., we can obey the path of a node to return to the node) [129, 130].

The network analysis enables us to estimate dynamic relationship patterns and evaluate the network structure to identify key network functionality [130]. For instance, variables such as tension, functional activity, social pressure, and the nutritional content of drink and food reflect nodes in the network, and the positive and negative relationships between these nodes are edges. Nodes, edges, and networks are also often called vertices, links, and graphs, respectively. Cross-sectional and longitudinal time-series data may be used to approximate networks; networks may also be evaluated at community or person levels. Cross-sectional community data can expose group-level conditional independence relationships. Individualized networks focused on time series and may provide insights into a particular individual across time. Moreover, the networks of various communities may be linked. Overall, network research represents a wide variety of computational methods for analyzing multiple network structures [130].

Network analysis has been used in various research fields ranging from psychology [130] to bioinformatics [4, 7, 131, 132]. Bringmann et al. (2016) described the logic of network approach, related approaches, and visualization and provided an analytical illustration to demonstrate the association between everyday emotional variations and neuroticism. The findings indicate that people with elevated neuroticism have a denser relational network relative to their less neurotic counterparts. This impact is particularly prominent for the negative emotion network, consistent with previous research that observed a denser network in stressed subjects than in healthy subjects. In summary, the author demonstrates how network methodology can provide new methods for psychology examining complex processes. Cordeddu et al. (2009) described a novel gene that can trigger “Noonan-like syndrome” when mutated [133]. SHOC2 was discovered by introducing the previously identified genes, namely *KRAS, SOS1*, *RAF1,* and *PTPN11,* in the Genes2Networks tool [134]. Zaidel-Bar et al. (2007) published a signaling network composed of the molecular components and associations of focal adhesions [135].

Earlier studies have revealed that two genes within a molecular network having higher-order topological similarities are most likely to interact with each other and may be associated with the similar or same phenotypes [136–138]. Additionally, disease phenotypic data may also enhance the accuracy of key gene prediction associated with rare disease phenotypes [139]. This information has led to various computational tools for predicting candidate genes for inherited diseases [139]. Dezső et al. (2009) developed a novel approach based on the shortest path betweenness for prioritizing candidate genes within protein–protein networks. Gene having the shortest distance from the disease node is considered as key candidate gene [140]. Wu et al. (2008) developed a regression model that on the basis of correlation scores quantifies the relationship among gene nearness and phenotype likenesses within the PPI network and identifies best possible disease-associated candidate genes [141]. Li and Patra (2010) developed a novel approach for prioritizing disease-associated key genes by expanding the random walk with restart algorithm on a heterogeneous network created via linking the gene network as well as the phenotype similarity network employing the known gene-phenotype connections [142]. Zhu et al. (2013) hypothesized another method that detects disease-associated key genes based on the cosine angle of corresponding diffusion profiles and linear correlation coefficient. The diffusion profile was defined as the "stationary distribution of all genes under a random walk with restart on the phenotype similarity network" [143]. Every gene's diffusion profile was retrieved by flattening the probability distribution over the PPI network while beginning a walk from the gene. Thus, integrative analysis of biological system via networks analysis [139, 144, 145] may help us to predict a disease/trait-associated key gene and protein–protein interaction [44], gene co-expression (Co-Ex) networks [146, 147], and metabolic interaction (MI) networks [148, 149] more effectively.

Miele et al. (2019) recommended nine tips to deter common falls and to strengthen the study of biologists' network results, namely (1) at first, only formulation of questions; subsequent application of networks, (2) classify the network data properly, (3) using specialized tools for network analysis, (4) be mindful that network visualization can be useful but deceptive, (5) ignore using metrics blindly; instead grasp formulas, (6) resist blind usage clustering algorithm; evaluate their disparity alternatively, (7) do not pick the simple way to model networks, (8) review data on different network layers, and (9) immerse yourself in the network literature outside your domain. These nine tips may serve as a path for a data analyst to get a foot in the network data analysis entrance. However, these guidelines are not limited, and special attention may be given to a few issues, like diffusion on networks. Nevertheless, the nonspecialist must try to positively answer research problems by learning network analysis more comprehensively [150]. Thus, statistics play a key role in bioinformatics, and their right usage may help us to answer biological problems more precisely.

2.4 Conclusion and Future Perspective

In conclusion, numerous fascinating advancements in biotechnology have produced huge quantities of diverse kinds of big data, which demand powerful and suitable computational statistical tools with biological expertise and computer algorithms. Statisticians have played a leading role in bioinformatics, helping researchers develop robust design and analysis methods to derive meanings of biological knowledge from the rich treasure of multiplatform genomics. Their profound understanding of the scientific method and complexity, and ambiguity has enabled them to play a vital role in this undertaking. Various statistic techniques, like probability and Bayes's theory, hypothesis testing and significance, clustering and classification, multidimensional analysis and visualization, statistical models, experimental designs, statistical resampling techniques, and statistical network analysis are continuously employing in a wide range of bioinformatics analysis [1]. However, there is still scope for further work and potential improvements. For instance, bioinformatics is an emerging field, and the science community needs innovative approaches to incorporate knowledge through various channels and obtain more systematic insights into the underlying molecular biology. These approaches need to combine statistical rigor while connection creation and analytical reliability for scaling up to large data settings and outcome interpretability, which will help the researcher understand them precisely. Thus, in the near future, there is an urgent requirement that the statistical community must find more efficient ways of integrating this knowledge into the modeling method, which could contribute to stronger forecasts and findings and increased interpretability of the data.

Conflict of Interest None.

Additional Information Figure 2.1. (CC BY 4.0) [45] and Fig. 2.2 (A) CC BY 2.0 [151] have been reused under Creative Commons Attribution licenses.

References

1. Lee JK. Road to statistical bioinformatics. In: Statistical bioinformatics [internet]. Hoboken: John Wiley & Sons, Ltd.; 2010. p. 1–6. [cited 2020 Sep 13]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470567647.ch1.
2. Morris JS, Baladandayuthapani V. Statistical contributions to bioinformatics: design, Modeling, structure learning, and integration. Stat Modelling. 2017;17:245–89.
3. Gupta MK, Sarojamma V, Reddy MR, Shaik JB, Vadde R. Computational biology: toward early detection of pancreatic Cancer. CRO [Internet]. 2019;24. [cited 2019 Nov 21]. Available from: http://www.dl.begellhouse.com/journals/439f422d0783386a,4e35fd64043789bc,3d24f17d426f6f57.html
4. Gupta MK, Vadde R, Gouda G, Donde R, Kumar J, Behera L. Computational approach to understand molecular mechanism involved in BPH resistance in Bt- rice plant. J Mol Graph Model. 2019;88:209–20.
5. Gupta MK, Vadde R. Genetic Basis of Adaptation and Maladaptation via Balancing Selection. Zoology 2019; 136125693.
6. Gupta MK, Vadde R. Divergent evolution and purifying selection of the type 2 diabetes gene sequences in Drosophila: a phylogenomic study. Genetica [Internet]. 2020 . [cited 2020 Aug 29]; https://doi.org/10.1007/s10709-020-00101-7.
7. Gouda G, Gupta MK, Donde R, Kumar J, Parida M, Mohapatra T, et al. Characterization of haplotypes and single nucleotide polymorphisms associated with Gn1a for high grain number formation in rice plant. Genomics. 2020;112:2647–57.
8. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: the perils of multiple testing. Perspect Clin Res. 2016;7:106.
9. Ranganathan P, Pramesh CS, Buyse M. Common pitfalls in statistical analysis: “P” values, statistical significance and confidence intervals. Perspect Clin Res. 2015;6:116.
10. Gupta SK. The relevance of confidence interval and P-value in inferential statistics. Indian J Pharmacol. 2012;44:143–4.
11. Akobeng AK. Confidence intervals and p-values in clinical decision making. Acta Paediatr. 2008;97:1004–7.
12. du Prel J-B, Hommel G, Röhrig B, Blettner M. Confidence interval or p-value?: part 4 of a series on evaluation of scientific publications. Dtsch Arztebl Int. 2009;106:335–9.
13. Cabral HJ. Multiple comparisons procedures. Circulation. 2008;117:698–701.
14. Drachman D. Adjusting for multiple comparisons. J Clin Res Best Pract. 2012;8(7):1–3.
15. Moon KR, van Dijk D, Wang Z, Gigante S, Burkhardt DB, Chen WS, et al. Visualizing structure and transitions in high-dimensional biological data. Nat Biotechnol. 2019;37:1482–92.

16. Roweis ST, Saul LK. Nonlinear dimensionality reduction by locally linear embedding. Science. 2000;290:2323–6.
17. Tenenbaum JB, de Silva V, Langford JC. A global geometric framework for nonlinear dimensionality reduction. Science. 2000;290:2319–23.
18. Chen L, Buja A. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. J Am Stat Assoc. 2009;104(485):209–19.
19. Moon T, Stirling W. Mathematical methods and algorithms for signal processing. PAP/CDR edition. Upper Saddle River: Pearson; 1999.
20. van der Maaten L, Hinton G. Visualizing Data using t-SNE. J Mach Learn Res. 2008;9:2579–605.
21. Amir ED, Davis KL, Tadmor MD, Simonds EF, Levine JH, Bendall SC, et al. Vi SNE enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia. Nat Biotechnol. 2013;31:545–52.
22. Linderman GC, Rachh M, Hoskins JG, Steinerberger S, Kluger Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data [internet]. Nat Methods. 2019. [cited 2020 Sep 21]. Available from: https://pubmed.ncbi.nlm.nih.gov/30742040/
23. Mendelson A. Foreword. In: Gaster B, Howes L, Kaeli DR, Mistry P, Schaa D, editors. Heterogeneous Computing with OpenCL [Internet]. Boston: Morgan Kaufmann; 2012. p. 7–9. [cited 2020 Sep 21]. Available from: http://www.sciencedirect.com/science/article/pii/B9780123877666000487.
24. Rajasundaram D, Selbig J. More effort - more results: recent advances in integrative "omics" data analysis. Curr Opin Plant Biol. 2016;30:57–61.
25. Mei B, Wang Z. An efficient method to handle the "large p, small n" problem for genomewide association studies using Haseman-Elston regression. J Genet. 2016;95:847–52.
26. Kosorok MR, Ma S. Marginal asymptotics for the large "p", small "n" paradigm: with applications to microarray data. Ann Statist. 2007;35:1456–86.
27. Okut H. Bayesian regularized neural networks for small n big p data. In: Artificial neural networks - models and applications [internet]. IntechOpen; 2016. [cited 2020 Sep 22]; Available from: https://www.intechopen.com/books/artificial-neural-networks-models-and-applications/bayesian-regularized-neural-networks-for-small-n-big-p-data.
28. Brown BJ, Fearn T, Vannucci M. The choice of variables in multivariate regression: a non-conjugate Bayesian decision theory approach. Biometrika. 1999;86:635–48.
29. Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith A, et al. Bayesian factor regression models in the "large p, small n" paradigm. Bayesian Statist. 2003;7:733–42.
30. Gianola D, Okut H, Weigel KA, Rosa GJ. Predicting complex quantitative traits with Bayesian neural networks: a case study with Jersey cows and wheat. BMC Genet. 2011;12:87.
31. Okut H, Gianola D, Rosa GJM, Weigel KA. Prediction of body mass index in mice using dense molecular markers and a regularized neural network. Genet Res (Camb). 2011;93:189–201.
32. Okut H, Wu X-L, Rosa GJM, Bauck S, Woodward BW, Schnabel RD, et al. Predicting expected progeny difference for marbling score in Angus cattle using artificial neural networks and Bayesian regression models. Genet Sel Evol. 2013;45:34.
33. Hoerl AE, Kennard RW. Ridge regression: biased estimation for nonorthogonal problems. Technometrics. 1970;12:55–67.
34. Tibshirani R. Regression shrinkage and selection via the Lasso. J Royal Statist Soc Series B (Methodological). 1996;58:267–88.
35. Won S, Choi H, Park S, Lee J, Park C, Kwon S. Evaluation of penalized and nonpenalized methods for disease prediction with large-scale genetic data [internet]. Biomed Res Int. 2015: e605891. [cited 2020 Sep 22]. Available from: https://www.hindawi.com/journals/bmri/2015/605891/
36. Chang KN, Zhong S, Weirauch MT, Hon G, Pelizzola M, Li H, et al. Temporal transcriptional response to ethylene gas drives growth hormone cross-regulation in Arabidopsis. elife. 2013;2: e00675.

37. Thunders M, Cavanagh J, Li Y. De novo transcriptome assembly, functional annotation and differential gene expression analysis of juvenile and adult E. fetida, a model oligochaete used in ecotoxicological studies. Biol Res. 2017;50:7.
38. Atwal GS, Kinney JB. Learning quantitative sequence–function relationships from massively parallel experiments. J Stat Phys. 2016;162:1203–43.
39. Glick M, Klon AE, Acklin P, Davies JW. Enrichment of extremely noisy high-throughput screening data using a naïve Bayes classifier. J Biomol Screen. 2004;9:32–6.
40. Minoche AE, Dohm JC, Himmelbauer H. Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and genome Analyzer systems. Genome Biol. 2011;12: R112.
41. Fischer-Hwang I, Ochoa I, Weissman T, Hernaez M. Denoising of aligned genomic data. Sci Rep. 2019;9:15067.
42. Kinney JB, Atwal GS. Parametric inference in the large data limit using maximally informative models. Neural Comput. 2014;26:637–53.
43. Kinney JB, Tkačik G, Callan CG. Precise physical models of protein–DNA interaction from high-throughput data. PNAS. 2007;104:501–6.
44. Baralis E, Fiori A. Exploring heterogeneous biological data sources. In: 2008 19th international workshop on database and expert systems applications; 2008. p. 647–51.
45. Alyass A, Turcotte M, Meyre D. From big data analysis to personalized medicine for all: challenges and opportunities. BMC Med Genomics [Internet]. 2015;8. [cited 2020 Dec 9]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4482045/
46. Team RC. R: a language and environment for statistical computing, vol. 2014. Vienna: R Foundation for Statistical Computing; 2014.
47. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13:2498–504.
48. Heymann S. Gephi. In: Alhajj R, Rokne J, editors. Encyclopedia of social network analysis and mining [internet]. New York: Springer; 2018. p. 928–41. . [cited 2020 Dec 12]. https://doi.org/10.1007/978-1-4939-7131-2_299.
49. Urpa LM, Anders S. Focused multidimensional scaling: interactive visualization for exploration of high-dimensional data. BMC Bioinformatics. 2019;20:221.
50. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:180203426 [cs, stat] [Internet]. 2018. [cited 2019 Dec 31]; Available from: http://arxiv.org/abs/1802.03426
51. Hill CB, Czauderna T, Klapperstück M, Roessner U, Schreiber F. Metabolomics, standards, and metabolic Modeling for synthetic biology in plants. Front Bioeng Biotechnol [Internet]. 2015;3. [cited 2020 Sep 30]. Available from: https://www.frontiersin.org/articles/10.3389/fbioe.2015.00167/full
52. Morandini P. Rethinking metabolic control. Plant Sci. 2009;176:441–51.
53. Friedrich A, Kenar E, Kohlbacher O, Nahnsen S. Intuitive web-based experimental Design for High-Throughput Biomedical Data [internet]. Hindawi: BioMed Res Int; 2015. p. e958302. [cited 2020 Sep 30]. Available from: https://www.hindawi.com/journals/bmri/2015/958302/
54. Busetto AG, Hauser A, Krummenacher G, Sunnåker M, Dimopoulos S, Ong CS, et al. Near-optimal experimental design for model selection in systems biology. Bioinformatics. 2013;29:2625–32.
55. Khan IA, Fraser A, Bray M-A, Smith PJ, White NS, Carpenter AE, et al. ProtocolNavigator: emulation-based software for the design, documentation and reproduction biological experiments. Bioinformatics. 2014;30:3440–2.
56. Westfall PH, Young SS. Resampling-based multiple testing: examples and methods for p-value adjustment. 1st ed. New York: Wiley-Interscience; 1993.
57. Rudas T. Probability theory. In: Peterson P, Baker E, McGaw B, editors. International encyclopedia of education [internet]. 3rd ed. Oxford: Elsevier; 2010. p. 378–82. [cited 2020

Sep 23]. Available from: http://www.sciencedirect.com/science/article/pii/B9780080448947013592.

58. Nakajima T. Probability in biology: overview of a comprehensive theory of probability in living systems. Prog Biophys Mol Biol. 2013;113:67–79.
59. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge: Cambridge University Press; 1998.
60. Bishop MJ, Thompson EA. Maximum likelihood alignment of DNA sequences. J Mol Biol. 1986;190:159–65.
61. Churchill GA. Stochastic models for heterogeneous DNA sequences. Bull Math Biol. 1989;51:79–94.
62. Liu JS, Lawrence CE. Bayesian inference on biopolymer models. Bioinformatics. 1999;15:38–52.
63. Liu JS, Neuwald AF, Lawrence CE. Bayesian models for multiple local sequence alignment and Gibbs sampling strategies. J Am Stat Assoc. 1995;90:1156–70.
64. Zhou Q, Liu JS. Modeling within-motif dependence for transcription factor binding site predictions. Bioinformatics. 2004;20:909–16.
65. Narlikar L, Gordân R, Ohler U, Hartemink AJ. Informative priors based on transcription factor structural class improve de novo motif discovery. Bioinformatics. 2006;22:e384–92.
66. Schmidler SC, Liu JS, Brutlag DL. Bayesian segmentation of protein secondary structure. J Comput Biol. 2000;7:233–48.
67. Lunter G, Miklós I, Drummond A, Jensen JL, Hein J. Bayesian coestimation of phylogeny and sequence alignment. BMC Bioinformatics. 2005;6:83.
68. Boys RJ, Henderson DA, Wilkinson DJ. Detecting homogeneous segments in DNA sequences by using hidden Markov models. J Royal Statist Soc Series C. 2000;49:269–85.
69. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82:711–32.
70. Boys RJ, Henderson DA. A Bayesian approach to DNA sequence segmentation. Biometrics. 2004;60:573–81.
71. Green PJ, Mardia KV. Bayesian alignment using hierarchical models, with applications in protein bioinformatics. Biometrika. 2006;93:235–54.
72. Wilkinson DJ. Bayesian methods in bioinformatics and computational systems biology. Brief Bioinform. 2007;8:109–16.
73. Black M. Bayesian inference for gene expression and proteomics edited by Kim-Anh do, Peter Müller, Marina Vannucci. Int Stat Rev. 2007;75:433–4.
74. Jansen R, Yu H, Greenbaum D, Kluger Y, Krogan NJ, Chung S, et al. A Bayesian networks approach for predicting protein-protein interactions from genomic data. Science. 2003;302:449–53.
75. Nguyen CD, Gardiner KJ, Nguyen D, Cios KJ. Prediction of protein functions from protein interaction networks: a Naïve Bayes approach. In: Ho T-B, Zhou Z-H, editors. PRICAI 2008: trends in artificial intelligence. Berlin: Springer; 2008. p. 788–98.
76. Geng H, Lu T, Lin X, Liu Y, Yan F. Prediction of protein-protein interaction sites based on naive Bayes classifier [internet]. Biochem Res Int. 2015:e978193. [cited 2020 Sep 26]; Available from: https://www.hindawi.com/journals/bri/2015/978193/
77. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian networks to analyze expression data. J Comput Biol. 2000;7:601–20.
78. Kontoghiorghes EJ, editor. Handbook of parallel computing and statistics. 1st ed. Boca Raton: Chapman and Hall/CRC; 2005.
79. Mitrophanov AY, Borodovsky M. Statistical significance in biological sequence analysis. Brief Bioinform. 2006;7:2–24.
80. Vilardell M, Sánchez-Pla A. Hypothesis testing approaches to the exon prediction problem. Bioinformatics. 2006;22:3003–8.
81. Yates PD, Mukhopadhyay ND. An inferential framework for biological network hypothesis tests. BMC Bioinformatics. 2013;14:94.

82. Manda P, Freeman MG, Bridges SM, Jankun-Kelly T, Nanduri B, McCarthy FM, et al. GOModeler- a tool for hypothesis-testing of functional genomics datasets. BMC Bioinformatics. 2010;11:S29.
83. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014;15:504–18.
84. Pond SLK, Frost SDW, Muse SV. HyPhy: hypothesis testing using phylogenies. Bioinformatics. 2005;21:676–9.
85. Ge Y, Sealfon SC, Speed TP. Multiple testing and its applications to microarrays. Stat Methods Med Res. 2009;18:543–63.
86. Mieth B, Kloft M, Rodríguez JA, Sonnenburg S, Vobruba R, Morcillo-Suárez C, et al. Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. Sci Rep. 2016;6:36671.
87. Ramsköld D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. PLoS Comput Biol. 2009;5:e1000598.
88. Mou T, Deng W, Gu F, Pawitan Y, Vu TN. Reproducibility of methods to detect differentially expressed genes from single-cell RNA sequencing. Front Genet [Internet]. 2020;10. [cited 2020 Sep 27]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6979262/
89. Kim S, Schliekelman P. Prioritizing hypothesis tests for high throughput data. Bioinformatics. 2016;32:850–8.
90. Andreopoulos B, An A, Wang X, Schroeder M. A roadmap of clustering algorithms: finding a match for a biomedical application. Brief Bioinform. 2009;10:297–314.
91. Pirim H, Ekşioğlu B, Perkins A, Yüceer Ç. Clustering of high throughput gene expression data. Comput Oper Res. 2012;39:3046–61.
92. Oyelade J, Isewon I, Oladipupo F, Aromolaran O, Uwoghiren E, Ameh F, et al. Clustering algorithms: their application to gene expression data. Bioinform Biol Insights. 2016;10:237–53.
93. Chandrasekhar T, Thangavel K, Elayaraja E. Effective clustering algorithms for gene expression data. arXiv:12014914 [cs, q-bio] [Internet]. 2012. [cited 2020 Sep 27]; Available from: http://arxiv.org/abs/1201.4914
94. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. IEEE Trans Knowl Data Eng. 2004;16:1370–86.
95. Qin ZS. Clustering microarray gene expression data using weighted Chinese restaurant process. Bioinformatics. 2006;22:1988–97.
96. Yu H, Liu Z, Wang G. An automatic method to determine the number of clusters using decision-theoretic rough set. Int J Approx Reason. 2014;55:101–15.
97. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci U S A. 1999;96:2907–12.
98. Du Z, Wang Y, Ji Z. PK-means: a new algorithm for gene clustering. Comput Biol Chem. 2008;32:243–7.
99. Jin X, Han J. Partitional clustering. In: Sammut C, Webb GI, editors. Encyclopedia of machine learning [internet]. Boston: Springer; 2010. p. 766. . [cited 2020 Sep 28]. https://doi.org/10.1007/978-0-387-30164-8_631.
100. Kerr G, Ruskin HJ, Crane M, Doolan P. Techniques for clustering gene expression data. Comput Biol Med. 2008;38:283–93.
101. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326:289–93.
102. Serra F, Stefano MD, Spill YG, Cuartero Y, Goodstadt M, Baù D, et al. Restraint-based three-dimensional modeling of genomes and genomic domains. FEBS Lett. 2015;589:2987–95.
103. Shalon D, Smith SJ, Brown PO. A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. Genome Res. 1996;6:639–45.

104. He M, Lin Y, Xu Y. Identification of prognostic biomarkers in colorectal cancer using a long non-coding RNA-mediated competitive endogenous RNA network. Oncol Lett. 2019;17:2687–94.
105. Pereverzeva M, Murray SO. Luminance gradient configuration determines perceived lightness in a simple geometric illusion. Front Hum Neurosci [Internet]. 2014;8. [cited 2020 Sep 29]. Available from: https://www.frontiersin.org/articles/10.3389/fnhum.2014.00977/full
106. O'Donoghue SI, Baldi BF, Clark SJ, Darling AE, Hogan JM, Kaur S, et al. Visualization of biomedical data. Ann Rev Biomed Data Sci. 2018;1:275–304.
107. Merlotti A, Faria do Valle I, Castellani G, Remondini D. Statistical modelling of CG interdistance across multiple organisms. BMC Bioinformatics. 2018;19:355.
108. Özcan Şimşek NÖ, Özgür A, Gürgen F. Statistical representation models for mutation information within genomic data. BMC Bioinformatics. 2019;20:324.
109. Baghalian K, Hajirezaei M-R, Schreiber F. Plant metabolic Modeling: achieving new insight into metabolism and metabolic engineering. Plant Cell. 2014;26:3847–66.
110. Yonekura-Sakakibara K, Fukushima A, Saito K. Transcriptome data modeling for targeted plant metabolic engineering. Curr Opin Biotechnol. 2013;24:285–90.
111. Yuan JS, Galbraith DW, Dai SY, Griffin P, Stewart CN. Plant systems biology comes of age. Trends Plant Sci. 2008;13:165–71.
112. Papp B, Notebaart RA, Pál C. Systems-biology approaches for predicting genomic evolution. Nat Rev Genet. 2011;12:591–602.
113. Lee SY, Park JM, Kim TY. Chapter four - application of metabolic flux analysis in metabolic engineering. In: Voigt C, editor. Methods in enzymology [internet]. Academic Press; 2011. p. 67–93. [cited 2020 Sep 30]. Available from: http://www.sciencedirect.com/science/article/pii/B9780123851208000048.
114. Collakova E, Yen JY, Senger RS. Are we ready for genome-scale modeling in plants? Plant Sci. 2012;191–192:53–70.
115. Kruger NJ, Masakapalli SK, Ratcliffe RG. Strategies for investigating the plant metabolic network with steady-state metabolic flux analysis: lessons from an Arabidopsis cell culture and other systems. J Exp Bot. 2012;63:2309–23.
116. Sweetlove LJ, Fell D, Fernie AR. Getting to grips with the plant metabolic network. Biochem J. 2008;409:27–41.
117. Schallau K, Junker BH. Simulating plant metabolic pathways with enzyme-kinetic models. Plant Physiol. 2010;152:1763–71.
118. Good PI. Resampling methods: a practical guide to data analysis [internet]. 3rd ed. Basel: Birkhäuser; 2006. [cited 2020 Sep 30]. Available from: https://www.springer.com/gp/book/9780817643867
119. Fieberg JR, Vitense K, Johnson DH. Resampling-based methods for biologists. PeerJ [Internet]. 2020;8. [cited 2020 Oct 2]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7211410/
120. Molinaro AM, Simon R, Pfeiffer RM. Prediction error estimation: a comparison of resampling methods. Bioinformatics. 2005;21:3301–7.
121. Gana Dresen IM, Boes T, Huesing J, Neuhaeuser M, Joeckel K-H. New resampling method for evaluating stability of clusters. BMC Bioinformatics. 2008;9:42.
122. Zhang K, Zhao H. Assessing reliability of gene clusters from gene expression data. Funct Integr Genomics. 2000;1:156–73.
123. Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc Natl Acad Sci U S A. 2001;98:13790–5.
124. Kerr MK, Churchill GA. Bootstrapping cluster analysis: assessing the reliability of conclusions from microarray experiments. Proc Natl Acad Sci U S A. 2001;98:8961–5.
125. Dudoit S, Fridlyand J. Bagging to improve the accuracy of a clustering procedure. Bioinformatics. 2003;19:1090–9.
126. Ma'ayan A. Introduction to network analysis in systems biology. Sci Signal. 2011;4:tr5.

127. Watts DJ, Strogatz SH. Collective dynamics of “small-world” networks. Nature. 1998;393:440–2.
128. Barabási AL, Albert R. Emergence of scaling in random networks. Science. 1999;286:509–12.
129. Bringmann LF, Pe ML, Vissers N, Ceulemans E, Borsboom D, Vanpaemel W, et al. Assessing temporal emotion dynamics using networks. Assessment. 2016;23:425–35.
130. Hevey D. Network analysis: a brief overview and tutorial. Health Psychol Behav Med. 2018;6:301–28.
131. Gupta MK, Behara SK, Vadde R. In silico analysis of differential gene expressions in biliary stricture and hepatic carcinoma. Gene. 2017;597:49–58.
132. Gupta MK, Behera SK, Dehury B, Mahapatra N. Identification and characterization of differentially expressed genes from human microglial cell samples infected with Japanese encephalitis virus. J Vector Borne Dis. 2017;54:131–8.
133. Cordeddu V, Di Schiavi E, Pennacchio LA, Ma’ayan A, Sarkozy A, Fodale V, et al. Mutation in SHOC2 promotes aberrant protein N-myristoylation and underlies Noonan-like syndrome with loose anagen hair. Nat Genet. 2009;41:1022–6.
134. Berger SI, Posner JM, Ma’ayan A. Genes2Networks: connecting lists of gene symbols using mammalian protein interactions databases. BMC Bioinformatics. 2007;8:372.
135. Zaidel-Bar R, Itzkovitz S, Ma’ayan A, Iyengar R, Geiger B. Functional atlas of the integrin adhesome. Nat Cell Biol. 2007;9:858–67.
136. Oti M, Brunner HG. The modular nature of genetic diseases. Clin Genet. 2007;71:1–11.
137. Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L. The human disease network. PNAS National Academy of Sciences. 2007;104:8685–90.
138. Ideker T, Sharan R. Protein networks in disease. Genome Res. 2008;18:644–52.
139. Luo J, Liang S. Prioritization of potential candidate disease genes by topological similarity of protein–protein interaction network and phenotype data. J Biomed Inform. 2015;53:229–36.
140. Dezső Z, Nikolsky Y, Nikolskaya T, Miller J, Cherba D, Webb C, et al. Identifying disease-specific genes based on their topological significance in protein networks. BMC Syst Biol. 2009;3:36.
141. Wu X, Jiang R, Zhang MQ, Li S. Network-based global inference of human disease genes. Mol Syst Biol. 2008;4:189.
142. Li Y, Patra JC. Genome-wide inferring gene-phenotype relationship by walking on the heterogeneous network. Bioinformatics. 2010;26:1219–24.
143. Zhu J, Qin Y, Liu T, Wang J, Zheng X. Prioritization of candidate disease genes by topological similarity between disease and protein diffusion profiles. BMC Bioinformatics. 2013;14(Suppl 5):S5.
144. Köhler S, Bauer S, Horn D, Robinson PN. Walking the Interactome for prioritization of candidate disease genes. Am J Hum Genet. 2008;82:949–58.
145. Boitard S, Rodríguez W, Jay F, Mona S, Austerlitz F. Inferring population size history from large samples of genome-wide molecular data - An approximate Bayesian computation approach. PLoS Genet. 2016;12:e1005877.
146. Stuart JM, Segal E, Koller D, Kim SK. A gene-coexpression network for global discovery of conserved genetic modules. Science. 2003;302:249–55.
147. Prieto C, Risueño A, Fontanillo C, De las Rivas J. Human gene coexpression landscape: confident network derived from tissue transcriptomic profiles. PLoS One. 2008;3:e3911.
148. Wagner A, Fell DA. The small world inside large metabolic networks. Proc Biol Sci. 2001;268:1803–10.
149. Tanaka R. Scale-rich metabolic networks. Phys Rev Lett. 2005;94:168101.
150. Miele V, Matias C, Robin S, Dray S. Nine quick tips for analyzing network data. PLoS Comput Biol. 2019;15:e1007434.
151. Tsui IFL, Chari R, Buys TPH, Lam WL. Public databases and software for the pathway analysis of Cancer genomes. Cancer Inform. 2007;3:379–97.

Introduction of the Databases of Rice

3

Gayatri Gouda, S. Sabarinathan, Ravindra Donde,
Goutam Kumar Dash, Menaka Ponnana, Manoj Kumar Gupta,
Ramakrishna Vadde, Lambodar Behera, and Trilochan Mohapatra

Abstract

Rice is one of the world's most important agricultural crops and a model plant that is widely studied. The completion of the complete rice (*Oryza sativa*) genome sequence through high-throughput experimental platforms has resulted in a huge amount of data being generated and specialized databases, and bioinformatic tools for data processing, analysis, efficient organization, and visualization have been developed. In this chapter, we address a set of biological databases that host rice-specific sequence, genetic variation, gene expression, pathways, and gene–

S. Sabarinathan, Ravindra Donde, Goutam Kumar Dash, Menaka Ponnana and Manoj Kumar Gupta contributed equally with all other contributors.

G. Gouda (✉) · R. Donde · G. K. Dash · M. K. Gupta · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

T. Mohapatra
Secretary (DARE) and Director General (ICAR), Government of India, New Delhi, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_3

interactome data from various genomic and proteomic sources, and aid in data analysis and visualization.

Keywords

Database · Rice · Genomics · Proteomics · Gene expression

Abbreviations

BAR	Bio-Analytic Resource
CSRDB	Cereal Small RNA Database
DIPOS	Database of Interacting Proteins in *Oryza sativa.*
ESTs	Expressed Sequence Tags
FSTs	Flanking Sequence Tags
GEO	Gene Expression Omnibus
GRASSIUS	Grass Regulatory Information Services
IRGSP	International Rice Genome Sequencing Project
MCDRP	Manually Curated Database of Rice Proteins
MPIC	Mitochondrial Protein Import Components
PDB	Protein Data Bank
PLANEX	PLAnt co-EXpression Database
PlantDHS	Plant DNaseI Hypersensitive Site Database
PmiRKB	Plant miRNA Knowledge Base
PMRD	Plant microRNA Database
PRIN	Predicted Rice Interactome Network
RAP-DB	Rice Annotation Project Database
RED	Rice Expression Database
RGAP	Rice Genome Annotation Project
RICD	Rice Indica cDNA Database
RiceSTIFDB	Stress-Responsive Transcription Factor Database
RiceVarMap	Rice Variation Map
RKD	Rice Kinase Database
ROAD	Rice Oligonucleotide Array Database
SNPs	single nucleotide polymorphisms
TIGR	The Institute for Genomic Research

3.1 Introduction

Rice (*Oryza sativa* L.) is an important crop from Gramene family because most of the human population mainly depend on rice as their food [1]. However, recent climate change has resulted in drastic decrease in the yield and has also affected the rice genome at molecular level. Previously, it was reported that all the agronomic traits are controlled by various genes and considering that, to date, numerous genes and its sequences have been identified and used in the breeding program to improve the yield and quality of rice [2]. Additionally, since rice has a small genome size, it is often used as a model organism for the monocot plant species. Thus, there is always a quest to unmask the genetic structure and composition of rice genome. Though the recent advancement of high-throughput sequencing technologies has been able to solve these issues, they had led to the accumulation of huge amount of biological data. This in turn demands the development of various bioinformatic databases and tools, which may store and analyze data efficiently. Considering this, to date many rice-specific databases are developed that contain heterogeneous biological data ranging from genomics to proteomics and are regularly exploited to study and evaluate numerous agricultural-related information such as gene mapping, gene identification, functional characterization, gene expression study, structure prediction, protein structure, and function prediction. Thus, in this chapter authors have described in brief about various rice databases and information they contain.

3.2 Genomic Database

Gramene is a comparative genomic database of plants that incorporates knowledge regarding genetic charts, sequences, gene markers, proteins, pathways, and phenotypes [3]. One may search or explore the database to identify genes and phenotypes that share similar characteristics. Different plant organisms may also be correlated and differentiated considering the identical genes, genomes, mechanisms, and phenotypes. From the Gramene database, one may also know the position of genes in the chromosome and its function [4]. Using the genome browser module of the Gramene database, the annotation of rice genome is possible, which provides information on SNPs, indels, and markers for the respective genes. The rice genome also acts as the reference for genome-wide comparative study. In order to decide whether rice plants are linked to other cereals, the researcher uses a number of sequence-based tags including the sequences of expressed sequence tags (ESTs), flanking sequence tags (FSTs), proteins, cDNAs, BACs and BAC ends, SNPs, microarray probes, and repeat sequences [5].

The gene archive takes advantage of the genetic colinearity (synteny) between rice and other main crop plant genomes to annotate genome of organisms whose own plants have not been sequenced [5]. For performing annotation, Gramene employs several modules like BLAST tool, marker module, and QTL modules. BLAST tool aligns the query sequence with the reference genome available for the particular crop with many species. For *Oryza,* the database stored information on 11 wild species of

Oryza such as *Oryza barthii, Oryza glaberrima, Oryza glumipatula, Oryza longistaminata, Oryza punctate, Oryza rufipogan, Oryza brachyantha, Oryza sativa indica, Oryza meridionalis, Oryza sativa japonica, and Oryza sativa f spontanea.* The BLAST alignment results provide information regarding the details of the alignment gene with E-value and identity percentage of gene with reference and the orientation of the gene in either positive or negative strand. BLAST has been discussed in detail in Chap. 7 of the book. QTL module provides information about QTLs of rice and several other cereals (*described below in QTL section*). Marker module provides information on markers available for all the genes/QTLs of the crops. The detailed information about microsatellite or SSR markers of rice genome with mapped chromosome position with location is also available, which in turn enables us to easily identify the gene and their nearest gene location. These data provide a gateway to find out the marker position of a particular gene with respect to various traits, like abiotic and biotic traits in rice.

Another widely used rice genomic database is “Rice Annotation Project Database (RAP-DB)” [6, 7]. The information about the rice genome is generated through the “International Rice Genome Sequencing Project (IRGSP)” and is stored for public use in RAP-DB [6, 7]. In this database, the Japanese rice Nipponbare (build IRGSP 1.0) is considered as a reference genome for all the rice genotypes. The genome sequencing was carried out through high-throughput Illumina sequencing platform. RicyerDB is another rice genomic database that contains information regarding the yield-related genes of rice. Both genomic and proteomic data are available in this database, which make researcher to easily get the gene information and the pathways involved for gene expression in rice [8]. The tools of RicyerDB provide base for the development of a shared platform for browsing and visualizations of yield-related genes. Another utility allows the user to search for a single gene and offers insight into the roles and positions of biological processes. The study of protein–protein interaction and protein–protein interaction network construct is also possible through this database. This database also stores information from various sources such as RAP-DB, Rice Genome Annotation Project (RGAP), NCBI, and UniProt and STRING databases.

RiceGAAS database is also known as “Rice Genome Automated Annotation System.” It allows to execute the genomic data of rice for public use [21]. The data are collected from various sources such as IRGSP [70] and submitted to the public domain database DDBJ (https://www.ddbj.nig.ac.jp/services-e.html) and provided all the information regarding Gene entry, gene homology identification, prediction of long terminal repeats, and gene models. The RiceRelativesGD is another gene database aimed at providing a genetic resource useful for rice breeding. In 2019, Mao et al. [71] developed this database for identifying the genetic information of close rice relative species. This database contains genomic information from 13 separate rice relatives such as *O. sativa* (*japonica* group), *O. sativa* (*indica* group), *O. rufipogon, O. barthii, O. glumaepatula, O. meridionalis, O. nivara, O. punctata, O. brachyantha, Leersia perrieri, O. glaberrima, Zizania latifolia, and Echinochloa crus-galli* which are accessible to the public. Their study provides knowledge on genes specific for different functions such as stress and photosynthesis that are used in breeding program.

3.3 QTL/Gene Database

QTL databases are used to identify the gene function that are used in breeding program [28]. For instance, Q-TARO database provides detailed information of rice QTLs. One important feature is that it enlists a table comprising all of the QTLs and their genetic parameters such as their trait or trait type, population, mapping, accuracy (LOD value), and map location of the QTL. Another important feature is the genome browser displaying genomic positions of QTLs. Q-TARO also specifically displays the colinear spatial structures of QTLs and QTL regions on the rice genome [27].

For plants, Gramene is another quantitative information resource that incorporates data through different data domains. In Gramene QTL database, QTLs are identified as a part of gene, which is associated with a particular phenotype. The QTL database includes the world's largest online repository of QTL data for rice. QTLs initially published on individual genetic maps have been systematically aligned to the rice sequence using flanking markers as anchors, where they can be searched as normal genomic features. It enables the analysis of QTLs in colinear regions in other cereals and allows researchers, to distinguish sequences and QTLs correlated with related traits or phenotypes across a broad variety of plant species. Researchers can identify whether a QTL colocalizes with other QTLs and can integrate data from different studies to enhance the accuracy of a QTL location. It provides plant biologists and geneticists a way to investigate the interaction between genomic variation and diverse modes of phenotypic variation [28].

To improve QTL-based candidate gene recognition and gene expression study, PlantQTL-GE database has developed [29]. This database contains information about chromosomes and details on gene expression in microarray data and ESTs and genetic markers of *Arabidopsis thaliana* and *Oryza sativa.* Another database, namely the Institute for Genomic Research (TIGR) database, contains DNA, RNA, and protein sequence of plant, human, and microbes. This database consists of repetitive DNA sequence of 12 plant species, namely *Arabidopsis, Hordeum, Brassica, Glycine, Lotus, Oryza, Triticum, Lycopersicon, Medicago, Sorghum, Solanum, and Zea*. The repeated sequences within each database are further classified into subcategories namely groups and subclasses dependent on sequence and structure similarities [72]. Sequence similarity can also be checked for the downloaded files and are accessible from different sources [73].

3.4 Single Nucleotide Polymorphism (SNP) Database

SNPs, which comprise the most abundant type of genetic variation, are used in genetic studies [74]. SNPs play key role while studying gene mapping, diversity, and evolutionary variation among populations. They also play an important role in designing markers to identify the genetic variation occurring in the contemporary genome and that it has been transferred from the wild type. While other forms of variation including indels, microsatellites, variation of copy number, and epigenetic markers remain important to consider and can cause disease, in genetic study, SNPs are largely the simplest to determine, and are the most useful and commonly used markers. One of the most important SNP databases is dbSNP (http://www.ncbi.nlm.nih.gov/SNP), in which SNP identifiers (SNPids) or rsIDs are used to identify SNPs. This database arranges the nucleotide sequence based on their variation in sequence to differentiate between two sequences and search the polymorphism through nucleotide substitution, insertion, deletions, and nucleotide repeats [75].

By the sequencing project of 3000 rice genome, it has been possible to identify SNPs through the alignment of gene of interest with the reference genome in rice [76, 77]. This result provides the information of SNPs that are synonymous or nonsynonymous to the particular genome. The 3000-rice genome project has been described in detail in Chap. 5. Rice Variation Map database is developed for genomic variation study [30]. Yan et al. [78] developed a database of SNP for rice, namely IC4R, that provides SNP information of 18 billion reads. For commercial rice verities, they provide SNP barcode to easily assess the SNPs using seven machine learning-based methods that are DT, KNN, NB, ANN, RF, LR-M, and LR-O algorithms in the Python Sklearn Library (https://scikit-learn.org/stable/). To identify the genetic variation during various stress conditions, Rice Stress-Resistant SNP (RSRS) database has been developed [35]. Recently, Alexandrov and their team developed the SNP-Seek Database using 60 billion SNP reads, which provide SNP information by discarding all the indels in any of the genomic region to find out the structural genetic variations [33, 79].

3.5 Transcriptome Database

With the introduction of high-throughput technology like next-generation sequencing platforms, it is possible to identify the gene expression at the molecular level. Data on gene expression have proven invaluable to genome annotation programs [80]. The functional genomic study helps the researcher to find out the function of rice genes that control various traits. To study these gene functions, several databases are developed through bioinformatic tool where all the repository information of genes with specific trait of interest is available. For instance, TENOR (Transcriptome ENcyclopedia of Rice) database comprises of large-scale mRNA sequencing (mRNA-Seq) data extracted from rice. This database includes details on rice transcriptomes, such as transcript structures and expression profiles, as well as data on coexpression and data on cis-regulatory elements for each gene in 1 kb upstream regions. Since specifying the ability of plants to adapt to different growing conditions is a key issue in plant science, understanding the regulatory networks of genes associated with environmental changes is of great interest [81]. The team developed the database by using mRNA-Seq data under 10 abiotic stress condition such as high, low, and extremely low cadmium; high and low phosphate; high salinity; drought; osmotic; cold; and flood and under two conditions of plant hormone treatments (jasmonic acid and ABA). Earlier, Oono et al. used the TENOR database to identify the transporters for cadmium tolerance to study the gene expression under cadmium concentration [40]. TENOR offers three kinds of search systems. First, under various conditions, users can provide one or more transcript IDs for scanning functional annotations and patterns of expression. In addition, users can search genes for functional annotation keywords as well. This shows both partially and completely matched results. Second, the "GBrowse" genome browser helps users to check for transcripts with a transcript ID or genomic coordinate. In this search, the user can get all the information regarding the novel transcript structures viz. both annotated and unannotated with their characteristics. In the third search, the user can search for a collection of plant stress hormone-responsive genes through reactive expression patterns, defining the path of transition (suppression or induction), experimental circumstances, sampled tissues, and time points, in addition to fold change (FC) thresholds and statistical importance of changes in expression level.

Rice Expression Database (RED) is interactive rice gene expression profile database completely extracted from RNA-Seq data. RED provides a detailed list of 284 high-quality RNA-Seq results, includes a wide range of gene expression models, and encompasses a wide range of plant development stages. RED consists of a collection of genes unique to housekeeping and tissue and creates coexpression networks dynamically for gene(s) of interest. RiceArrayNet is another database that provides information in terms of correlation coefficients on coexpression between genes in rice. The correlation coefficient shows the coexpression pattern of genes in the rice genome [37]. Lee and the team developed this database that provides the correlation data in three different ways: First, gene coexpression is visualized in the form of cluster or network; second, the coexpression is visualized in scatter plot; and third, the gene coexpression is visualized in the histogram. Another recent Internet-based database for plant gene analysis is the "PLAnt co-EXpression database (PLANEX)." It includes freely accessible GeneChip details collected from the "Gene Expression Omnibus (GEO)" of NCBI. PLANEX is a database for genome-wide coexpression, which helps genes from a wide range of experimental designs to be functionally identified [42]. PLANEX describes "Pearson's correlation coefficients (PCCs; r-values)" distributed for a specified microarray platform, contributing to a single organism from a gene of interest. The PLANEX database offers a correlation database, a cluster network, and an analysis of enrichment test results for eight plant organisms such as *Arabidopsis thaliana, Triticum aestivum, Glycine max, Vitis vinifera, Hordeum vulgare, Solanum lycopersicum, Oryza sativa, and Zea mays*. The cluster network of coexpressed genes is developed by PLANEX, which is calculated using the k-mean method. Genevestigator is another advanced

Web-based framework developed to use modern data mining concepts and groundbreaking algorithms to conduct molecular expression meta-analysis. This database is focused on the systematic, large-scale combination of normalized and quality-controlled expression data with ontology-based experimental background variables such as anatomy, development, perturbation, or genetic background This large-scale combination of data and meta-data gives new insight into transcriptomes' spatio-temporal response design and helps users to answer concerns that cannot be answered by evaluating a single experiment [50]. Other important gene expression databases are discussed in Table 3.1.

3.6 Protein Database

The protein database provides all the information of the gene that will encode protein. For functional genomics, proteome study related to genome sequence data is useful. These genome sequences help researcher to identify the genes that are expressed in protein developed through alternate splicing and post-transcriptional modification. Rice proteome study is possible through leaves, embryos, endosperms, roots, branches, shoots, and calluses, which provide the detailed mutation of gene through various environmental conditions. Many databases have been developed for proteome study in rice. For instance, OryzaPG-DB, a shotgun proteogenomic-dependent rice proteome database, integrates the genomic features of data from experimental shotgun proteomics. This version of the database was developed from the results of 27 nanoLC-MS/MS runs on a mass spectrometer of hybrid ion trap–orbitrap, providing high precision for the study of tryptic digests from undifferentiated cultured rice cells. Through searching the product ion spectra against the Michigan State University, protein, cDNA, transcript, and gene databases, peptides were detected and mapped to the rice genome. These peptides were occupied by approximately 3200 genes, and 40 of them incorporated novel genomic characteristics. The users may search, import, or browse the chromosome, gene, protein, cDNA, or transcript database and download the modified annotations in standard GFF3 format with PNG format visualization.

The "Mitochondrial Protein Import Components (MPIC)" database contains searchable details on the plant and non-plant mitochondrial protein import apparatus. An in silico study was performed to compare the mitochondrial protein import apparatus of 24 organisms representing different lines of *Saccharomyces cerevisiae*, algae to *Homo sapiens* and plants, including *Oryza sativa, Arabidopsis thaliana*, and other newly sequenced plant species. In the MPIC DB, each of these species has been thoroughly scanned and manually constructed for analysis. The database provides a user-friendly graphical map, enabling users to find their appropriate import component. The MPIC DB offers a robust database to promote thorough investigation of mitochondrial protein import machinery and to identify conservation and divergence patterns that might have been skipped [64].

Manually Curated Database of Rice Proteins (MCDRP) is a database for rice protein mainly focused on the reported experimental data [68]. Another database,

Table 3.1 Description of various database used in rice research

Database type	Database Name	Function	Reference
Genomic	Gramene	Provides tools to perform powerful comparative analyses through a wide range of species of plants.	[3]
	Phytozome	Provides an overview of the evolution of each plant gene at the sequence level, gene function, gene family, and organization of the genome	[9]
	Ensemble plant	Provides information on visualizing, mining, and analyzing plant genome	[10]
	RAP-DB	A hub for rice genomics.	[11]
	RGKbase	Evolutionary biology and comparative genomics	[12]
	GreenPhylDB	Comparative and functional genomics in plants	[13]
	PlantGDB	Browsing features of genomes that combine all relevant EST and cDNA data for emerging gene models	[14]
	PGDBJ	Provides a method for plant genome annotation.	[15]
	MIPS PlantsDB	Provides central data integration network for annotation of structural and functional gene	[16]
	OryGenesDB	Provides positional information of flanking sequence tags (FSTs) of insertional mutagens	[17]
	PGSB	Provides keyword search for gene identifier and functional gene description of plant	[18]
	RMD	For new gene identification and regulatory element	[19]
	Rice Phylogenomics database	Functional study of rice genome	[20]
	RiceGAAS	For executing reliable analysis from sequence of rice genome	[21]
	RAD	Provides the new information on manual annotation and a systematic structural study of rice genome	[22]
	RGKbase	For genome assembly and annotation	[12]
	Ricebase	Phenotypic interpretation, gene annotation	[23]

(continued)

Table 3.1 (continued)

Database type	Database Name	Function	Reference
	OryzaGenome	Provides the variant information with hyperlinks to Oryzabase	[24]
	DroughtDB	Facilitates the recognition, study, and characterization of genes associated with the resistance of drought stress	[25]
	RiceMetaSys	Information on genes for salt and drought stress	[26]
QTL	Q-TARO	Provides all the information of rice QTLs	[27]
	Gramene	Provides information on genetic and physical position of QTL	[28]
	PlantQTL-GE	QTL-based gene identification	[29]
SNP	Rice variation map (RiceVarMap)	Provides comprehensive information of SNPs and indels	[30]
	PmiRKB (Plant miRNA Knowledge Base)	Investigate the metabolism of miRNA precursors and gene regulation	[31]
	RICD (Rice Indica cDNA database)	Provides cDNA resource with comprehensive information for comparative genomics. And functional analysis of *indic*a subspecies	[32]
	Rice SNP-seek database	Quick retrieving of SNP alleles for all varieties in a given genome region	[33]
	HapRice	Identification of polymorphic SNPs between any two accessions to rice	[34]
	Rice stress-resistant SNP database	Characterization of SNPs for various stresses	[35]
Expression	Rice expression database (RED)	Information on important biological processes and mechanism	[36]
	RiceArrayNet	Provides information in form of correlation coefficients on coexpression between genes	[37]
	RiceXPro	Provides information on gene expression of important agronomic trait in rice	[38]
	Rice oligonucleotide Array database (ROAD)	Provides detailed profiles of gene expression for all rice genes	[39]
	TENOR	Expression study in various environmental stress	[40]
	OryzaExpress	Gene expression network and omic annotation using large-scale microarray data	[41]

(continued)

Table 3.1 (continued)

Database type	Database Name	Function	Reference
	PLANEX	For functional identification of genes	[42]
	CoP	Identifies the association of coexpressed genes with metabolic pathways	[43]
	PlaNet	Predicts the identity of functional homologs.	[44]
	ExPath	Pathway enrichment study	[45]
	Plant microRNA database (PMRD)	To study the function and their target genes of miRNAs	[46]
	ATTED-II	For functional identification and study of regulatory relationships among genes.	[47]
	IsomiR Bank database	Provides target prediction and enrichment analysis to evaluate the effects of isomiRs	Zhang et al. [48]
	Rice kinase database (RKD)	Expression analysis for rice kinases using NGS data	[49]
	Genevestigator	Gene expression analysis using microarray data	[50]
	Bio-analytic resource (BAR)	Transcriptome and protein–protein interaction study	[51]
	PlantTFDB	Provides information of transcription factors	[52]
	CSRDB (Cereal Small RNA Database)	A tool of smRNA to identify single mature transcript for the recognition of specific target sites	[53]
	PtRFdb	Characterization and validation of plant tRFs	[54]
	RiceSTIFDB (stress-responsive transcription factor database)	Accelerates study of rice TFs in functional genomics and identifies the regulatory processes underlying abiotic stress responses	[55]
	GRASSIUS (grass regulatory information services)	Integrates TFs and gene promoter information	[56]
	Plant MPSS (massively parallel signature sequencing) databases	Measure the degree of expression of the majority of genes under specified conditions	[57]
	PlantDHS (plant DNaseI hypersensitive site database)	Detection of distant cis-regulatory DNA elements (CREs) that are located away from promoters	[58]

(continued)

Table 3.1 (continued)

Database type	Database Name	Function	Reference
	PlantAPA (alternative polyadenylation)	Alternative modulation of polyadenylation and gene expression in plants	[59]
	RiceFREND	Provides a platform for gene function prediction	[60]
	RiceSRTFDB	Accelerates the study of rice TFs in functional genomics and identifies the regulatory processes underlying responses to abiotic stress	[55]
	RECoN	Characterization of abiotic stresses on genome-scale and regulatory mechanism study	[61]
	RTFDB	Identification of gene expression related to stress which is generally tissue-specific	[62]
Protein	UniProtKB	The measurable properties of each protein, like isoelectric point, molecular weight, and expression and experimentally defined properties like sequences of amino acids	[63]
	PDB (protein data Bank)	Collection of protein 3D structure	http://www.wwpdb.org/
	MPIC (mitochondrial protein import components) database	Comparative study of machinery for importing mitochondrial protein from plants and animals	[64]
	OryzaPG-DB (Oryza ProteoGenomic Database)	Provides models of peptide-based expression along with the corresponding genomic origin, including novelty annotation for each peptide	[65]
	DIPOS (database of Interacting Proteins in *Oryza sativa*)	Provides comprehensive information of rice proteins	[66]
	PRIN (predicted Rice Interactome network)	Provides novel insights into gene networks and functional coordination of genes	[67]
	MCDRP (manually Curated database of rice proteins)	Annotation of rice protein	[68]

(continued)

Table 3.1 (continued)

Database type	Database Name	Function	Reference
Gene ontology and pathway	Planteome	It provides a set of reference and species-specific ontologies, gene, and phenotype annotation for plants.	[69]
	RiceCyc	Provides information about the substrates, enzymes, metabolites, reactions, and pathways of primary as well as intermediary metabolism within rice.	https://bigd.big.ac.cn/databasecommons/database/id/3305
	KEGG rice	Provides information about rice-associated pathway	https://www.genome.jp/kegg-bin/show_organism?org=dosa
	IntAct rice	Provides a publicly available, open-source information platform, and protein interaction data analysis tools.	https://www.ebi.ac.uk/intact/home.xhtml

namely the “Database of Interacting Proteins in Oryza sativa (DIPOS),” offers detailed knowledge on interacting proteins in rice, where two statistical approaches are used to model interactions, i.e., interologs and domain-based methods. Of 27,746 proteins, DIPOS comprises 14,614,067 pair-related associations, covering around 41 percent of the entire proteome of *Oryza sativa* [66].

3.7 Gene Ontology and Pathway Database

To broaden our understanding of biological processes in plants and to explain how biological functions develop, it is important to recognize certain diversified pathway. An application of reference and species-specific ontologies for plants and annotations to genes and phenotypes are provided by the Planteome project (http://www.planteome.org). Ontologies serve as common standards for the semantic integration of data from plant genomics, phenomics, and genetics from a large and growing dataset. There Plant Ontology, Plant Trait Ontology, and Plant Experimental Conditions Ontology, developed by the Planteome project, together with Gene Ontology, Biological Interest Chemical Entities, Phenotype and Attribute Ontology, and others, are the reference ontologies [69]. The platform also offers access to species-specific crop ontologies established around the world by different plant breeding and research communities. Out of 95 plant taxa, annotated with reference ontology terms, developers offer integrated data on plant traits, phenotypes, and gene function and expression. In order to facilitate community engagement, the Planteome project has developed a plant gene annotation platform, Planteome Noctua. All Planteome ontologies are freely accessible and are managed for sharing and monitoring revisions and new queries at the Planteome GitHub platform (https://

github.com/Planteome). From the ontology browser (http://browser.planteome.org/amigo) and our data archive, the annotated data are readily available.

RiceCyc is another directory of rice metabolic pathway network [82]. It is a glimpse of the main and intermediate metabolism of substrates, enzymes, metabolites, reactions, and pathways in rice. Version 3.3 of RiceCyc contains 316 pathways and 6643 peptide-coding genes mapped to 2103 enzyme-catalyzed and 87 transport reactions regulated by protein. Annotations given by the KEGG and Gramene databases enriched the original functional annotations of rice genes with InterPro, Enzyme Commission (EC) numbers, MetaCyc, and Gene Ontology. Employing the Pathologic module of Pathway Software, pathway inferences and network diagrams were first predicted based on MetaCyc reference networks and plant pathways from the Plant Metabolic Network. This was enriched by manually inserting metabolic pathways and explicitly reported gene functions for rice. From pathway diagrams to the relevant genes, metabolites, and chemical structures, the RiceCyc database is hierarchically browsable. Users may also upload transcriptomic, proteomic, and metabolomic data to visualize expression trends in a simulated cell using the OMIC Viewer integrated application. RiceCyc enables comparative pathway research, coupled with additional species-specific pathway databases housed in the Gramene project [82]. Another database, namely IntAct, offers a publicly distributed, open-source information structure and molecular interaction data review. All interactions are produced through curation of literature or direct contributions from users and are publicly accessible (https://www.ebi.ac.uk/intact/home.xhtml).

3.8 Conclusion and Future Perspectives

In conclusion, we attempted to catalog numerous Web data services and tools available for rice research. Some of them, although few are recent and small-scale repositories, are well known and commonly utilized. It is obvious with the growing number of databases that there is an overwhelming amount of data accessible on the site, connected to almost any area of rice science. However, it has not been effectively investigated despite possessing such a vast amount of diverse data, since many biology researchers or prospective consumers are unfamiliar with all the possible tools to find and interpret the data [83]. Different databases often have different data exchange formats and protocols, which makes it difficult to integrate them into one place. In an ideal situation, a single platform should be available for all databases in a single domain of interest, where a user can use APIs and ontologies to search all the respective databases with a single query and compare the results; e.g., Araport [84] is one such initiative. In order to improve the legitimacy of their data, certain databases are now merging connections to other databases of similar types of data, which is the first move in offering a single forum. This maximizes the utilization of usable data in current assets, which may aid in the prevention of duplication. It provides small databases with greater coverage and may offer a broader image collectively, as small databases typically concentrate on one particular element and provide comprehensive details.

Conflicts of Interest None.

References

1. Wang J, Xu H, Li N, Fan F, Wang L, Zhu Y, et al. Artificial Selection of Gn1a Plays an Important role in Improving Rice Yields Across Different Ecological Regions. Rice [Internet]. 2015;8(1):37. Dec [cited 2019 Feb 7]. Available from: http://www.thericejournal.com/content/8/1/37
2. Zhang Q-J, Zhu T, Xia E-H, Shi C, Liu Y-L, Zhang Y, et al. Rapid diversification of five Oryza AA genomes associated with rice adaptation. Proc Natl Acad Sci. 2014 Nov 18;111(46): E4954–62.
3. Ware DH, Jaiswal P, Ni J, Yap IV, Pan X, Clark KY, et al. Gramene, a tool for grass genomics. Plant Physiol. 2002 Dec 1;130(4):1606–13.
4. Youens-Clark K, Buckler E, Casstevens T, Chen C, DeClerck G, Derwent P, et al. Gramene database in 2010: updates and extensions. Nucleic Acids Res. 2011 Jan 1;39(suppl_1): D1085–94.
5. Jaiswal P, Ni J, Yap I, Ware D, Spooner W, Youens-Clark K, et al. Gramene: a bird's eye view of cereal genomes. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D717–23.
6. Sasaki T. The map-based sequence of the rice genome. Nature. 2005 Aug;436(7052):793–800.
7. Project RA. The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res. 2008 Jan 1;36(suppl_1):D1028–33.
8. Jiang J, Xing F, Zeng X, Zou Q, Ricyer DB. A database for collecting Rice yield-related genes with biological analysis. Int J Biol Sci. 2018 May 22;14(8):965–70.
9. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, et al. Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res. 2012 Jan;40(Database issue):D1178–86.
10. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomics data. Methods Mol Biol Clifton NJ. 2016;1374:115–40.
11. Ohyanagi H, Tanaka T, Sakai H, Shigemoto Y, Yamaguchi K, Habara T, et al. The Rice Annotation Project Database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D741–4.
12. Wang D, Xia Y, Li X, Hou L, Yu J. The Rice genome knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology. Nucleic Acids Res. 2013 Jan;41(Database issue):D1199–205.
13. Conte MG, Gaillard S, Lanau N, Rouard M, Périn C. GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Res. 2008 Jan;36(Database issue):D991–8.
14. Dong Q, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D354–9.
15. Asamizu E, Ichihara H, Nakaya A, Nakamura Y, Hirakawa H, Ishii T, et al. Plant genome DataBase Japan (PGDBj): a portal website for the integration of plant genome-related databases. Plant Cell Physiol. 2014 Jan 1;55(1):e8.
16. Nussbaumer T, Martis MM, Roessner SK, Pfeifer M, Bader KC, Sharma S, et al. MIPS PlantsDB: a database framework for comparative plant genome research. Nucleic Acids Res. 2013 Jan 1;41(D1):D1144–51.
17. Droc G, Ruiz M, Larmande P, Pereira A, Piffanelli P, Morel JB, et al. OryGenesDB: a database for rice reverse genetics. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D736–40.
18. Spannagl M, Nussbaumer T, Bader KC, Martis MM, Seidel M, Kugler KG, et al. PGSB PlantsDB: updates to the database framework for comparative plant genome research. Nucleic Acids Res. 2016 Jan 4;44(D1):D1141–7.
19. Zhang J, Li C, Wu C, Xiong L, Chen G, Zhang Q, et al. RMD: a rice mutant database for functional analysis of the rice genome. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D745–8.

20. Jung K-H, Cao P, Sharma R, Jain R, Ronald PC. Phylogenomics databases for facilitating functional genomics in rice. Rice. 2015 Jul 30;8(1):26.
21. Sakata K, Nagamura Y, Numa H, Antonio BA, Nagasaki H, Idonuma A, et al. RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res. 2002 Jan 1;30(1):98–102.
22. Ito Y, Arikawa K, Antonio BA, Ohta I, Naito S, Mukai Y, et al. Rice annotation database (RAD): a contig-oriented database for map-based rice genomics. Nucleic Acids Res. 2005 Jan 1;33(Database issue):D651–5.
23. Edwards JD, Baldo AM, Mueller LA. Ricebase: a breeding and genetics platform for rice, integrating individual molecular markers, pedigrees and whole-genome-based data. Database [Internet]. 2016 Jan 1;2016(baw107). [cited 2021 Jan 8] https://doi.org/10.1093/database/baw107.
24. Ohyanagi H, Ebata T, Huang X, Gong H, Fujita M, Mochizuki T, et al. OryzaGenome: genome diversity database of wild Oryza species. Plant Cell Physiol. 2016 Jan 1;57(1):e1.
25. Alter S, Bader KC, Spannagl M, Wang Y, Bauer E, Schön C-C, et al. DroughtDB: an expert-curated compilation of plant drought stress genes and their homologs in nine species. Database [Internet]. 2015 Jan1;2015(bav046). [cited 2021 Jan 8] https://doi.org/10.1093/database/bav046.
26. Sandhu M, Sureshkumar V, Prakash C, Dixit R, Solanke AU, Sharma TR, et al. RiceMetaSys for salt and drought stress responsive genes in rice: a web interface for crop improvement. BMC Bioinformatics. 2017 Sep 30;18(1):432.
27. Yonemaru J, Yamamoto T, Fukuoka S, Uga Y, Hori K, Yano M. Q-TARO: QTL annotation Rice online database. Rice. 2010 Sep;3(2):194–203.
28. Ni J, Pujar A, Youens-Clark K, Yap I, Jaiswal P, Tecle I, et al. Gramene QTL database: development, content and applications. Database [Internet]. 2009 Jan 1;2009(bap005). [cited 2021 Jan 5] https://doi.org/10.1093/database/bap005.
29. Zeng H, Luo L, Zhang W, Zhou J, Li Z, Liu H, et al. PlantQTL-GE: a database system for identifying candidate genes in rice and Arabidopsis by gene expression and QTL information. Nucleic Acids Res. 2007 Jan 1;35(suppl_1):D879–82.
30. Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res. 2015 Jan 28;43(D1):D1018–22.
31. Meng Y, Gou L, Chen D, Mao C, Jin Y, Wu P, et al. PmiRKB: a plant microRNA knowledge base. Nucleic Acids Res. 2011 Jan 1;39(suppl_1):D181–7.
32. Lu T, Huang X, Zhu C, Huang T, Zhao Q, Xie K, et al. RICD: a rice indica cDNA database resource for rice functional genomics. BMC Plant Biol. 2008 Nov 26;8(1):118.
33. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 2015 Jan 28;43(D1):D1023–7.
34. Yonemaru J, Ebana K, Yano M. HapRice, an SNP haplotype database and a web tool for Rice. Plant Cell Physiol. 2014 Jan 1;55(1):e9.
35. Tareke Woldegiorgis S, Wang S, He Y, Xu Z, Chen L, Tao H, et al. Rice stress-resistant SNP database. Rice. 2019 Dec 23;12(1):97.
36. Xia L, Zou D, Sang J, Xu X, Yin H, Li M, et al. Rice expression database (RED): an integrated RNA-Seq-derived gene expression database for rice. J Genet Genomics Yi Chuan Xue Bao. 2017 May 20;44(5):235–41.
37. Lee T-H, Kim Y-K, Pham TTM, Song SI, Kim J-K, Kang KY, et al. RiceArrayNet: a database for correlating gene expression from transcriptome profiling, and its application to the analysis of Coexpressed genes in Rice. Plant Physiol. 2009 Sep;151(1):16–33.
38. Sato Y, Antonio BA, Namiki N, Takehisa H, Minami H, Kamatsuki K, et al. RiceXPro: a platform for monitoring gene expression in japonica rice grown under natural field conditions. Nucleic Acids Res. 2011 Jan;39(Database issue):D1141–8.
39. Cao P, Jung K-H, Choi D, Hwang D, Zhu J, Ronald PC. The Rice oligonucleotide Array database: an atlas of rice gene expression. Rice. 2012 Jul 19;5(1):17.

40. Oono Y, Yazawa T, Kanamori H, Sasaki H, Mori S, Handa H, et al. Genome-Wide Transcriptome Analysis of Cadmium Stress in Rice [Internet], vol. 2016. Hindawi: BioMed Research International; 2016. p. e9739505. [cited 2020 Dec 28]. Available from: https://www.hindawi.com/journals/bmri/2016/9739505/
41. Hamada K, Hongo K, Suwabe K, Shimizu A, Nagayama T, Abe R, et al. OryzaExpress: an integrated database of gene expression networks and omics annotations in rice. Plant Cell Physiol. 2011 Feb;52(2):220–9.
42. Yim WC, Yu Y, Song K, Jang CS, Lee B-M. PLANEX: the plant co-expression database. BMC Plant Biol. 2013 May 20;13(1):83.
43. Ogata Y, Suzuki H, Sakurai N, Shibata D. CoP: a database for characterizing co-expressed gene modules with biological information in plants. Bioinformatics. 2010 May 1;26(9):1267–8.
44. Mutwil M, Klie S, Tohge T, Giorgi FM, Wilkins O, Campbell MM, et al. PlaNet: combined sequence and expression comparisons across plant networks derived from seven species. Plant Cell. 2011 Mar 1;23(3):895–910.
45. Chien C-H, Chow C-N, Wu N-Y, Chiang-Hsieh Y-F, Hou P-F, Chang W-C. EXPath: a database of comparative expression analysis inferring metabolic pathways for plants. BMC Genomics. 2015 Jan 21;16(2):S6.
46. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, et al. PMRD: plant microRNA database. Nucleic Acids Res. 2010 Jan;38(Database issue):D806–13.
47. Obayashi T, Nishida K, Kasahara K, Kinoshita K. ATTED-II updates: condition-specific gene Coexpression to extend Coexpression analyses and applications to a broad range of flowering plants. Plant Cell Physiol. 2011 Feb;52(2):213–9.
48. Zhang Y, Zang Q, Xu B, Zheng W, Ban R, Zhang H, et al. Isomi R Bank: a research resource for tracking IsomiRs. Bioinformatics. 2016 Jul 1;32(13):2069–71.
49. Chandran AKN, Yoo Y-H, Cao P, Sharma R, Sharma M, Dardick C, et al. Updated Rice Kinase Database RKD 2.0: enabling transcriptome and functional analysis of rice kinase genes. Rice. 2016 Aug 19;9(1):40.
50. Zimmermann P, Laule O, Schmitz J, Hruz T, Bleuler S, Gruissem W. Genevestigator transcriptome meta-analysis and biomarker search using Rice and barley gene expression databases. Mol Plant. 2008 Sep;1(5):851–7.
51. Waese J, Provart NJ. The bio-analytic resource: data visualization and analytic tools for multiple levels of plant biology. Curr Plant Biol. 2016 Nov 1;7–8:2–5.
52. Guo A-Y, Chen X, Gao G, Zhang H, Zhu Q-H, Liu X-C, et al. PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res. 2008 Jan 1;36(suppl_1):D966–9.
53. Johnson C, Bowman L, Adai AT, Vance V, Sundaresan V. CSRDB: a small RNA integrated database and browser resource for cereals. Nucleic Acids Res. 2007 Jan;35(Database issue): D829–33.
54. Gupta N, Singh A, Zahra S, Kumar S. PtRFdb: a database for plant transfer RNA-derived fragments. Database [Internet]. 2018 Jan 1, 2018;(bay063). [cited 2021 Jan 11] https://doi.org/10.1093/database/bay063.
55. Priya P, Jain M. RiceSRTFDB: a database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. Database J Biol Databases Curation. 2013;2013(bat027)
56. Yilmaz A, Nishiyama MY, Fuentes BG, Souza GM, Janies D, Gray J, et al. GRASSIUS: a platform for comparative regulatory genomics across the grasses. Plant Physiol. 2009 Jan;149 (1):171–80.
57. Nakano M, Nobuta K, Vemaraju K, Tej SS, Skogen JW, Meyers BC. Plant MPSS databases: signature-based transcriptional resources for analyses of mRNA and small RNA. Nucleic Acids Res. 2006 Jan 1;34(Database issue):D731–5.
58. Zhang T, Marand AP, Jiang J. PlantDHS: a database for DNase I hypersensitive sites in plants. Nucleic Acids Res. 2016 Jan 4;44(D1):D1148–53.

59. Wu X, Zhang Y, Li QQ. PlantAPA: A Portal for Visualization and Analysis of Alternative Polyadenylation in Plants. Front Plant Sci [Internet]. 2016;7. [cited 2021 Jan 11]. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2016.00889/full.
60. Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, et al. RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Res. 2013 Jan;41 (Database issue):D1214–21.
61. Krishnan A, Gupta C, MMR A, Pereira A. RECoN: Rice Environment Coexpression Network for Systems Level Analysis of Abiotic-Stress Response. Front Plant Sci [Internet]. 2017;8. [cited 2021 Jan 8]. Available from: https://www.frontiersin.org/articles/10.3389/fpls.2017.01640/full.
62. Chandran AKN, Moon S, Yoo Y-H, Gho Y-S, Cao P, Sharma R, et al. A web-based tool for the prediction of rice transcription factor function. Database [Internet]. 2019 Jan 1 ;2019(baz061). . [cited 2021 Jan 26] doi:https://doi.org/10.1093/database/baz061.
63. Komatsu S. Rice proteome database based on two-dimensional polyacrylamide gel electrophoresis: its status in 2003. Nucleic Acids Res. 2004 Jan 1;32(90001):388D–392.
64. Murcha MW, Narsai R, Devenish J, Kubiszewski-Jakubiak S, Whelan J. MPIC: a mitochondrial protein import components database for plant and non-plant species. Plant Cell Physiol. 2015 Jan 1;56(1):e10.
65. Helmy M, Tomita M, Ishihama Y. OryzaPG-DB: Rice proteome database based on shotgun Proteogenomics. BMC Plant Biol. 2011 Apr 12;11(1):63.
66. Sapkota A, Liu X, Zhao X-M, Cao Y, Liu J, Liu Z-P, et al. DIPOS: database of interacting proteins in Oryza sativa. Mol BioSyst. 2011 Sep;7(9):2615–21.
67. Zhu P, Gu H, Jiao Y, Huang D, Chen M. Computational identification of protein-protein interactions in Rice based on the predicted Rice Interactome network. Genomics Proteomics Bioinformatics. 2011 Oct 1;9(4):128–37.
68. Gour P, Garg P, Jain R, Joseph SV, Tyagi AK, Raghuvanshi S. Manually curated database of rice proteins. Nucleic Acids Res. 2014 Jan 1;42(Database issue):D1214–21.
69. Cooper L, Meier A, Laporte M-A, Elser JL, Mungall C, Sinn BT, et al. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res [Internet]. 2017 Nov 23;46(D1). [cited 2021 Jan 11], Available from: https://www.osti.gov/pages/biblio/1625561-planteome-database-integrated-resource-reference-ontologies-plant-genomics-phenomics.
70. Sasaki T, Burr B. International Rice genome sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol. 2000 Apr;3(2):138–41.
71. Mao L, Chen M, Chu Q, Jia L, Sultana MH, Wu D, et al. RiceRelativesGD: a genomic database of rice relatives for rice research. Database [Internet]. 2019 Jan 1;2019(baz110). [cited 2021 Jan 27] https://doi.org/10.1093/database/baz110.
72. Ouyang S, Buell CR. The TIGR Plant Repeat Databases: a collective resource for the identification of repetitive sequences in plants. Nucleic Acids Res. 2004 Jan 1;32(suppl_1):D360–3.
73. Lee Y, Tsai J, Sunkara S, Karamycheva S, Pertea G, Sultana R, et al. The TIGR gene indices: clustering and assembling EST and known genes and integration with eukaryotic genomes. Nucleic Acids Res. 2005 Jan 1;33(Database Issue):D71–4.
74. Johnson AD. SNP bioinformatics: a comprehensive review of resources. Circ Cardiovasc Genet. 2009 Oct;2(5):530–6.
75. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. 2001 Jan 1;29(1):308–11.
76. Fuentes RR, Chebotarov D, Duitama J, Smith S, la Hoz JFD, Mohiyuddin M, et al. Structural variants in 3000 rice genomes. Genome Res. 2019 May 1;29(5):870–80.
77. Li J-Y, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience. 2014 May 28;3:8.
78. Yan J, Zou D, Li C, Zhang Z, Song S, Wang X. SR4R: an integrative SNP resource for genomic breeding and population research in Rice. Genomics Proteomics Bioinformatics. 2020 Apr 1;18 (2):173–85.

79. Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. 2017 Jan 4;45 (D1):D1075–81.
80. Yao W, Li G, Yu Y, Ouyang Y. funRiceGenes dataset for comprehensive understanding and application of rice functional genes. GigaScience [Internet]. 2018 Jan 1;7(gix119). [cited 2021 Jan 11] https://doi.org/10.1093/gigascience/gix119.
81. Kawahara Y, Oono Y, Wakimoto H, Ogata J, Kanamori H, Sasaki H, et al. TENOR: database for comprehensive mRNA-Seq experiments in Rice. Plant Cell Physiol. 2016 Jan 1;57(1):e7.
82. Dharmawardhana P, Ren L, Amarasinghe V, Monaco M, Thomason J, Ravenscroft D, et al. A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice N Y N. 2013 May 29;6(1):15.
83. Garg P, Jaiswal P. Databases and bioinformatics tools for rice research. Curr Plant Biol. 2016 Nov 1;7–8:39–52.
84. Krishnakumar V, Hanlon MR, Contrino S, Ferlanti ES, Karamycheva S, Kim M, et al. Araport: the Arabidopsis information portal. Nucleic Acids Res. 2015 Jan;43(Database issue):D1003–9.

4 Brief Insight into the Evolutionary History and Domestication of Wild Rice Relatives

Manoj Kumar Gupta, Gayatri Gouda, S. Sabarinathan, Ravindra Donde, Goutam Kumar Dash, Menaka Ponnana, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

Plant domestication has significantly influenced the growth of human society. The domestication of rice lists amongst the most significant historical breakthroughs. However, the sources and domestication methods are debatable. Thus, in this chapter, authors attempted to understand in brief about genetic diversity in rice, as well as a description of the processes about the domestication of rice began and at which location rice was domesticated. Information retrieved

S. Sabarinathan, Ravindra Donde, Goutam Kumar Dash and Menaka Ponnana contributed equally with all other contributors.

M. K. Gupta · G. Gouda · R. Donde · G. K. Dash · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_4

from the published literature to date suggests that the rice genus *Oryza* is a small genus comprising about 25 species, but it has incredible adaptive capabilities to differing ecological circumstances. Within the genus Oryza, two distinct domestication events have occurred—one in Asia and another in Africa. In Asia, wild rice *O. rufipogon* is popular, which was cultivated about 9000 years ago. In Africa, the wild rice Oryza, namely *O. glaberrima*, was independently domesticated around 3000 years ago. During domestication, plant experience decreased nucleotide diversity, enhanced linkage disequilibrium and modified population frequencies of polymorphic nucleotides within the domestication-related genes. In the near future, the information presented in this chapter may aid in enhancing rice's yield.

Keywords

Rice · Domestication · Evolution · Domestication traits

Abbreviations

AFR	Africa
ASN	Asia
BB	Bacterial blight
BPH	Brown plant hopper
CMS	Cytoplasmic male sterility
GLH	Green leaf hopper
IND	*indica*
JP	*japonica* (JP)
KP	Korean Peninsula
ML	Maximum Likelihood
OG	*O. glaberrima*
OR	*O. rufipogon*
OS	*O. sativa*
RYMV	Rice yellow mottle virus.
Shb	Sheath blight
WBPH	White-backed plant hopper

4.1 Introduction

One of the most significant innovations in human history has been identified as the domestication of plants and animals. Hunter-gathering communities started cultivating plant species as a primary fibre and food source, starting at the beginning of the Holocene subsequently the last major glacial phase, ~120 hundred years ago, and nowadays, we depend on domesticated species for our sustainability. Sedentary communities that generated due to shift in agriculture brought about through domestication, giving birth to towns that ultimately contributed to several modern human cultural characteristics, including literature, state creation and organized religion. All of this came about as a consequence of human reliance on a wide variety of domesticated animal (and plant) types and as a result of shifts in global human populations and ecological behaviour [1].

Since, during domestication, wild species are introduced to new selective conditions due to human agriculture and usage, domesticated crop species are the product of an evolutionary mechanism [2]. It is a technique of transformation of speciation and/or organisms that takes place when the reproduction, as well as the dispersal of one species, is regulated via another species in order to fulfil the needs of the later, especially for food [1]. Crop domestication is a unique plant–human co-evolution scenario wherein plant species evolve and propagate under human-manipulated conditions. As domestication also improvises domesticated crops and animal fitness, the outcome of this mutualism is not one-sided. Domestication process brings about drastic changes within-population levels and an extension of the distribution of domesticated species beyond their initial geographical centres of origin [3]. It is estimated that from about 120–160 taxonomic families [4, 5], there are 1000–2500 semi-domesticated and totally domesticated plant organisms, and all of these species are comparatively new, having only evolved after the Neolithic, and probably lived for only a few centuries in some instances.

For over 160 years, the domesticated organism has been the subject of evolutionary research. In 1859, Charles Darwin described in detail about the domesticated species in his book, namely the ‘Origin of Species’. In that book, he emphasized the variation between races, the similarity between offspring and ancestors, and the transformative role of selection in the distinction of species [6]. In 1868, he subsequently wrote, ‘*Variation of Plants and Animals under Domestication*’ to further explore some of the themes he generated in the ‘Origin of Species’ [7]. Since Darwin, ‘there has been great interest in the study of domestication and crop evolution, both to advance our understanding of the evolutionary process and to support the breeding of better crops to meet new adaptive challenges. The recent origin of crop species, the wealth of information on their genetics, their human association and relatively good paleontological (i.e. archaeological) record allow us to use domesticated species as models for the early stages of species formation and population divergence, and to probe the mode and tempo of various evolutionary processes’ [1].

With fast population growth and the challenge of climate change, there is an immediate need for a sustainable, global plan to ensure fair and sustainable food stability. As per the United Nations Food and Agriculture Organization, nearly 70% more food will be produced over the next three decades, sufficient to feed almost 9 billion people by 2050. Due to the fact that rice (*Oryza sativa* L.) can be a major staple crop that provides the primary source of nutrition for nearly half of the global population [8], there is always a quest to look for alternatives that may enhance the rice yield. Studies on *Oryza* genome evolution have been underway for several decades. The processes through which rice was domesticated and diversified are essential in understanding our modern way of life [9]. Recently, one study suggested that wild rice possesses many propagation-related characteristics, including the durability of prostrate development, the rate of seed shattering, the open panicle layout, the seed awning and the skill of outcrossing. The lack of shattering of the seeds was a crucial trait for the emergence of the rice cultivars since it directly supported the ancient seed gatherers, which would enable them to properly gather seeds [8]. Thus, the key aim of this chapter is to include an overview about genetic diversity in the genus *Oryza* and how and where domestication of rice started, including information about domestication trait, which in turn may help in increase grain yield.

4.2 Natural Diversity of the Genus *Oryza*

The rice genus *Oryza* is a small genus comprising about 25 species, but it has incredible adaptive capabilities to differing ecological circumstances. In contrast to other cereal plants, *Oryza* holds a distinct phylogenetic position within the subfamily Erythrina. The genus *Oryza* was described in 1753 by Linnaeus. In 1910, the number of chromosomes in rice haploid genome was established by Kuwada. The characters and subspecies of the genus *Oryza* were not well-defined until the 1960s. The features of the group typically include bisexual spikelets, rudimentary sterile lemmas and narrow, herbaceous, linear leaves having scabrous margins. Traditional nomenclature of *Oryza* species has changed at a sluggish rate since the 1960s. In 1963, Tateoka examined the range of species around the group based on research in the world's major herbaria and a field study in Asia (ASN) and Africa (AFR). He defined the essential classes of species inside the genus; subsequently, he named these groups as species complexes. The genus can be broadly classified into four species complexes, namely *O. officinalis* complex, *O. sativa* (OS) complex, *O. meyeriana* complex and *O. ridleyi* complex [10]. Polyploidization and other evolutionary events are the main reason for the speciation of *Oryza* species [11]. All of these species are strongly linked to one another (Fig. 4.1). Of these species, *OS* is grown worldwide, although Oryza spp., Oryza grandiglumis and *Oryza alta* are present in the central and southern parts of the Americas. The majority of this species biologically resides in Southern and Southern-East AFR [10].

These wild species are capable of harbouring several beneficial genes, especially for resistance against major biotic and abiotic stressors [10]. Useful trait of *O. sativa* is cultigen and high yielding. *O. nivara* is resistance to bacterial blight (BB), grassy stunt virus. *O. rufipogon* is resistance to BB, blast, brown plant hopper (BPH), tungro virus; moderately tolerant to sheath blight (Shb), tolerance to aluminium and soil acidity, increased elongation under deep water; source of cytoplasmic male sterility (CMS), and yield enhancing loci. *O. breviligulata* is resistance to green leaf hopper (GlH), BB; drought avoidance; tolerance to heat and drought. *O. glaberrima* is cultigen; tolerance to drought, acidity, iron toxicity, p-deficiency; resistance to

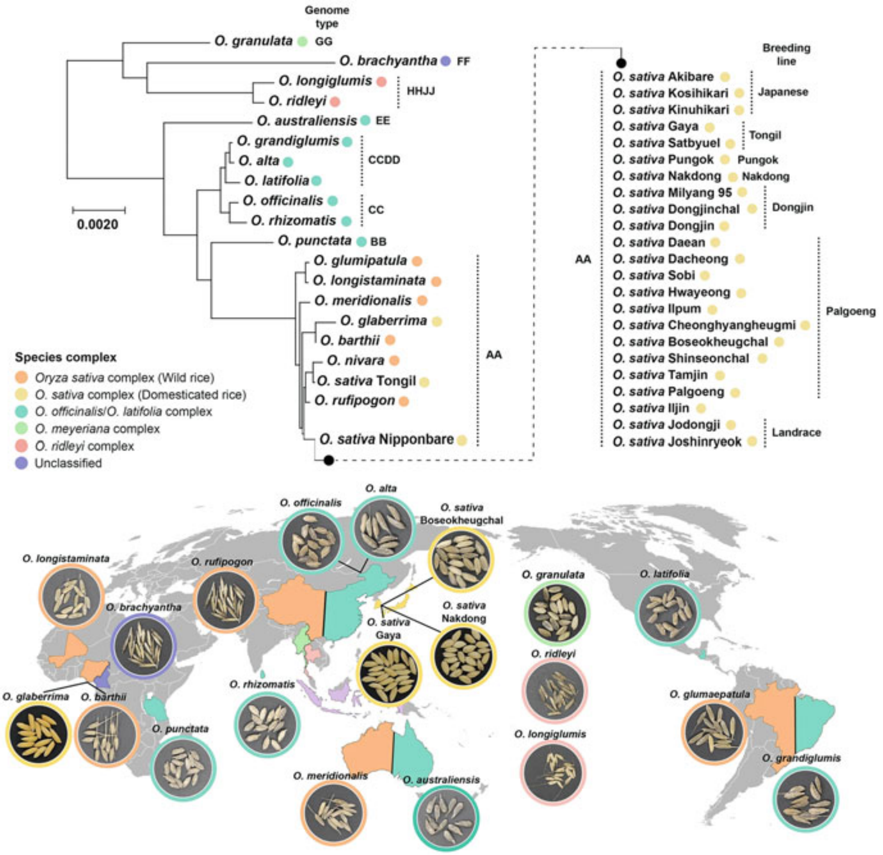


Fig. 4.1 Phylogenetic tree of samples and geographic distribution. The phylogenetic tree of *Oryza* genotypes was constructed based on the chloroplast full sequence acquired from NCBI. RAxML program was used to draw a Maximum Likelihood (ML) tree with 1000 bootstraps. OS cultivars were added separately to the tree to fully show the seed samples regardless of phylogenetic distance. The map indicates the diverse worldwide distribution of wild and domesticated *Oryza* spp. (Adapted from [11])

BB, blast, rice yellow mottle virus (RyMv), African gall midge, nematodes, weed competitiveness. *O. longistaminata* is resistance to stem borer, nematodes, BB, drought avoidance. *O. glumaepatula* have elongation ability, source of CMS and are tolerance to heat. *O. meridionalis* is drought avoidance, have elongation ability and are tolerance to drought and heat. *O. punctata* is resistance to BPH, BB, zigzag leafhopper; tolerance to heat and drought. *O. minuta* is resistance to blast, BB, GlH, BPH. *O. officinalis* is resistance to BPH, GlH, BB, thrips, WBPH, stem rot and tolerance to heat. *O. rhizomatis* is tolerant to drought avoidance, resistance to blast and tolerance to heat. *O. eichinger* is resistance to BPH, white-backed plant hopper (WBPH) and GlH. *O. latifolia* is resistance to BPH and BB and has high biomass production. *O. alta* is resistance to striped stemborer, high biomass production. *O. grandiglumis* have high biomass production. *O. australiensis* is resistance to BPH, BB and blast; drought avoidance; and tolerance to heat and drought. Both *O. granulata* and *O. meyeriana* are shade tolerance and adaptation to aerobic soil. *O. longiglumis* is resistance to blast and BB. *O. ridleyi* is resistance to blast, whorl maggot, tungro virus, BB and stem borer. *O. coarctata* is tolerant to salinity and stoloniferous. *O. schlechteri* is stoloniferous. *O. brachyantha* is resistance to yellow stem borer, BB, whorl maggot and leaf folder and tolerance to laterite soil. *Leersia perrieri* is shade-tolerant and stoloniferous [12]. However, a number of detectable phenotypic variations occur between *OS* and its wild kin [13]. Most natural wild rice seeds have long awns and extreme shattering. The domesticated variety has short awns and reduced the shattering as the goal is more to get the seeds out of the paddy than get the seeds dispersed. Keeping the seeds dormant allows for the greatest number of seeds to germinate for the maximum amount of time before achieving maturity. But, for the commercially grown crop, viability is being decreased as time goes by. Wild species have a red pigment inside of their pericarp and seed coat, which distinguish themselves from domesticated species, but certain African domesticated varieties retained their natural red pigment. Seeds in the wild have a dark straw hue, but seeds in domesticated are straw-coloured. Mating habits also vary, e.g. *O. glaberrima* (OG) and OS are almost completely inbreeding, while O. barthii and OR are partially outcrossing, with estimates varies between 10 and 50%. Domesticated grains are of a variable scale, and wild grains are generally small. The wild ancestors used to have an open panicle having less secondary branches carrying a few kernels, now that the wild relatives have been bred for a more compact and tightly packed panicle, and the number of kernels has risen [13].

4.3 Rice Domestication

Though there are differences, 'these phenotypes are not perfectly partitioned between wild and cultivated plants. While we refer to domestication "events," it is important to remember that domestication was a process that occurred over an extended period of time. Genetic loci that were selected from existing genetic variation in the wild species may appear fixed within domesticated rice but will show variation within the wild varieties. Although domestication traits are not favoured by natural selection, many of these traits are polygenic. A single allele promoting a more domesticated phenotype could be masked in the wild by a dominant allele at the same locus, or by alleles at other loci in the pathway until a chance combination of different pre-existing wild alleles produces a plant with a domestication phenotype' [14]. The genotype contributing to a domestic phenotype will not last long in the wild; however, the parents contributing to the domestic phenotype may also bear wild phenotypes and would not be selected against. Positive mutations that arose after domestication might be missing from the gene pool of wild, but early cultivars will continuously be filled with these mutations. As gene flow amongst domestic and wild rice persists, the image gets much more complex [13]. Thus, it is highly required to understand the features of domestication trait and how it helps in the domestication of various rice.

4.3.1 Domestication Trait

An agronomic trait is a domestication trait that causes crops to be grown in a more effective and profitable fashion. This trait is what essentially separates our domesticated animals from their wild equivalents. Compared to their ancestors, modern rice cultivars have been very good in breeding various cultivars. In a normal variant community, cultivars vary significantly in different characteristics, along with hull colour, pericarp colour, plant architecture, shattering, awn and grain scale. Usually, crops are domesticated due to having at least a subset of traits needed for domestication [15, 16]. In rice, this syndrome requires broad seeds to maximize the overall yield, strong relative resource distribution, highly determinate, development and apical dominance, and nonshattering seeds, both of which were favoured by our ancestors. Although the domestication syndrome dominates, certain advantageous alleles were first fixed by our ancestors since they helped in the collection of plants for production and crop characteristics [17]. These alleles have been modified over a long period of time by low-level introgression with indigenous species. This is how domesticated plants are distinguished from their wild ancestors.

Certain genes that decide how crops are domesticated have been established and disassembled, helping them to be properly understood. If a gene is important to a domestication phenotype, it is likely to have decreased nucleotide diversity, an elevated degree of association disequilibrium and changing population frequencies of polymorphisms within the gene and its related areas [18, 19]. Studies connecting phenotypes with genotypes will help researchers classify the genes responsible for phenotypes. QTL mapping is an efficient and accurate method for identifying domesticated genes linked to undesirable traits. In maize, the teosinte glume architecture (tga) gene controls inflorescence structure variants [20].

A domestication gene, *tb1*, is employed for root apical superiority in maize as comparison to its progenitor, teosinte [15]. The increased expression levels for maize have been attributed to natural human selection [21]. The Qlocus in wheat is a dominant genetic feature that regulates several characteristics, including how the grain shatters and how the chaff covers grain. It is a trans-acting transcriptional regulator specific to plants located in the AP2 band with a single amino acid change that impacts dimerization [22].

Shattering is a readily identifiable characteristic in rice that might have been easily chosen for throughout our ancestor's evolution. This enables the shattering of rice plants to be propagated more effectively. Nevertheless, qSH1is controlling shattering gene. It is an important QTL that affects shattering, encoding a transcription factor with a homeobox, and this is a persistent feature in all but one of the *japonica* (JP) subspecies. A single nucleotide in a cis-regulatory feature that regulates the shattering region causes the causative mutation [23]. The long awn rice is unacceptable during processing and storage, which was picked from the domestic populations with short awns or no awns. Analyses of the variation at the An-1 locus suggest artificial selection is reducing the genetic diversity of rice cultivars. Additionally, An-1 has a pleiotropic effect on the length and number of grains in rice [24]. Similarly, the *An-2/LABA1* gene is a domestication gene that has a minor impact on the longevity of the awn and grain yield [25, 26]. When combined with *An-1, An-2/LABA1* may have an additive effect, resulting in a longer awn height in rice. The *GAD1/RAE2* gene produces a tiny secretory signal peptide that is a member of the 'EPIDERMAL PATTERNING FACTOR-LIKE' family that is responsible for the increased number of grains per panicle, shorter grains and awnless phenotype in cultivated rice [27–29].

PROG1 is a significant QTL that plays a role in tiller initiation and tillering amount in rice. A single mutation is responsible for a single amino acid replacement in *sativa*, causes the trait shifts. This substitution occurs at the protein's functional C terminus and has a major influence in activation of transcription [30, 31]. *OsLG1,* which controls a fundamental morphological change in rice panicle shape and a greater effect on pollination and seed-shedding behaviours, is encoded via the 'SQUAMOSA promoter-binding protein' (SBP) domain and controls laminar junction and ligule development. It was discovered that the gene and its upstream region are crucial in understanding seed structure, resulting in open panicles similar to those found in the wild parent. Combining the effects of gene expression and phenotype contrasts, it can be inferred that a 9.3 kb region upstream of *OsLG1* controls its expression [32].

4.3.2 Evolutionary Origins

Within the genus *Oryza*, two distinct domestication events have occurred—one in ASN and once in AFR. In ASN, wild rice OR is popular, which was cultivated about 9000 years ago. In AFR, the wild rice *Oryza*, namely OG, was independently domesticated around 3000 years ago. Archaeological research has indicated that the third domestication in South America was confirmed to have existed during pre-Columbian times, but this crop is no longer grown [33].

4.3.3 The Origin of Asian Cultivated Rice

Since domesticated rice serves as the source of human food, the production and origin from the wild species *Oryza rufipogon* are a subject of great concern, primary in South and East ASN. The early emphasis was to trace the origin of rice cultivation. Considering this, to date, various areas were suggested, including India, the Yangtze River area in China, southern China, the coastal swamp forest ecosystems throughout South-east ASN and the so-called 'belt region' with its large variety of *Oryza* species grown along the southern Himalayas slope [34]. Although there are few pieces of research on the order of rice domestication that might have occurred ~10,000 y ago, the origin of cultivation in the other areas of the planet, such as western ASN, occurred ~10,000 y ago [35]. In the past, a study in rice origins and dispersal has gained immensely from the production of a substantial quantity of new scientific evidence from archaeological location, itself motivated by the implementation of new analytical procedures that can better diagnose rice and better track its early history. In addition, the recent widespread usage of flotation in Southern and Eastern ASN has culminated in the recovery of rich macrobotanical remnants from some significant locations [36–38]. Phytolith research has also proven valuable for detecting microscopic remnants of plants to the species level, even in extremely early deposits (e.g. Pleistocene) where husks and grains do not occur. This development helps us to classify domesticated, or at the very minimum cultivated, rice occurring outside the areas of wild rice cultivation having enough precision for distinguishing the two main rice subspecies, namely *japonica* (JP) and *indica* (IND), from each other [39–41].

Researchers now agree that the Yangtze River region within China is the prime site of rice cultivation because of recent geological discoveries. However, whether JP or IND has single or multiple origins is an unresolved controversy both in the arenas of both genetics and archaeology [42, 43]. The answer of this issue depends on studies the archaeologists have completed, which will decide the cultivation sources of IND or JP. Present debates in archaeology concentrate mainly on basic issues about the source of rice production within China and the domestication duration. Cultivation and eventual domestication are progressively seen as becoming considerably more separated in time than once assumed, as a horizon of what is called ‘predomestication cultivation’ often spanning thousands of years is being gradually documented in the Old World, and this also appears valid for rice. Furthermore, several recent reports have shown that there is no definite line break amongst agriculture and hunting-gathering and that the transition amongst the two is not a transformative move but rather a gradual phase of quantitative and qualitative changes that could have taken ~1000 years [35, 43]. These issues are linked to hypotheses from anthropology and archaeology about how agriculture and food processing evolved in the world [44]. We now examine new archaeological information that discusses the question of how rice expanded into China, the Japanese archipelago, Korean Peninsula (KP) and India.

4.3.4 China

Recently, archaeological sediments have been entirely flotation for the extraction of macrobotanical remnants of plants, and several burned plants remnants have been collected for research. They are produced from a range of different crops like rice [43]. The oldest rice remnants discovered within China are at sites in Jiangxi Province, China [45], and Zhejiang Province, China [46]. The cultural remnants of these sites are hypothesized to be existed nearly 10,000 BP ago, but it is pertinent to note that the cultural deposits within the Xianrendong cave site bear a comparatively long sequence; the lower layer was recently dated to around 20,000 BP [47]. This knowledge poses the likelihood that an additional date will also be required for the rice remnants, which were discovered at this location. Shangshan is an early Neolithic site with pit and house features along with items of stone and ceramic. The cultural assemblage and accumulation of cultural objects from the ‘Shangshan era’ (11,000 to 9000 BP) and the ‘Kuahuqiao’ period (8000 to 7000 BP) may be distinguished [48]. However, after various soil samples were floated, only 1 charred rice grain and 1 rice spikelet base were retrieved. Any of these items have to do with Kuahuqiao society. A few grains of rice were found from a Shangshan archaeological horizon. Rice traces were also contained in other geological material. To provide an example, rice husks are frequently found in the paste of pottery shards and were widely found in pottery shards dated to both times. An immense amount of rice husks was also contained in these early layers of the region. The data show that Shangshan people allegedly exploited rice intensively.

Jiahu was a permanent site dated by radiocarbon studies to the Stone Age amongst 9000 to 7800 BP. Flotation of a large number of soil samples was performed, undertaken using a large number of burned plant remnants, and rice grains were retrieved [34]. Other plants contained in the region include soya bean, water chestnut, lotus roots and acorn (Quercus sp.). Zhao’s study on the Jiahu rice implies that it may have been domesticated, since its grain phenotypic features, including size and form, seem to be quite close to current domesticated rice. The finding at another site of Hemudu site in the 1970s rendered a big international news item. Because of the damp atmosphere at the field, organic materials kept in good shape [34]. Many plants remain were uncovered of which rice was the most important. Some scholars also indicated that the Hemudu people may have had a sophisticated rice community [49]. However, nothing was understood about the anatomy of the rice that might reveal whether or not it is still wild or domesticated. The paper discusses the question of whether the rice was the food resource of preference at Hemudu, despite the availability of other edible wild plants.

The Tianluoshan site offered researchers an outstanding opportunity to address these concerns. Its location is just 7 km from Hemudu, and its archaeological facilities are virtually similar to Hemudu [34]. A sampling technique was employed to retrieve plant remnants and provide water filtering and flotation. Greater than 200 soil samples have been analysed so far, and several plant remnants have been retrieved, including Euryale ferox, rice, water chestnut, acorns, Diospyros sp., Ziziphus jujuba and numerous weed seeds. The most significant work is a systematic review of the rice spikelet bases performed via Fuller and the team [50]. The findings demonstrate that the Tianluoshan rice consisted of a high proportion of shattering, WT spikelets, which indicates that the phase of rice domestication was not yet complete in the Hemudu era, i.e. sometime around 6500 BP. The research revealed that rice was one of the most valuable foods at Tianluoshan, but they also demonstrated that rice was one of the most important crops at the location. Also, rice farming did not supplant hunting-gathering as the normal lifestyle of the local Tianluoshan people. Wild items, such as acorns, were also central to prehistoric people’s diets [34].

4.3.5 Korea

Rice is not native to the KP; therefore, rice study focuses on how the plant came to exist there and the potential routes of diffusion. It was believed that rice farming was imported to Korea from China during the 'Early Mumun' era (3400–2800 BP). Based on earlier accounts, it is likely that rice arrived in Korea during the 'Chulmun' era (7500–4000 BC) [51]. Archaeological findings have shed light on this mystery. Archaeological data now suggest that Chulmun subsistence was focused on an agricultural economy all along. However, Chulmun agriculture was a dryland farming method centred not on grains, but apparently on millets, including both broomcorn millet (*Panicum miliaceum*) and foxtail (*Setaria italica*), and legumes like adzuki bean (*Vigna angularis*), soya bean (*Glycine max*) and other crops [40]. No macrobotanical fossils have been found to have been discovered. Despite this, rice is commonly accepted in pottery shards from this time. Future research may discuss the position rice plays in this environment. Examination of the plant remnants indicated that rice cultivation was present in the past and also implied that rice farming existed. This fact indicates that rice cultivation originated in the KP, which claims that rice agriculture diffused through the KP. The path of transmission is not yet well understood.

4.3.6 Japan

Like the KP, wild rice does not exist within Japan presently, and perhaps never did. In general, it is assumed that rice agriculture first arrived within Japan mostly during 'Yayoi' era, which started in ~2800 BP [34]. The early rice farming within Japan seems to have been close to the farm-level rice farming in the KP. While rice agriculture possibly started in the 'Yayoi' era, it is more probable that rice was domesticated within Japan earlier during the 'Jomon' period, near about 4000 BP. The rice seed impressions on the 'Jomon' era pottery were recorded by the use of the scanning electronic microscopic unit. That shows that domesticated rice was imported into Japan before the 'Yayoi' period, but that is not obvious how important rice was in the subsistence of the Japanese people. This rice seems to have been a part of a dryland agriculture scheme, as numerous crops of dryland farming have been identified in sites dating back to the late and middle 'Jomon', including legumes (soya bean and adzuki beans) and barnyard millet (*Echinochloa utilis*). It is important to observe that the KP and Japan have certain parallels about the early growth of agriculture. The beginning of cultivation in these two regions was originally a dryland farming scheme with the basic crops being millets and beans. The arrival of rice to these two regions consisted of late time intervals and paddy rice farming systems. This technology soon substituted the conventional dryland farming system.

4.3.7 India

The past of JP and IND within India involves the domestication of these plants and their eventual expansion into several areas of the world. Either OR is a wild relative of *Oryza* and well distributed in India today, and was possibly present during the Pleistocene [52]. The country has a significant number of long series, with strong plant records, such as those in the Ganges Valley in India. OR and *O. nivara* are known to occur 9000 years ago [52, 53]. It is now known that the Indian subcontinent was once a centre of cultivation, with significant regions in the Ganges Plain and the Deccan Plateau. Before crops were planted, there were domesticated wild plants, such as mung beans and small-seeded grasses [53]. The question of the origin of IND rice in India has long been discussed, and recent studies have given additional information about the topic. It now appears that an independent root of cultivation of ancestral IND and proto-IND rice existed in the Ganges Plains, but the plant was fully domesticated only when it was domesticated from JP which brought from China thousands of years ago [53]. Consumption of IND started a long time ago, at least 8000 years ago. By 5000 BC, the herb was grown and seems to have become a staple product [53].

4.3.8 Independent Domestication of African Rice

Linguistic data indicate that OG is of African origin. In some West African language groups, the words for rice are antedating the Portuguese-derived words for Asian food such as Malo, Maro and Mano [54, 55]. Archaeologists have discovered rice grains dated from the 1800 BC to 800 BC in Ganjigana situated in north-east Nigeria. These ancient figurines were found in 1800 BC to 800 BC. At the nearby site of Kursakata, archaeologists have unearthed burnt grains of rice from 1200 BC up to AD 100 [56]. Often, there is little proof that the grains from these sites are domesticated and therefore not wild rice. The first known domesticated OG dates back to 300 BC and is now located at Jenne-Jeno, Mali, on the Inland Niger Delta [57]. Analysis of molecular evidence, starting with isozyme studies and verified by single nucleotide polymorphism and simple sequence repeat data, suggests that African rice is special. It is a near relative of O. barthii [58, 59]. The centre of diversity for OG is possibly the upper Niger River Delta. Porteres theorized that the OG was first grown in the floodwaters over the span of many years. Rice farming then increased utilizing nonfloating cultivars, which contributed to further selections that led to the planting of upland fields watered only by rainfall. Earlier studies have reported that ‘Asian rice was introduced into OG’s range after the initial domestication and the two species are now sown side by side in West AFR. Recently, breeders have crossed OS and OG, combining the stress-tolerance traits of OG with the yield potential of OS’ [13].

4.3.9 A Single Origin and Multiple Introgressions

An analysis of crop genomes revealed that the causal mutation within a domestication gene is set in the cultivars, triggering reduced genetic diversity, which is referred to as a 'selective sweep' [60]. Detailed research on the rice genome variants of wild rice OR accessions and cultivated OS varieties helped researchers to better understand the phylogenetic relationships amongst cultivated and wild rice and established the signatures of selection in rice domestication. This systemic comparison described 55 selective sweeps that were used in the domestication of plants [19]. If a gene was involved in a domestication phenotype, then there would be decreased nucleotide diversity, enhanced linkage disequilibrium and modified population frequencies of polymorphic nucleotides within the gene and related areas, which would provide proof to determine rice origins and domestication. Centred on rice genome variation research, cultivated rice persists as a single species that has been further separated into several subspecies over time. Few advantageous mutations arose in wild rice strains, which are then picked and propagated to produce new proto-JP-like varieties, which have been distributed to other areas of ASN. The JP descendant plants were distributed to other locations in ASN. The IND varieties were extracted from the proto-japonica-like varieties via the crosses to the OR lines emerging in a genetically homogenous community following several stages of crosses and acquisitions (detail information is provided in Ref. [60]). The more prominent mutations that have been fixed with their flanking regions with low genetic diversity offer clear proof to locate their history [60].

4.3.10 Unmasking the Origin of Rice Domestication Employing Molecular Data

Several research employing molecular markers have been conducted on the variety of rice (both the cultivated types and the wild relatives) [61]. A detailed survey of OS was performed by Glaszmann [62] utilizing isozyme signs, *sativa*. The genetic differentiation at the molecular level of both the *indica* and *japonica* forms was clearly seen in this research. With several other molecular markers, like RFLP [63], AFLP [64] and SSR [65], this finding has since been verified. However, it should be stressed that this genetic separation into two independent gene pools does not interfere with the theory of single domestication, since it may be the product of a good selection, following domestication, for the two distinct forms of plants. The key issue goes unresolved: When did the gene pool radiation in both *indica* and *japonica* take place? Recently, several studies on the timing of the distinction between the two groups, based on the study of transposable elements, have enabled us to determine that two separate domestication events were the product of the *indica* and *japonica* types. Mobile genetic sequences that are present in most living species are transposable components [61]. They can be categorized into two primary classes (I and II). Class I components, typically referred to as retrotransposons, are transposed by a copy and paste process in the form of mRNA. The class II elements transpose in the form of DNA and thus by a process of cutting and pasting. Class I components, retrotransposons, are of special importance in the study of the past of a crop as a genetic marker. Next, they are incorporated into the genome at random. Secondly, it is irreversible to insert them ('an element can only transpose by first being transcribed into a mRNA, which is in turn reverse-transcribed into DNA and can reinsert in another location into the genome'). Consequently, if two accessions have the same element in an orthologous role, it may be inferred that they come from a shared ancestor with the same element as well. Third, LTR retrotransposons, a special form of retrotransposons, may be employed to date radiation occurrences in a specific evolutionary lineage: the two LTRs (long terminal repeats) flanking the retrotransposon are purely similar in sequence when incorporated into the genome. And, over time, these two sequences diverge from one another [61]. By converting the divergence rate into a time period (molecular clock concept), the degree of divergence between the two LTRs of a given object can be converted into an insertion date. Genomic palaeontology [66] was named after this definition. It was introduced to rice for tentatively dating the radiation amongst the gene pools of *japonica* and *indica*. This date was calculated by the researchers at 200000 years, unquestionably prior to domestication (10,000 years ago). A similar research by Ma et al. [67] contributed to a similar finding, but their approximate radiation date between gene pools was 400,000 years (because the authors employed a different rate of molecular clock). See two experiments specifically illustrate that there is at least two centres for OS domestication in Asia.

In the rice domestication analysis, the next stage is to uncover the site of these domestication areas. This could be done by checking for insertions of LTR retrotransposon typical to both *indica* and *japonica* forms between the cultivated form and its wild relative. Some preliminary studies have shown that between the *japonica* varieties and some Chinese accessions of the wild relative *O. rufipogon*, multiple insertions are normal, which is compatible with the theory of a domestication core in the Yangtze River basin (Ishii, pers. comm.). In the case of the *indica* gene pool, the first screening results revealed that there are no definitive findings were obtained from *O. rufipogon* accessions from the southern hills of the Himalayas. Several writers have indicated that a diffuse process ranging from Nepal to Thailand was the domestication of *indica*-type rice. If this is the situation, perhaps the first occurrence of domestication in this area of Asia would be challenging or perhaps hard to find [61].

4.4 Benefits of Rice Domestication

Earlier genetic analysis on domestication provides some perspectives on potential domestication strategies [60]. Researching genetic changes in crops are a continuing effort. In order to face the demands of global climate change, crops must be generated at an incredibly high pace. Genetic differences favourable for those characteristics have been picked by modern breeding [68]. The high yields and grain chemical composition of modern cultivars are achieved by the synthesis of many genes associated with both domestications and improvement. The great variety of wild rice species, which have far more natural allelic variance than domesticated rice, would encourage breeding to boost its efficiency in many agronomic traits. Using cultivated rice as a base, scientists have discovered the genetic variation in wild rice, landraces and elite varieties, and created valuable tools for our comprehension of the genetic function of agronomically significant traits. The low level of genetic variability in cultivated rice would be a constraint in further changes in rice genetic diversity, and the high level of allelic variance in wild rice will be a significant resource in rice breeding that can be reintroduced into the gene pools of existing elite varieties. Taken together, both the rice domestication studies and the detailed analyses of rice genetic diversity would significantly support the research of rice gene expression in agronomic traits.

4.5 Conclusion and Future Perspective

In conclusion, the domestication process of rice took a long time, which contributed to the appearance of a new species, *Oryza sativa*, over the last 10,000 years. Research studies into the domestication phase and cloning of genes linked to domestication have uncovered further nuances than was traditionally not believed. Clearly, cultivated rice and its progenitor wild rice vary in their signature characteristics owing to domestication. Genetic analysis on OS-OR genome complex is an approach that shows special roots of rice domestication. An emerging body of research on the domestication of African rice brings into doubt the common understanding that rice domestication originated from the production of Asian rice. Genome-wide study on OS and OR species is focused on genetic variation for rice breeding. New breeding techniques can be encouraged by the use of precision gene alteration on superior parent lines culminating in the development of the ideal phenotype. Identifying the genes conferring essential traits can improve rice breeding by discovering which traits are better developed together.

Conflict of Interest None.

Additional Information Figure 4.1 [11] (CC0 1.0) Has Been Reused under Creative Commons Attribution Licenses

References

1. Purugganan MD. Evolutionary insights into the nature of plant domestication. Curr Biol. 2019 Jul 22;29(14):R705–14.
2. Purugganan MD, Fuller DQ. The nature of selection during plant domestication. Nature. 2009 Feb;457(7231):843–8.
3. Diamond J. Evolution, consequences and future of plant and animal domestication. Nature. 2002 Aug 8;418(6898):700–7.

4. Meyer RS, DuVal AE, Jensen HR. Patterns and processes in crop domestication: an historical review and quantitative analysis of 203 global food crops. New Phytol. 2012;196(1):29–48.
5. Milla R, Bastida JM, Turcotte MM, Jones G, Violle C, Osborne CP, et al. Phylogenetic patterns and phenotypic profiles of the species of plants and mammals farmed for food. Nat Ecol Evol. 2018 Nov;2(11):1808–17.
6. Darwin C. In: Murray J, editor. On the origin of species by means of natural selection, or, the preservation of favoured races in the struggle for life; 1859. 522 p.
7. Darwin CR. Variation of plants and animals under domestication. 1868 . [cited 2020 Dec 31]; Available from: https://agris.fao.org/agris-search/search.do?recordID=US201300282879.
8. Amarasinghe YPJ, Kuwata R, Nishimura A, Phan PDT, Ishikawa R, Ishii T. Evaluation of domestication loci associated with Awnlessness in cultivated Rice, Oryza sativa. Rice. 2020 Apr 28;13(1):26.
9. Chen E, Huang X, Tian Z, Wing RA, Han B. The genomics of Oryza species provides insights into Rice domestication and Heterosis. Annu Rev Plant Biol. 2019 Apr 1;70:639–65.
10. Singh PK, Venkatesan K, Swarnam TP. Chapter 12 - Rice Genetic Resources in Tropical Islands. In: Sivaperuman C, Velmurugan A, Singh AK, Jaisankar I, editors. Biodiversity and Climate Change Adaptation in Tropical Islands [Internet]. Academic Press; 2018. p. 355–84. [cited 2021 Jan 3]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128130643000120.
11. Kim H, Lee KK, Jeon J, Harris WA, Lee Y-H. Domestication of Oryza species eco-evolutionarily shapes bacterial and fungal communities in rice seed. Microbiome. 2020 Feb 14;8(1):20.
12. Sanchez PL, Wing RA, Brar DS. The Wild Relative of Rice: Genomes and Genomics. In: Zhang Q, Wing RA, editors. Genetics and Genomics of Rice [Internet], Plant Genetics and Genomics: Crops and Models. New York: Springer; 2013. p. 9–25. . [cited 2021 Jan 3]. https://doi.org/10.1007/978-1-4614-7903-1_2.
13. Sweeney M, McCouch S. The complex history of the domestication of Rice. Ann Bot. 2007 Oct 1;100(5):951–7.
14. Agnoun Y, Biaou S, Sié M, Vodouhè R, Ahanchédé A. The African Rice Oryza glaberrima Steud: knowledge distribution and prospects. Int J Biol. 2012 Jun 28;4(3):158.
15. Doebley J, Stec A, Hubbard L. The evolution of apical dominance in maize. Nature. 1997 Apr 3;386(6624):485–8.
16. Fuller DQ. Contrasting patterns in crop domestication and domestication rates: recent Archaeobotanical insights from the Old World. Ann Bot. 2007 Oct 1;100(5):903–24.
17. Ross-Ibarra J, Morrell PL, Gaut BS. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci U S A. 2007 May 15;104(Suppl 1):8641–8.
18. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated Rice. PLoS Genet. 2007 Sep 28;3(9):e163.
19. Huang X, Kurata N, Wei X, Wang Z-X, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012 Oct 25;490(7421):497–501.
20. Wang H, Nussbaum-Wagler T, Li B, Zhao Q, Vigouroux Y, Faller M, et al. The origin of the naked grains of maize. Nature. 2005 Aug 4;436(7051):714–9.
21. Wright SI, Bi IV, Schroeder SG, Yamasaki M, Doebley JF, McMullen MD, et al. The effects of artificial selection on the maize genome. Science. 2005 May 27;308(5726):1310–4.
22. Simons KJ, Fellers JP, Trick HN, Zhang Z, Tai Y-S, Gill BS, et al. Molecular characterization of the major wheat domestication gene Q. Genetics. 2006 Jan;172(1):547–55.
23. Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, et al. An SNP caused loss of seed shattering during rice domestication. Science. 2006 Jun 2;312(5778):1392–6.
24. Luo J, Liu H, Zhou T, Gu B, Huang X, Shangguan Y, et al. An-1 encodes a basic helix-loop-helix protein that regulates awn development, grain size, and grain number in rice. Plant Cell. 2013 Sep;25(9):3360–76.

25. Hua L, Wang DR, Tan L, Fu Y, Liu F, Xiao L, et al. LABA1, a domestication gene associated with long, barbed awns in wild Rice. Plant Cell. 2015 Jul;27(7):1875–88.
26. Gu B, Zhou T, Luo J, Liu H, Wang Y, Shangguan Y, et al. An-2 encodes a Cytokinin synthesis enzyme that regulates awn length and grain production in Rice. Mol Plant. 2015 Nov 2;8 (11):1635–50.
27. Yano K, Yamamoto E, Aya K, Takeuchi H, Lo P-C, Hu L, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016 Aug;48(8):927–34.
28. Bessho-Uehara K, Wang DR, Furuta T, Minami A, Nagai K, Gamuyao R, et al. Loss of function at RAE2, a previously unidentified EPFL, is required for awnlessness in cultivated Asian rice. Proc Natl Acad Sci U S A. 2016 Aug 9;113(32):8969–74.
29. Jin J, Hua L, Zhu Z, Tan L, Zhao X, Zhang W, et al. GAD1 encodes a secreted peptide that regulates grain number, grain length, and awn development in Rice domestication. Plant Cell. 2016 Oct 1;28(10):2453–63.
30. Jin J, Huang W, Gao J-P, Yang J, Shi M, Zhu M-Z, et al. Genetic control of rice plant architecture under domestication. Nat Genet. 2008 Nov;40(11):1365–9.
31. Tan L, Li X, Liu F, Sun X, Li C, Zhu Z, et al. Control of a key transition from prostrate to erect growth in rice domestication. Nat Genet. 2008 Nov;40(11):1360–4.
32. Ishii T, Numaguchi K, Miura K, Yoshida K, Thanh PT, Htun TM, et al. OsLG1 regulates a closed panicle trait in domesticated rice. Nat Genet. 2013 Apr;45(4):462–5.
33. Choi JY, Zaidem M, Gutaker R, Dorph K, Singh RK, Purugganan MD. The complex geography of domestication of the African rice Oryza glaberrima. PLoS Genet. 2019 Mar 7;15(3): e1007414.
34. Gross BL, Zhao Z. Archaeological and genetic insights into the origins of domesticated rice. PNAS. 2014 Apr 29;111(17):6190–7.
35. Zeder MA. The origins of agriculture in the near east. Curr Anthropol. 2011 Oct 1;52(S4): S221–35.
36. Watson PJ. In pursuit of prehistoric subsistence: a comparative account of some contemporary flotation techniques. Midcont J Archaeol. 1976;1(1):77–100.
37. Crawford GW. Paleoethnobotany of the Kameda peninsula Jomon. University of Michigan Press; 1983. 215 p.
38. Pearsall DM. Paleoethnobotany: a handbook of procedures. 3rd ed. Walnut Creek, California: Routledge; 2015. 513 p.
39. Fuller DQ. Agricultural origins and Frontiers in South Asia: a working synthesis. J World Prehist. 2006 Mar 1;20(1):1–86.
40. Lee G-A. The transition from foraging to farming in prehistoric Korea. Curr Anthropol. 2011 Oct 1;52(S4):S307–29.
41. Gu Y, Zhao Z, Pearsall DM. Phytolith morphology research on wild and domesticated rice species in East Asia. Quat Int. 2013 Feb 21;287:141–8.
42. Fuller DQ, Sato Y-I. Japonica rice carried to, not from, Southeast Asia. Nat Genet. 2008 Nov;40 (11):1264–5.
43. Zhao Z. New Archaeobotanic data for the study of the origins of agriculture in China. Curr Anthropol. 2011 Oct 1;52(S4):S295–306.
44. Price TD, Bar-Yosef O. The origins of agriculture: new data, new ideas: an introduction to supplement 4. Curr Anthropol. 2011 Oct 1;52(S4):S163–74.
45. Zhijun Z. The middle Yangtze region in China is one place where rice was domesticated: phytolith evidence from the Diaotonghuan cave, Northern Jiangxi. Antiquity. 1998 Dec;72 (278):885–97.
46. Jiang L, Liu L. New evidence for the origins of sedentism and rice domestication in the lower Yangzi River, China. Antiquity. 2006 Jun;80(308):355–61.
47. Wu X, Zhang C, Goldberg P, Cohen D, Pan Y, Arpin T, et al. Early pottery at 20,000 years ago in Xianrendong cave, China. Science. 2012 Jun 29;336(6089):1696–700.

48. Jiang L. The Shangshan Neolithic site in Pujiang County, Zhejiang: new evidence of rice civilization in the lower Yangtze River region. Gudai Wenming Yanjiu Zhongxin Tongxun. 2005;7:51–5.
49. Ellis JR, Pashley CH. Burke* JM, McCAULEY DE. High genetic diversity in a rare and endangered sunflower as compared to a common congener. Mol Ecol. 2006;15(9):2345–55.
50. Fuller DQ, Qin L, Zheng Y, Zhao Z, Chen X, Hosoya LA, et al. The domestication process and domestication rate in Rice: spikelet bases from the lower Yangtze. Science. 2009 Mar 20;323 (5921):1607–10.
51. Ahn S-M. The emergence of rice agriculture in Korea: archaeobotanical perspectives. Archaeol Anthropol Sci. 2010 Jun 1;2(2):89–98.
52. Fuller DQ, Allaby RG, Stevens C. Domestication as innovation: the entanglement of techniques, technology and chance in the domestication of cereal crops. World Archaeol. 2010 Mar 1;42(1):13–28.
53. Fuller DQ. Finding plant domestication in the Indian subcontinent. Curr Anthropol. 2011 Oct 1;52(S4):S347–62.
54. Fage JD, Oliver RA. Papers in African Prehistory. CUP Archive; 1970. 348 p.
55. Blench R. Archaeology, language, and the African past. Lanham: Rowman Altamira; 2006. 392 p.
56. Klee M, Zach B, Neumann K. Four thousand years of plant exploitation in the Chad Basin of northeast Nigeria I: The archaeobotany of Kursakata. Veg History Archaeobot. 2000;9:223–37.
57. McCouch SR, Sweeney M, Li J, Jiang H, Thomson M, Septiningsih E, et al. Through the genetic bottleneck: O. rufipogon as a source of trait-enhancing alleles for O. sativa. Euphytica [Internet]. 2007. [cited 2021 Jan 4]; Available from: https://agris.fao.org/agris-search/search.do?recordID=US201300754269.
58. Second G. Origin of the genic diversity of cultivated rice (*Oryza spp*.): study of the polymorphism scored at 40 isozyme loci. Japan J Gen. 1982;57(1):25–57.
59. Semon M, Nielsen R, Jones MP, McCouch SR. The population structure of African cultivated rice oryza glaberrima (Steud.): evidence for elevated levels of linkage disequilibrium caused by admixture with O. sativa and ecological adaptation. Genetics. 2005 Mar;169(3):1639–47.
60. Chen E, Huang X, Han B. How can rice genetics benefit from rice-domestication study? Natl Sci Rev. 2016 Sep 1;3(3):278–80.
61. Panaud O. The molecular bases of cereal domestication and the history of rice. C R Biol. 2009 Feb 1;332(2):267–72.
62. Glaszmann JC. Isozymes and classification of Asian rice varieties. Theoret Appl Genet. 1987 May 1;74(1):21–30.
63. Wang ZY, Tanksley SD. Restriction fragment length polymorphism in Oryza sativa L. Genome [Internet]. 2011 Feb 15. [cited 2021 Jan 25]; Available from: https://cdnsciencepub.com/doi/abs/10.1139/g89-563.
64. Prashanth SR, Parani M, Mohanty BP, Talame V, Tuberosa R, Parida A. Genetic diversity in cultivars and landraces of Oryza sativa subsp. indica as revealed by AFLP markers. Genome [Internet]. 2011 Feb 15. [cited 2021 Jan 25]; Available from: https://cdnsciencepub.com/doi/abs/10.1139/g02-003.
65. Garris AJ, Tai TH, Coburn J, Kresovich S, McCouch S. Genetic structure and diversity in Oryza sativa L. Genetics. 2005 Mar 1;169(3):1631–8.
66. Vitte C, Ishii T, Lamy F, Brar D, Panaud O. Genomic paleontology provides evidence for two distinct origins of Asian rice (Oryza sativa L.). Mol Gen Genomics. 2004 Dec 1;272(5):504–11.
67. Ma J, Devos KM, Bennetzen JL. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in Rice. Genome Res. 2004 May 1;14(5):860–9.
68. Wang Z-Y, Zheng F-Q, Shen G-Z, Gao J-P, Snustad DP, Li M-G, et al. The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. Plant J. 1995;7 (4):613–22.

3000 Genome Project: A Brief Insight

5

Manoj Kumar Gupta, Gayatri Gouda, Ravindra Donde, S. Sabarinathan, Goutam Kumar Dash, N. Rajesh, Menaka Ponnana, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

The main food of half the world's population is rice, *Oryza sativa* L. By 2030, rice production must increase by at least 25% to meet global food demand of ever growing human population. In order to reduce the impact of climate change and arable land loss and ensure stable global food supplies, accelerated genetic gains during rice improvement are highly required. Since this process is complicated, we first need to have detail information regarding the genetic diversity of the

S. Sabarinathan, Goutam Kumar Dash, N. Rajesh and Menaka Ponnana contributed equally with all other contributors.

M. K. Gupta · G. Gouda · R. Donde · G. K. Dash · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Dept of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_5

oryza sp. gene pool, the correlation of diverse alleles with essential traits of rice, and the systematic use of the rich genetic diversity through employing methods that adopt expertise in rice improvements through breeding strategies. Considering this, in 2014, an international sequencing project of 3000 rice genomes was published. These details information may help us to detect novel alleles associated with important phenotypes of rice by employing various bioinformatics or genetic methods. It will also help us to unmask the *O. Sativa* genomic diversity more precisely. This project also encouraging the global rice community to employ data present in the 3000 rice genomes project for establishing various global public rice genetic/genomic database, which in turn will promote rice-breeding technology in the future. Thus, in this chapter, authors made an attempt to understand in brief about the various germplasms employed in 3000 genome project and the genetic diversity of *O. sativa*, which, in the near future, may help us to increase grain yield of rice.

Keywords

3000 rice genome project · Population structure · Genetic diversity · Structural variation

Abbreviations

3RGP	3000 rice genomes project
CAAS	Chinese Academy of Agricultural Sciences
CNCGB	China National Crop Gene Bank
ICS	Institute of Crop Sciences
IRGC	International Rice Genebank Collection
IRRI	International Rice Research Institute
MAF	Minor allele frequency
NTE	Nontransposable elements
SNP	Single nucleotide polymorphism
SVs	Structural variants
TE	Transposable elements

5.1 Introduction

For majority of the world's poor, rice (*Oryza sativa* L.) offers most of the regular calories in their diet. However, due to the continuous growth of the human population, there is a continuous demand for food crops, including rice [1]. The global rice production is estimated to upsurge by 25% or more by the next decade (Seck et al., 2012). Thanks to the Green Movement, where in addition to selective breeding, plant breeders are continuously exploiting genetic diversity of the rice plant to enhance grain yield [1]. Researchers are also continuously looking for approaches where we can reduce the scale of our farms, e.g., less water and less land, to improve their efficiency and flexibility under the increasingly extreme environmental pressures that would be triggered by climate change. Thus, cereal grains would have to continue to grow in a way to enable them to sustain more resilience by genetic modifications for enhancing yield capacity and quality [2]. Since this process is complicated, we first need to have more information regarding the genetic diversity of the *oryza sp.* gene pool, the correlation of diverse alleles with essential rice traits, and the systematic use of this rich genetic diversity by implementing methods that adopt expertise in rice improvements through breeding strategies [2]. Considering this, in 2014, a group of researchers across the world established the "3,000 rice genomes project" (3RGP), which provides detail insight into the genetic diversity of ~3000 *O. sativa* genomes across various geographical location [2]. This data is an important source for discovering novel alleles for developmental and stress-related rice phenotypes. It may also aid in unmasking the level of diversity in *O. sativa* genome at a more in-depth level. Thus, in this chapter, authors made an attempt to understand in brief about the various germplasms employed in 3000 genome project and the genetic diversity of *O. sativa*, which, in the near future, may help us to increase grain yield of rice.

5.2 Germplasms and Sequencing of 3GRP

In 3RGP, for sequencing ~3000 germplasm accessions were selected, which is mainly comprised of 534 and 2466 accession from the "China National Crop Gene Bank" (CNCGB) in the "Institute of Crop Sciences, Chinese Academy of Agricultural Sciences" (CAAS) and the "International Rice Genebank Collection" (IRGC) at the "International Rice Research Institute" (IRRI), respectively [2]. The 2466 accessions provided via IRRI reflect a panel randomly chosen from 12,000 *O. sativa*, which in turn were selected from >101,000 rice accessions in the IRRI genebank; considering factors like eco-cultural type, the country of origin, and varietal grouping, while restricting redundant data from each country, and supplemented through distinct, nominated entries from IRRI and the "Centre de Coopération Internationale en Recherche Agronomique pour le Développement". The 534 accessions that have been contributed via CAAS comprised of a core selection of 246 accessions selected from ~932 accessions generated in the similar manner from the 61,470 lineages of *O. sativa* conserved within the CNCGB, and 288 distinct accessions that had been chosen on the basis of their isozyme activity, and employed as parental lines within the international rice molecular breeding network [2]. Overall, the 3000 sampled rice accessions considered in 3RGP were obtained from 89 different regions/countries. Of all, 33.9%, 25.6%, and 17.6% came from Southeast Asia, South Asia, and China, respectively [2]. Employing Illumina-based next generation technology and Os-Nipponbare-Reference-IRGSP-1.0 (IRGSP-1.0) [3], the 3RGP data were generated that have an average sequencing depth of 14X, average mapping rates, and genome coverage of 94.0% and 92.5%, respectively. Raw sequencing data are available from DDBJ (accession ERP005654), GigaDB (http://gigadb.org/dataset/200001), and EBi & NCBI (accession PRJEB6180).

5.3 Genome Size, Population Structure, and Genetic Diversity

With the aid of biotechnology, the objective of rice breeding is not only to increase crop productivity but also to improve the quality characteristics by mutation. To date, many enormous efforts and rapid progress have been made in the rice breeding programs, and remarkable achievements have been achieved. As a consequence, new varieties of rice with higher yield and quality have been developed and released [4, 5]. In several crops, including rice, mutation serves as an effective approach for producing rice varieties with desired traits. Mutation may either be induced with physical agent or naturally. When induced naturally, it may be transmitted from generation to generation. At present, the mutation serves as the most successful plant breeding approach in line with transgenic breeding and recombinant breeding, in particular during sexual production [4]. Genetic variation, as the key component of germplasm, is a normal source for rice breeding to fulfill current food requirements. Earlier studies have reported that the "higher the level of genetic variation in the population, the more valuable it is as a resource used in the breeding program" [4]. DNA markers and genetic engineering may serve as a reliable source for detecting genetic diversity in various plants [6]. They can also detect the differentiation amongst individuals, accessions, and characterization of novel germplasms at the molecular level, which in turn can be used for plant breeding [5, 7].

Original phylogenetic analyses of 3RGP [2] found that the 3000 accessions were specifically divided into two main groups: *indica* and *japonica*, two tiny varietal groups: aus/boro and basmati/sadri, plus an additional community (134) of intermediate forms (admixed). The *indica* group was the biggest and most representative group with 1760 (58.2%) memberships of five different subgroups with varying backgrounds. There were 843 (27.9 percent) accessions in the *japonica* group, which had two well-differentiated subgroups, 388 temperate *japonicas* and 455 *tropical japonicas*. The aus/boro group consists of 215 accessions and is more closely related to the *indica* group, while the *japonica* group is more closely related to the aromatic basmati/sadri group and consists of 68 accessions, mostly from South Asia [2]. Later population structure and diversity of 3RGP data through Wang and the team reported that genotype of 3RGP can be broadly classified into nine subpopulations (Fig. 5.1), majority of which can be linked by their geographic origins [8]. "There were four XI clusters (XI-1A from East Asia, XI-1B of modern varieties of diverse origins, XI-2 from South Asia, and XI-3 from Southeast Asia); three GJ clusters (primarily East Asian temperate (named GJ-tmp), Southeast Asian subtropical (named GJ-sbtrp), and Southeast Asian tropical (named GJ-trp)); and single groups for the mostly South Asian cA and cB accessions. Accessions with admixture components <0.65 within XI and GJ were classified as 'XI-adm' and 'GJ-adm', respectively, and

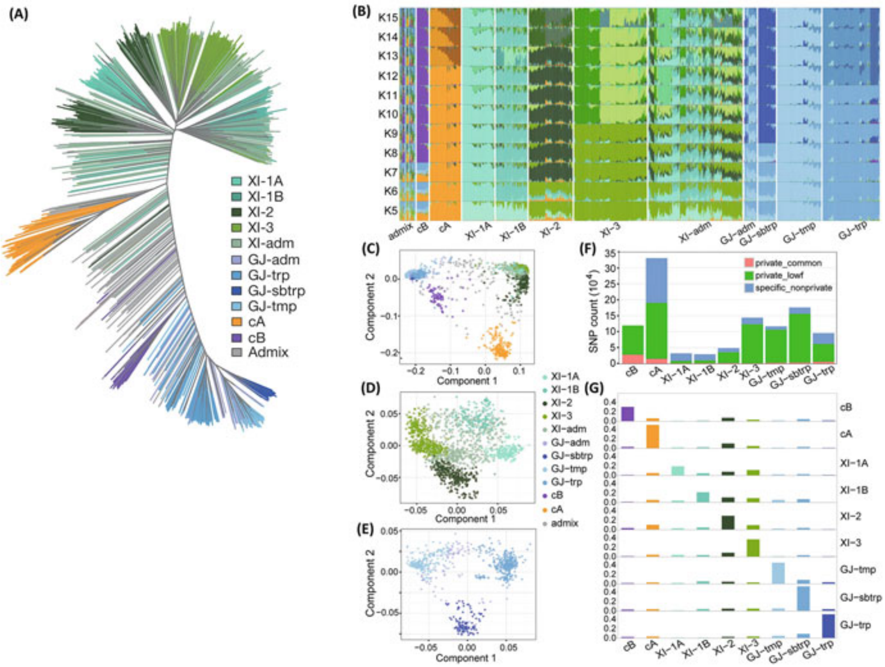


Fig. 5.1 (**a**) Unweighted neighbor-joining tree based on 3010 samples and computed on a simple matching distance matrix for filtered SNPs. (**b**) ADMIXTURE analyses for $k = 5$ to $k = 15$. (**c–d**) Multidimensional scaling plots for all ($n = 3010$) (**c**), XI ($n = 1786$) (**d**) and GJ ($n = 849$) (**e**) accessions. (**e**) Private and specific SNPs in each subpopulation. Private alleles are defined as being present in at least four accessions in a subpopulation and not found in other subpopulations; population-specific alleles are common in the subpopulation (≥20%) but of low frequency (<2%) in others. (f) Doubleton sharing—that is, SNPs shared by two accessions—within and between subpopulations, with values normalized by the sample sizes (Adapted from [8])

accessions that fell between major groups were classified as admixed" (Fig. 5.1b) [8].

Recent genome size and SNPs analysis of 3RGP genome via aligning with *O. sativa* cv. Nipponbare IRGSP 1.0 reference genome have reported that average mapping coverage of 3RGP genome is 92% (74.6–98.7%) [8]. They have also found over 29 million SNPs, and they are almost all bi-allelic. Filtering narrowed the data collection to a typical set of 17 million SNPs, which recorded the bulk (>99.9%) of SNPs having MAF > 0.25%. The majority (91%) and a half (56%) of the transposable element and nontransposable element (NTE) genes experience a large number of high-effect SNPs. Allele frequency profiles for SNPs with MAF greater than 10%, represented the broad variety of adaptations and demographic events (Fig. 5.2a). "Private" alleles were found to be more abundant in subpopulations cA & cB in comparison to other subpopulations (Fig. 5.1f). In comparison to other subpopulations, the IX subpopulations have a lower total number of "private alleles", most likely due to continuing gene transfer from natural hybridization as

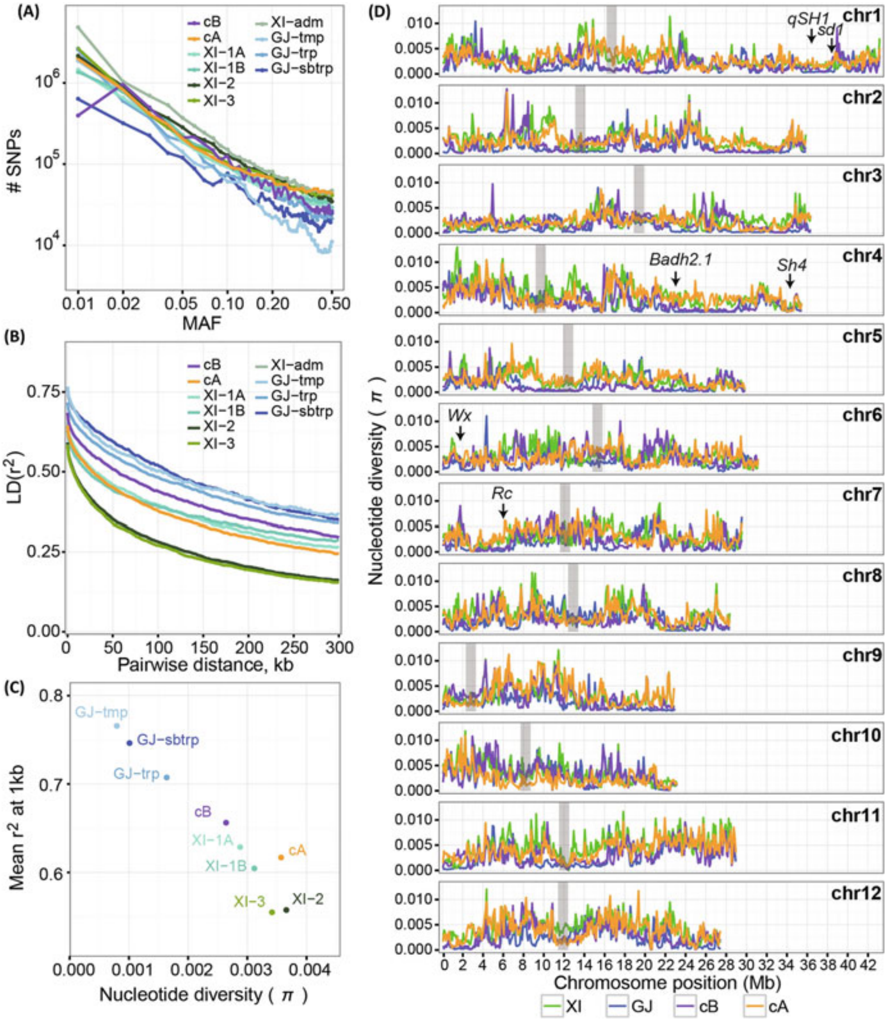


Fig. 5.2 (**a**) MAF histogram. (**b**) Genome-wide linkage disequilibrium. (**c**) Nucleotide diversity versus linkage disequilibrium. (**d**) Diversity scans (*π*) for all chromosomes for major groups (XI, GJ, cA, and cB) using 100-kb windows in which centromeric regions are highlighted in gray (Adapted from [8])

well as breeding. Same doubleton sharing pattern found between and within subpopulations [8]. They also reported that the link disequilibrium decay rates for combined subpopulations are greater in XI in comparison to GJ, with few variations amongst the two GJ subpopulations. However, when looking at all nine subpopulations, linkage disequilibrium decay differed across the nine subpopulations, with XI-3 & XI-2 showing much greater linkage disequilibrium decay rates than I-1B and IX-1A did (Fig. 5.2b) [8]. To explain how a gene is

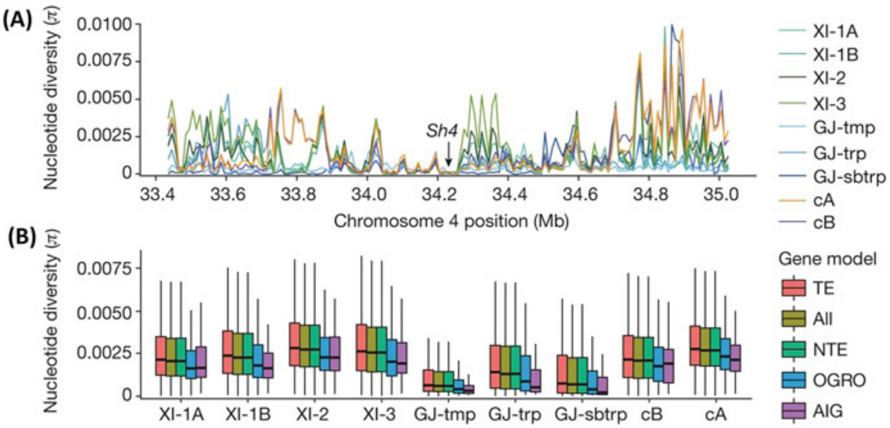


Fig. 5.3 (**a**) Differential nucleotide diversity between subpopulations at the *Sh4* locus on chromosome 4 using 10-kb sliding windows. (**b**) Box plots of the distribution of π in 100-kb regions surrounding gene models across the genome. Box plots are shown for $k = 9$ subpopulations for all 100-kb windows (All) ($n = 3728$ in total) and those containing genes annotated as transposable elements (TE) ($n = 3305$ windows), NTE ($n = 3709$), from the OGRO/QTARO database (OGRO) ($n = 828$) and the subset of 78 domestication-related genes (AIG) ($n = 61$ windows). Box plots show the median, box edges represent the first and third quartiles, and the whiskers extend to farthest data points within 1.5× interquartile range outside box edges (Adapted from [8])

controlled, a study was conducted to see how many genes are kept in certain regions of reduced diversity where the gene experiences little or no constrain (Fig. 5.2d). *Sh419* [9], which regulates nonshattering, displayed an analogous diversification trend overall subpopulations (Fig. 5.3a), suggests much longer selection then *qSH120* [10]. At the *sd121* locus, a decreased genetic variation existed on all major branches of the tree. This has a similar pattern as observed in *qSH1* [8]. However, greater diversity within the 100-kb regions existed in the XI, cA, and cB groups, while the GJ groups had decreased diversity and this represents the breeding past connected with the "Green Revolution" [8]. Other significant declines in diversity were witnessed at other essential loci. The *Wx23* [11] locus that influences amylose contents well as stickiness on cooking, the *Badh2.1* [12] locus that influences fragrance as well as their nearby regions are extremely variable in the XI, cA, and cB classes, which suggest complicated backgrounds for selection for various styles of eating qualities [8]. The *Rc25* locus [13] is very poor in diversity within all the various classes, and there are a lot of different diversity scenarios in XI, cA, and cB.

5.4 Structural Variations

Structural variants (SVs) identification and characterization have revolutionized the perception of the landscape of genetic variance in numerous organisms. A structural variant is usually characterized as a genome alteration (with respect to a reference genome) with a different number of copies (i.e., deletion, loss, and gain), chromosome position or orientation [14, 15]. Structural variations account for more differing base pairs in human genomes than SNPs; but, in plants, SV studies are still restricted. While less prevalent than SNPs, owing to their wider size and the likelihood of modifying gene composition, dose, or position, SVs have a greater capacity to influence activity [15]. Following the discovery that structural genetic heterogeneity in human genomes is widespread, several SV experiments have been undertaken in other animals, ranging from agriculturally significant to extinct ones [14, 15]. However, owing to the absence of high-quality reference genomes [14, 15] and rigorous approaches, all of which are needed to discover and genotype SVs, the discovery of SVs has traditionally lagged behind discovering single-nucleotide variants. Structural variations in plants are not recognized as polymorphisms that influence specific plants, rather as differentiating elements amongst cultivars/ accessions of one genus [16]. In order to find hundreds of SVs, maize is the first plant species to be thoroughly questioned. The link between SVs and plant phenotypes has already been shown by many experiments in plants [17]. Another study has suggested that early and late flowering are caused by the increased copy number of *Ppd-B1* genes and *Vrn-A1* in wheat, respectively [18].

Recently, SVs analysis focused only on 453 accessions having mapping depths >15× and sequencing depths >20×, because genome coverage stabilized when sequencing depths >20× was carried out in genome of 3RG. Result obtained revealed overall 93,683 SVs having 582 SVs > 500 kb. Per genome average SVs detected is 12,178 SVs per genome. The average sizes of the detected inversions, duplications, and deletions, are 105.1 ± 22.7 kb, 127.1 ± 19.4 kb, and 5.3 ± 0.6 kb, respectively (Figs. 5.4 and 5.5). SVs displayed very good XI–GJ distinction. On average, each XI, cA, and cB accession differed from Nipponbare RefSeq by 14,754 SVs, 12,997 SVs, and 7892 SVs, respectively. Overall SV sequence that differentiated amongst GI & XI accessions is ~71 Mb, and two GJ accessions is ~22 Mb [8].

Importantly, 1940 SVs interrupted protein-coding regions within GJ, whereas >6518 SVs occurred amongst XI and GJ accessions that disrupted protein-coding regions were detected (Fig. 5.4c). The SV phylogenetic tree shows similarity to the SNP tree. And the branch dividing XI, GJ, cA, and cB accessions obviously indicates a variety of variations between them (Fig. 5.4d). In comparison, 44.7% of all SVs and 41.0% of 582 wide SVs accounted for 41,957 major-group-unbalanced SVs that were unevenly distributed amongst XI, GJ, cA, and cB accessions (Fig. 5.4e).

5.5 Conclusion and Future Perspective

In conclusion, the completion of 3 K rice genome sequencing and preliminary research is only the first step in setting up an automated database knowledge network and specialized technologies to improve rice breeding [19]. This initiative would be close in context to the creation of the Arabidopsis Information Portal (AIP) [20]. The

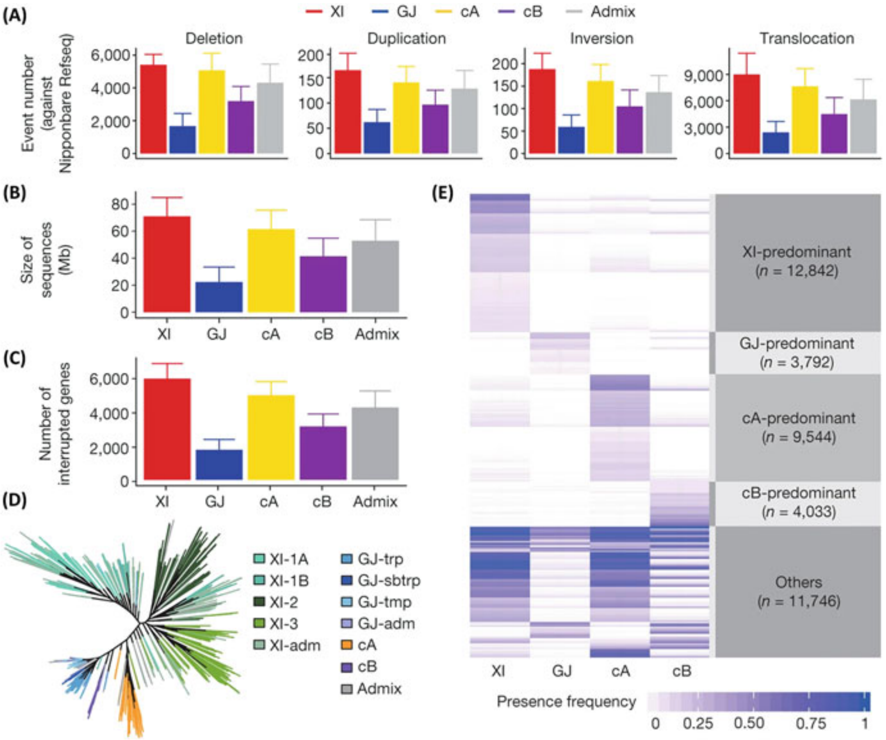


Fig. 5.4 (**a**) Number of deletions, duplications, inversions, and translocations. (**b**) Genome sizes affected by SVs. (**c**) Numbers of genes affected (included or interrupted) by the SVs. (**d**) Phylogenetic relationship of 453 rice accessions built from 10,000 randomly selected SVs. (**e**) Characterization of the 42,207 major-group-unbalanced SVs unevenly distributed among XI, GJ, cA, and cB on the basis of two-sided Fisher's exact tests. Bar plots in A-C are mean ± s.d. and numbers of accessions in XI, GJ, cA, cB, and admix are 303, 92, 33, 10, and 15, respectively (Adapted from [8])

"International Rice Informatics Consortium" (IRIC) under "Global Rice Science Partnership" (GRiSP) has been founded by IRRI. Discussions are also under way to formalize the IRIC consortium arrangement and technological dimensions of portal architecture, interoperability meta-data specifications, and persistent, diagnostic signatures of germplasm. The first priorities involve the curation of data on 3 K rice genomes and other public data, the description of reference genomes, the creation and archival of phenotyping datasets, a web-based framework or gateway and population structure resources, analyses of genome-wide interactions, and browsing of diversity. Even, it takes a long-term global initiative in rice functional genomics research to relate variability in the 3 K rice genome dataset to phenotypic variation as well as environmental adaptation. For a more full awareness of OS, future study could not only concentrate on recognizing and characterizing specific genes/alleles with a broad impact, but also on unique allelic combinations that underpin complicated features, genetic variation, and genes underlying significant

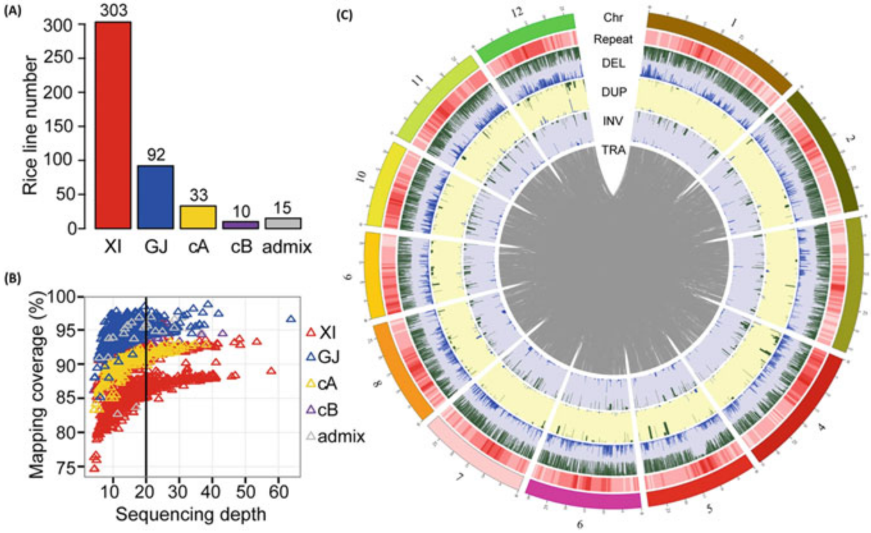


Fig. 5.5 (**a**) Number of accessions with sequencing depths ≥20× and mapping depth ≥15×. (**b**) Mapping coverage of the 3010 rice genomes to the Nipponbare RefSeq as a function of sequence depth. (**c**) Circular presentation of different types of structural variation detected in 453 high-coverage rice genomes when compared against the Nipponbare RefSeq. Chr, outermost circle represents 12 rice chromosomes with marks in Mb; Repeat, red heat map represents repeat content in 500-kb windows; DEL, green/blue color with inner/outer bars represents the average frequencies of deletions detected in XI and GJ; DUP, green/blue color with inner/outer bars represents the average frequencies of duplications detected in XI and GJ; INV, green/blue color with inner/outer bars represents the average frequencies of inversions detected in XI and GJ; TRA, gray color represents translocations across each genome with an average frequency > 0.3 in either XI or GJ. (Adapted from [8])

rice traits. A more comprehensive exploration and increased usage of rich genetic variation would be feasible with such value-added knowledge embedded into the database and access to suitable resources via the Web portal [21–23]. Although this project would certainly stimulate another round of rapid advancements in rice genetics, there are various challenges in extracting the most knowledge from sequence and phenomics data to build a global public information portal that will not only be useful for experimental analysis, but also for realistic rice breeding. These problems can be resolved by global initiatives in the field of rice science to ensure technological progress and the distribution of benefits to rice farmers and also to sustain human food protection. The task is immense and will demand extraordinary teamwork that transcends global, systemic, and personal ambitions.

Conflict of Interest None.

Additional Information Figures 5.1, 5.2, 5.3, 5.4, and 5.5 (CC BY 4.0) [8] have been reused under Creative Commons Attribution licenses.

References

1. Seck PA, Diagne A, Mohanty S, Wopereis MC. Crops that feed the world 7: Rice. Food Secur. 2012;4:7–24.
2. The 3 000 Rice Genomes Project. The 3,000 rice genomes project. GigaScience. 2014;3:7.
3. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6:4.
4. Tu Anh TT, Khanh TD, Dat TD, Xuan TD. Identification of phenotypic variation and genetic diversity in Rice (Oryza sativa L.) mutants. Agriculture. 2018;8:30.
5. Gouda G, Gupta MK, Donde R, Kumar J, Parida M, Mohapatra T, et al. Characterization of haplotypes and single nucleotide polymorphisms associated with Gn1a for high grain number formation in rice plant. Genomics. 2020;112:2647–57.
6. Ram SG, Thiruvengadam V, Vinod KK. Genetic diversity among cultivars, landraces and wild relatives of rice as revealed by microsatellite markers. J Appl Genet. 2007;48:337–45.
7. Edzesi WM, Dang X, Liang L, Liu E, Zaid IU, Hong D. Genetic diversity and elite allele Mining for Grain Traits in Rice (Oryza sativa L.) by association mapping. Front. Plant Sci. 2016;7:787.
8. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557:43–9.
9. Li C, Zhou A, Sang T. Rice domestication by reducing shattering. Science. 2006;311:1936–9.
10. Konishi S, Izawa T, Lin SY, Ebana K, Fukuta Y, Sasaki T, et al. An SNP caused loss of seed shattering during rice domestication. Science. 2006;312:1392–6.
11. Wang Z-Y, Zheng F-Q, Shen G-Z, Gao J-P, Snustad DP, Li M-G, et al. The amylose content in rice endosperm is related to the post-transcriptional regulation of the waxy gene. Plant J. 1995;7:613–22.
12. Chen S, Yang Y, Shi W, Ji Q, He F, Zhang Z, et al. Badh2, encoding betaine aldehyde dehydrogenase, inhibits the biosynthesis of 2-Acetyl-1-Pyrroline, a major component in Rice fragrance. The Plant Cell American Society of Plant Biologists. 2008;20:1850–61.
13. Sweeney MT, Thomson MJ, Pfeil BE, McCouch S. Caught red-handed: Rc encodes a basic helix-loop-helix protein conditioning red pericarp in rice. Plant Cell. 2006;18:283–94.
14. Escaramís G, Docampo E, Rabionet R. A decade of structural variants: description, history and methods to detect structural variation. Brief Funct Genomics. 2015;14:305–14.
15. Fuentes RR, Chebotarov D, Duitama J, Smith S, la Hoz JFD, Mohiyuddin M, et al. Structural variants in 3000 rice genomes. Genome Res. 2019;29:870–80.
16. Francia E, Pecchioni N, Policriti A, Scalabrin S. CNV and structural variation in plants: prospects of NGS approaches. In: Sablok G, Kumar S, Ueno S, Kuo J, Varotto C, editors. Advances in the understanding of biological sciences using next generation sequencing (NGS) approaches [internet]. Cham: Springer International Publishing; 2015. p. 211–32. [cited 2020 Dec 27]. Available from: https://doi.org/10.1007/978-3-319-17157-9_13.
17. Żmieńko A, Samelak A, Kozłowski P, Figlerowicz M. Copy number polymorphism in plant genomes. Theor Appl Genet. 2014;127:1–18.
18. Würschum T, Boeven PHG, Langer SM, Longin CFH, Leiser WL. Multiply to conquer: copy number variations at Ppd-B1 and Vrn-A1 facilitate global adaptation in wheat. BMC Genet. 2015;16:96.
19. Li J-Y, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. Gigascience. 2014;3:8.

20. International Arabidopsis Informatics Consortium. An international bioinformatics infrastructure to underpin the Arabidopsis community. Plant Cell. 2010;22:2530–6.
21. Li Z-K, Zhang F. Rice breeding in the post-genomics era: from concept to practice. Curr Opin Plant Biol. 2013;16:261–9.
22. Douchkov D, Baum T, Ihlow A, Schweizer P, Seiffert U. Microphenomics for interactions of barley with fungal pathogens. In: Genomics of plant genetic resources [internet]. Dordrecht: Springer; 2014. p. 123–48. [cited 2017 Aug 4]. Available from: http://link.springer.com/chapter/10.1007/978-94-007-7575-6_5.
23. McCouch SR, McNally KL, Wang W, Sackville HR. Genomics of gene banks: a case study in rice. Am J Bot. 2012;99:407–23.

Part II

Bioinformatics Analysis of Genomic and Proteomic Sequences

Databases and Bioinformatics Tools for Data Mining

6

Pallabi Pati, Sushil Kumar Rathore, and Manoj Kumar Gupta

Abstract

Data, information, and knowledge play an interesting role in human life. Huge repositories of data generated because of the recent development of technologies demand the development of novel tools and techniques that can retrieve more important information. Data mining is a kind of knowledge discovery technique that extracts useful information from heterogeneous biological data by employing various machine learning, artificial intelligent systems, and decision-making techniques. Thus, in this chapter, the authors attempted to understand how data mining approaches have revolutionized biological research. The topic of data mining is discussed in brief, including the application it has in bioinformatics. This chapter also illustrates some of the emerging problems and opportunities in data mining in bioinformatics by utilizing this analogy.

Keywords

Data Mining · Bioinformatics · Databases · KDD

P. Pati (✉)
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_6

Abbreviation

ANN	Artificial neural network
BLAST	Basic local alignment search tool
DBMS	Database management system
DDBJ	DNA databank of Japan
DNA	Deoxy Ribonucleic acid
EMBL	European molecular biology laboratory
ESTs	Expressed sequencing tags
GOLD	Genomes Online Database
INE	Integrated Rice Genome Explorer
INSDC	International nucleotide sequence database collaboration
IRGSP	International Rice Genome Sequencing Project
KDD	Knowledge discovery in database
KNN	K Nearest Neighbor
NCBI	National center for biotechnology information
PDB	Protein databank
RGP	Rice Genome Research Program
RNA	Ribonucleic acid
TBP	TATA box binding protein
TIGR	The Institute for Genomic Research
UCEs	Upstream control elements
VEP	Variant Effect Predictors

6.1 Introduction

The digital revolution has made it easy to record, process, store, distribute, and share digitized information. With major developments in computing and related technology and their ever-expanding use in various walks of life, vast quantities of data of various features continue to be gathered and processed in databases. If the amount of data in the world doubles every 20 months, it is possible that the size and number of databases will rise at a similar rate. Thus, it is really a challenge to discover information from this huge amount of data. Data mining is an effort to make sense of the abundance of knowledge embedded in this large data volume [1]. Data mining uses various techniques, models, or algorithms to analyze the vast amount of data stored in the databases. In data mining, a pattern or rule is discovered that helps in establishing a hidden relationship between variables. The key objective is to manipulate the computer's data processing ability with the capacity of humans to interpret patterns [2].

In 1989, Piatetsky-Shapiro coined the phrase "knowledge discovery in database" (KDD). Application of data mining and KDD is inevitable due to the huge size of data collected from different sources and difficulty in handling and analyzing these data manually. KDD can be seen as an inclusive method of extracting useful information from information, while data mining could be defined as the core of KDD, which involves algorithms that discover unknown patterns of data, construct models, and discovery [3]. In the last two decades, the development of various data mining techniques in various fields such as artificial intelligence, machine learning, soft computing, and statistics has led researchers to create and apply new data mining methodologies. Data mining generally works on information stored in the database, which may be interrelated or relevant and inconsistent and irrelevant sometimes. Thus, it requires an application that allows the administrator of that data to manage it in order to exploit and control the necessary data. Maintenance and manipulation of the database is known as database management systems (DBMS) [4]. For instance, “Integrated Rice Genome Explorer” (INE) is a database that helps us to integrate genetic as well as information regarding physical mapping with the genome sequences generated by the collaboration with the “International Rice Genome Sequencing Project” (IRGSP). These Databases contain a various tools to analyze and compare the genomic databases of rice and maize. This system is also very much helpful in the development of different kinds of grass crop species. This kind of comparative genomics analysis helps us gather updated knowledge regarding total structural and functional data of various basic plant genome, which will lead to a better era in the field of biological research by focusing on bioinformatics tools techniques. Considering the above information, in the current chapter, the authors tried to highlight the basics concept of databases and their standards and the benefits of DBMS. It then describes the principle of data mining and how data mining processes are useful in biological research.

6.2 The Databases Concept

Databases are having a significant influence on the increasing usage of computers. It is certainly fair to say that databases are in use in most areas where computing is required, including business, engineering, medicine, law, education, and library [4]. The word database is used so often that to define it, and we must define the concept of a database. A database is a list of many related data. By data, we mean factual information that can be recorded and that has implicit meaning. Consider the names, addresses, and phone numbers of your friends. You may have recorded this data on a personal computer and stored it in a database using a database management application, like Microsoft EXCEL or ACCESS. This is a set of information with a predetermined context and therefore is a data set. The previous concept of the database was very generalized; for instance, consider the list of terms this page includes as similar data. Therefore, this page may serve as a database. However, nowadays, the term database is generally used more specifically. A database should have an implicit set of properties [4].

A database can reflect few parts of the natural world, called the mini world, or the universe of discourse (UoD). The database reflects changes within the mini world. A database is a systematically organized data that holds some meaning which cannot be altered or deleted by the database administration. A random collection of data, even if very large, cannot be called a database. A database is created, designed, and populated with data for a specific purpose. It has both an intended group of users and some preconceived applications in which they may be interested. Thus, a database has a “source” from which details are collected (collection of records, memories, etc.), “any degree of contact” with events that “happening in the real world” (i.e., events that are true), and its information “used by an audience that is actively involved” (i.e., its users who frequently view, change, and erase the information contained in the database). For the purpose of regulatory compliance, a database can be of any size and complexity. On the one hand, it is likely some database may be limited (say, no more than a few hundred records). For example, a collection of names and addresses with a basic layout like: “John Smith” and “Kolkata”. On the other hand, there may be on the order of half a million books in large libraries, in different categories such as by author’s last name, subject, and book title. Each book could be arranged in alphabetical order within its respective category [4].

A database may be developed and managed either by anyone manually (by a physician), or by computer. Because the library card catalog is a manually edited database, it is an example of a database where someone can make mistakes. A computerized database may be established and sustained either by the services of the application program(s) designed for that purpose or by the services of a DBMS. Here error is minimal in comparison to a manual database. A DBMS allows users to build and retain databases that help us to keep track of items and arrange them logically for easier access to knowledge. As a comprehensive software system, the DBMS is a general-purpose system that facilitates the processes of defining, constructing, and manipulating the database for multiple purposes [4].

The concept of a database includes defining the types of data, its composition, and the rules for which data can be applied to the database. To build a database, first data is stored in some storage medium controlled by the DBMS, and then an interface is written to access this data. A database may be manipulated by making series of queries against a database to obtain specific data, updating the database to reflect changes in the mini-world, and generating reports from the data. In order to define a database, a programmer must first determine the data type for the data to be stored, then the configuration of the data, then what is the output data type of the data will be in the data structure, and finally, the configuration for that output data type is taken into consideration. The process of building the database is the process of storing the data itself on some controlled medium, like a hard drive or flash drive. In order to exploit a database, it needs to be queried to acquire requested data, which then needs to be modified to represent improvements within the mini world, which needs to be queried again to obtain more knowledge about the mini world and eventually needs to be used to generate a report or other kind of result data [4]. A general-purpose DBMS is not necessary use to implement a computerized database. We might compose our series of programs to build and preserve the database, in essence, developing our special-purpose DBMS applications (software that handles databases). Regardless of whether we use a general-purpose database system or not, we typically have to recruit a significant amount of software to manipulate the database [4].

6.3 Advantages of DBMS

DBMS allows end-users to build, view, edit, and erase data. It is a layer that links programs and data. Compared to the File-Based Data Management System, DBMS is a superior management application (https://www.tutorialspoint.com/). Few important advantages of DBMS are reducing data redundancy, sharing of data, data integrity, data security, privacy, backup and recovery, and data consistency. The file-based data storage systems spanned several files, each located in several separate places in a system and sometimes residing in another device in several locations. Due to this redundancy, multiple copies of the same file can lead to data redundancy. In a database, we can set a password in an encrypted way and preventing anyone from accessing our database so that they cannot alter our database's setup. As a result of this, there is no chance of encountering duplicate data. Users of a database may also share the data among themselves. However, there are multiple kinds of authorization to access the data. As with multiple layers of security, the data can only be disclosed depending on the class of authorizations. Remote users, who are working together, can also access the database at the same time, and they can also share the data they are looking at within the database.

Data integrity ensures data reliability and consistency. Data integrity is very important because multiple databases are stored on a single database server. All these databases contain information that is either visible to a lot of people or is connected to a lot of people. In order to ensure that the data used are accurate and consistent, it is essential for the data to be verified on multiple sources and it get exploited by different predefined users. Data security is a critical principle of database management. Only approved users should be given entry to the framework, and there should be a username and password for each authorized user. Unauthorized users are not permitted to access the database under any circumstances because it is highly illegal and against the security policies.

According to the laws, the privacy law in a database ensures that the approved users can only read, change, or erase the database's data. The information about the database is given and, once it is given, can be seen only by the user with the necessary authority. For example, some sites (like Facebook) require only one account and a certain password to create an account for a particular user. Others have their own username and password. A database management framework automatically also takes control of replication and recovery. The DBMS will remember all necessary changes and informs the user when it is time to back up their data. As well, in case of a system error or crash, it restores the database to the same state it was before the error occurred. The software configuration also guarantees that the anomalies cannot exist and that data redundancy is not a problem. To make some claims, all data will appear consistently across the database, and all users viewing the database will agree on all of it. It is not just an effective storage structure, but the versatility is still fantastic. If any improvements are made to it, they automatically go to all the users, and there is no data discrepancy.

6.4 Knowledge Discovery and Data Mining

The conventional way of converting data into information focused on doing manual review and evaluation by a domain specialist in order to identify valuable trends in data for decision support. For instance, early the work of Reeder & Feller employing methods was crucial in diagnosing and treating fever [5]. In 1996, this process was described by different steps starting from data selection, preprocessing, data transforming, data mining, and interpretation [6]. Data selection involves the previous know-how of and target of the application. The selection of a dataset or a subset of variables is made through ranking via selection technique [1, 7]. Data pattern processing is needed to enhance the actual data quality for mining. This also enhances the productivity of mining by decreasing the data processing time. Data preprocessing requires data cleaning, data transformation, data integration, compact representation, data reduction or data compression, etc. Data cleaning comprises operations like normalization, noise elimination, and missing data handling, redundancy reduction, etc. Real-world data is frequently erroneous, incomplete, and contradictory, likely due to technical errors or defects in the implementation of the system. It is important to clean up such low-quality data before data mining. Data integration plays a significant role. This operation involves the integration of various heterogeneous datasets created from various sources. Reduction and projection of data involve identifying useful features to represent the data (depending on the objective of the task) and using methods of reduction of dimensionality, discretization of features, and extraction (or transformation) of features. The application of data compression principles can help in data reduction, which has potential in the future to grow, especially in the field of multimedia dataset knowledge discovery. Data mining mainly involves classification, regression, clustering, description, image retrieval, the discovery of association rules and functional dependencies, rule extraction, etc. Interpretation deals with the deduction of patterns found and the possible visualization patterns of extracted information. To classify the genuinely interesting or useful patterns for the user, one may evaluate the extracted patterns automatically or semi-automatically. Using discovered knowledge, we integrating earlier generated knowledge into the performance system and taking knowledge-based acts [1, 7].

Thus, data mining is basically a subset of Knowledge Discovery. While the original notion was "Knowledge Discovery in Databases" (KDD), nowadays, in order to emphasize that Data Mining is an essential part of the knowledge discovery method, the current most common notion is "Knowledge Discovery and Data Mining" (KDD) (Fig. 6.1). It is important to note that KDD (knowledge discovery and data mining) is not simply a process but also encompasses the complete value-added chain from the data's extremely physical side to the very human side of knowledge—i.e., the latter characterized from a cognitive point of view: knowledge

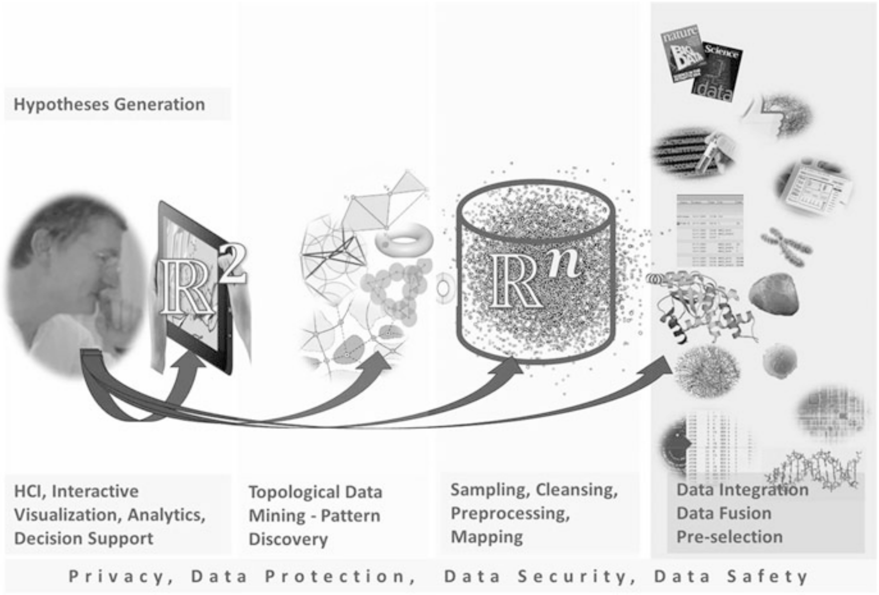

Hypotheses Generation
$\mathbb{R}^2$
$\mathbb{R}^n$
HCI, Interactive Visualization, Analytics, Decision Support
Topological Data Mining - Pattern Discovery
Sampling, Cleansing, Preprocessing, Mapping
Data Integration Data Fusion Pre-selection
Privacy, Data Protection, Data Security, Data Safety

Fig. 6.1 The general knowledge discovery process that is widely employed in the life sciences (Adapted from [7]

as a set of expectations [1, 7]. Recently, Holzinger describes the novel technique that extends the original definition by Fayyad and the team [6] by having an actual human make the decisions. As core theories of human-computer interaction, HCI & KDD, together with a novel approach, aims to bring all two together into this research project to advance knowledge in a specific context [7]. The core principle of HCI-KDD is to allow end-users to identify and classify previously unknown and potentially valuable and accessible knowledge. It is defined as the process of identifying new data patterns from unstructured data. The goal, in this case, is a visualization of data that was previously unseen. There is a specialist in the framework with clear domain expertise. By allowing them to interactively explore their data sets, they may be able to recognize, interpret, and appreciate valuable details, acquire new, and previously unknown information [7].

6.4.1 Data Mining

Initially, the data mining technique was widely used in economics. But nowadays, several data mining techniques are also used in the field of agricultural research. It helps to improve the prediction of the yield of the specific variety of cultivation that will be very much beneficial for the plant. As we all know the agricultural production largely depends on climatic condition, soil type, irrigation method, cultivation strategies, data mining will help us to predict the best dependencies or fitting model, which in turn may help us in increasing the grain yield. Typically, a data mining algorithm comprises three major components, namely, the model, preference criterion, and search algorithm [1]. A model includes parameters that must be calculated using a specific representational type or tool from the data for the chosen function. The preference criterion is the preference, based on the data given, for one model or set of parameters over another. The criterion is generally some sort of the model's goodness-of-fit function to the data, perhaps modified by a smooth term to prevent overfitting or to generate a model with too many degrees of freedom to be limited by the data given. The search algorithm is an algorithm for finding particular models or patterns and parameters, provided the data, model(s), and criterion of preference. The model-preference-search components are typically instantiated by a particular data mining algorithm [1]. Tasks for data mining are divided into two key categories: predictive and descriptive. Six key functions of data mining are described classifying, regression, clustering, modeling of dependencies, variance detection, and summarizing [6]. Classification, regression, and anomaly detection are categorized under the predictive category, whereas clustering, dependency modeling is categorized under the descriptive category. Predictive model forecasts use certain variables in the dataset to predict unknown values of other related variables while descriptive model classifies patterns or relationship and utilizes human-understandable pattern and trends in data [3].

Classification is part of the classical methodology of data mining that is based on machine learning. In a database, it finds mutual properties among a set of objects and categorizes them according to the classification model into diverse groups. Its primary objective is to scrutinize the training data and construct an accurate definition or model for each class using the data function. Statistical techniques such as decision trees, neural networks, and statistics are used in this process [3]. Regression describes the relationship between dependent and independent variables. Prediction is reached by endorsing regressions. Statistical regression is a mathematical model that relates the values of the dependent variable to the values of the other predictor or independent variable. The predicted variable in regression could be a continuous variable. Real-valued prediction variables in regression are mapped from elements of a learning function. Some of the widely used regression techniques include statistical regression, Neural Network, Support Vector Machine regression. More complex methods could also be used to predict future values, such as logistic regression, decision trees, or neural networks, and these techniques could also be combined to achieve better results [3].

Clustering is a technique of data mining that groups physical or abstract objects into related object classes. Clustering is a technique of dividing data sets (records/tuples/objects/samples) into multiple groups (clusters) based on predetermined similarities. The main objective of clustering is to find affinity-based groups (clusters) of objects so that there is a great resemblance to each other within individual clusters, while clusters are sufficiently diverse from each other. Clustering is a type of unsupervised learning in machine learning terminology [3]. Dependency Modeling (or association rule mining) is one of the finest recognized data mining techniques and is categorized under an unsupervised data mining technique that seeks to identify links or relationships between items or records belonging to a large dataset and identifies significant variables dependencies [3]. Anomaly detection is synonymous with the uncovering of the most significant changes or aberrations from standard behavior [3]. Although not part of data mining techniques, the summary is the result of these techniques and deals with the determination of a compact representation for a subset of data synonymously referred to as generalization or description [3]. Sequential Pattern is used over a business cycle to determine sequential patterns or associations or periodic events/trends between variable data fields [3].

6.4.2 Data Mining Architecture

The architecture of data mining can be mainly classified as below:

Knowledge Base It acts as the start of the entire process of data mining. It serves as a guide for looking for the resulting trends or evaluating their interestingness. This form of information can involve hierarchies of concepts that organize attributes or values into different abstraction stages.

Data Mining Engine It forms the main element of the mining framework, consisting of all the modules required to perform data mining tasks, such as characterization, prediction, cluster analysis, outlier analysis, and evolution analysis.

Pattern Evaluation Module This module is generally correlated with measures of interest. In order to stay focused on looking for interesting trends, it persistently interacts with the data mining engine. Many times, depending on the data mining method used, it uses thresholds to sieve out discovered patterns or can use the pattern evaluation module incorporated with the mining module.

User Interface The module acts as a link between users and the framework for data mining. It makes it simple and effective for users to communicate with the system without thinking about the convolutions behind the operation.

Data Sources (Www, Data Warehouse, Archive, Other Repositories) These are the actual data sources, and for efficient data mining, a huge amount of historical data is needed. In databases or data centers, companies usually store data. Often the data warehouse includes more than one database or text file, or spreadsheet. Another big source of data is www.

Server Database or Data Warehouse Includes concrete data that are set to be retrieved. Its main duty is to retrieve data at the request of users.

Other Processes Data must be cleaned and merged before it is passed on to the data warehouse server, as data are obtained from different sources and are in different formats such that it cannot be used directly for mining processes. The data need to be cleaned, integrated, and it is only important to pick and move on the secure data to the data warehouse server. Numerous techniques for cleaning, integration, and selection may be needed for the operation [3].

6.5 Databases for Biological Data Mining

In recent decades, a huge number of genome-scale experimental data sets have been made available. Thus, for storing and analyzing, several biology databases have appeared online. These databases can be classified according to data form, data processing techniques, data coverage scope, and database accessibility. These databases contain a wide range of data ranging from the genome of model and nonmodel plants (https://asia.ensembl.org/index.html) to protein information (https://www.rcsb.org/) (Table 6.1).

6.5.1 Databases for Genes, Genomes, and Variations

While breeding has been effective, the method of choice for farmers remained traditional, e.g., for studies contrasting two genes, A and B, in a test plant. With the aid of genomics and new sequencing techniques, scientists can study the underlying genetic makeup of plants, and these findings are helping us figure out how plant breeding contributes to the development of desired traits. Even though it is still in its infancy, Next-Generation Sequencing (NGS) technologies are permitting the mass sequencing of genomes and transcriptomes to produce a vast amount of genomic information. By means of bioinformatics technologies, the NGS data analysis, as evidenced by the huge collections of markers, allows the new genes and sequences discovery and the location and arrangement of their occurrence on the genome. By studying the gene expression level of a variety of breeds, breeders get an understanding of the molecular basis of complex traits. Genome-wide association studies, or GWAS, include TILLING and Eco-mutation in genome sequencing technologies, which can make it possible to scan mutant as well as germplasm collections for allelic variants in target genes [8]. It is very useful to re-sequence an organism’s genome more than once in order to find markers that can be used in high-throughput genotyping platforms like SNPs and SSRs, or the construction of a genetic map. These tools and resources make it much easier to study genetic diversity, which is important for maintaining germplasm, enhancing, and application. Also, they can be used to help identify some of the genes in those regions of the genome that might also be involved with the disease, and they can be used to find markers linked to those genes as well. New markers for quantifiable characteristics based on DNA, such as the ones mentioned above, are employed for marker-assisted selection, including breeding by design, marker-assisted backcross selection, and genome selection. Thus, advances in genomics provide breeders with novel tools

Table 6.1 List of few important Biological databases

Database Name	Description	Species	Link
BAR (bio-analytical resource for plant biology)	Provides a user-friendly interface for the exploration of gene expression data	Several plant species including *O. sativa*	http://bar.utoronto.ca
CoP database	Microarray data based integrated database for co- expressed genes and biological processes in plants	*Arabidopsis thaliana*, *Vitis vinifera Glycine max*, *Oryza sativa*, *Populus trichocarpa*, *Hordeum vulgare*, *Triticum aestivum*, and *Zea mays*.	http://webs2.kazusa.or.jp/kagiana/cop0911
CSRDB (cereal small RNA database)	Consists of large maize and rice datasets smRNA sequences provided by high performance pyrosequencing	*O. sativa* and maize	http://sundarlab.ucdavis.edu/smrnas
CyVerse (former iPlant collaborative)	Provides a strong computing platform allowing massive datasets and complex research to be discovered using the data	Plants, animals, and microbes	http://www.cyverse.org
DDBJ updated on daily bases	A repository of nucleotide sequence data	*O. sativa* and several organisms species	http://www.ddbj.nig.ac.jp
Diurnal	An internet-based site to keep track of diurnal and circadian genome wide expression profiles from results of model plants	Plant species	http://diurnal.mocklerlab.org
DroughtDB	Manually curated genes associated with stress response to drought	*O. sativa ssp. japonica cv. Nipponbare Zea mays*, *Arabidopsis thaliana*, *Sorghum bicolor*, *Hordeum vulgare*, *Brachypodium distachyon*, *Solanum lycopersicum*, *Secale cereale, and Aegilops tauschii*.	http://pgsb.helmholtz-muenchen.de/droughtdb
EMBL	Comprehensive compilation and annotation of nucleotide sequences	*O. sativa* and several organisms species	http://www.ebi.ac.uk/about

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	from available public sources		
Ensembl plants	Provides numerous genomic data sets and analysis and visualization tools for several plant species in the genome browser	*O. sativa* and other organism species	http://plants.ensembl.org/index.html
ExPath	It offers data on metabolic pathways inferred from transcriptomic data based on microarrays, gene annotation, and orthologous genes	*Oryza sativa, Arabidopsis thaliana, and Zea mays*	http://expath.itps.ncku.edu.tw
FamNet	Enables the user to retrieve data from one or more plant species linked to preserved structural-functional domains within proteins	Arabidopsis, *Oryza sativa*, *Medicago truncatula*, *Populus tremula*, *Hordeum vulgare*, *Glycine max*, *Nicotiana tabacum*, and *Triticum* spp	http://www.gene2function.de/famnet.html
Galaxy	A software framework that allows experimentalists to conduct complex large-scale research with only a web browser without informatics or programming skills		http://galaxyproject.org
GenBank updated on daily basis	NIH genetic sequence database, a repository of publicly available DNA sequences	*O. sativa* and other organism species	http://www.ncbi.nlm.nih.gov
Genevestigator	Provides powerful tools to explore gene expression across a wide variety of biological contexts	Arabidopsis, *Oryza sativa*, *Medicago truncatula*, *Populus tremula*, *Glycine max*, and *Triticum* spp	https://genevestigator.com/gv
Gramene	An open data resource for comparative functional genomics in cereals and other plant species	O. sativa and other plant species	http://www.gramene.org
GRASSIUS (grass regulatory information services)	It consists of a series of databases relating to the regulation and interaction of gene expression in grasses	*Zea mays, Oryza sativa, Saccharum* spp., and *sorghum bicolor*	www.grassius.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	with agronomic features. Includes transcription factors, promoters, transcription and co-regulators Factor-clones ORF		
GreenPhylDB	The database having catalog of gene families from various green plants	*O. sativa* and other plant species	http://www.greenphyl.org/cgi-bin/index.cgi
IsomiR Bank	Integrated resource that contains the sequence and expression of isomiRs	*Arabidopsis thaliana*, *Danio rerio*, *Homo sapiens*, *Mus musculus*, *Oryza sativa*, *Drosophila melanogaster*, *Zea mays,* and *Solanum lycopersicum*.	http://mcg.ustc.edu.cn/bsc/isomir
Mercator pipeline	Functional annotation of plant "omics" data	Arabidopsis, Chlamydomonas, rice	http://mapman.gabipd.org/web/guest/app/Mercator
MoChA ("molecular characteristics database for allergens")	Database of allergenic proteins acquired by bioinformatics methods or proof of binding to IgE. It has obtained accurate experimental genome, transcriptome, proteome data, and molecular properties	2000 organisms	http://lilab.life.sjtu.edu.cn:8080/mocha/main-7.9-2.html
MPIC ("mitochondrial protein import components") database	Searchable details on plant and nonplant mitochondria protein import equipment	*O. sativa* and 23 other organism species	http://www.plantenergy.uwa.edu.au/applications/mpic
NIASGBdb ("National Institute of Agrobiological sciences planttfdb database")	A database having information on simple sequence repeat (SSR) polymorphisms in plant genomes	*O. sativa* and other plant species	http://www.gene.affrc.go.jp/databases_en.php
OryGenesDB	A rice reverse genetics database, created with flanking sequence tags of different mutagens and data on functional genomics	*O. sativa ssp. indica and japonica*, and two other plant species	http://orygenesdb.cirad.fr/index.html

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
PDB (protein data Bank)	Worldwide archive of structural data of biological macromolecules	*O. sativa* and other organism species	http://www.rcsb.org/pdb
Phytozome	An annotated plant genome and gene familial data comparison center. Provides an overview of each plant gene's evolutionary history at the level of sequence, gene structure, gene family, and genome organization [70]	*O. sativa* and 64 other plant and algae species	http://www.phytozome.net
PLANEX (PLAnt co-expression) database	Have publicly available GeneChip data received from the gene expression omnibus	*Arabidopsis thaliana, Hordeum vulgare, Glycine max, Vitis vinifera, Oryza sativa, Triticum aestivum, Solanum lycopersicum,* and *Zea mays*	http://planex.plantbioinformatics.org
PlantAPA (alternative polyadenylation)	A internet based server for query, visualization, and analysis of poly (A) sites in plants, helping in profiling various cleavage sites and quantify expression pattern of poly(A) sites across different conditions	*Oryza sativa, Chlamydomonas reinhardtii, Medicago truncatula*, and *Arabidopsis thaliana*	http://bmi.xmu.edu.cn/plantapa
PlantArrayNet	Information on co-expressed genes using microarray-transcriptomic data	Rice, Arabidopsis, and *Brassica rapa*	http://arraynet.mju.ac.kr/arraynet
PlantDHS (plant DNase I hypersensitive site database)	Incorporate histone modification, transcription factor binding sites, RNA sequencing, genomic sequence, and nucleosome positioning/occupancy	*Arabidopsis thaliana, Oryza sativa, and Brachypodium distachyon*	http://plantdhs.org
PlantGDB	A database of sequence data from different plant species	*O. sativa* and other plant species	http://www.plantgdb.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
Plant homolog database	A database comprised of plant homologous genes	16 plant sp. Including 10 *Oryza* species	http://phd.big.ac.cn
Plant MPSS (massively parallel signature sequencing) databases	Information on the expression status of genes, and potential unique transcripts (antisense transcripts, alternative splice isoforms, and regulatory intergenic transcripts)	Grape. Arabidopsis, *Magnaporthe grisea,* and rice	http://mpss.udel.edu
Plant-PrAS (plant protein annotation suite) database	Database of properties related to physicochemical and structural information, and unique functional region in plant proteomes	*Arabidopsis thaliana, Glycine max, Populus trichocarpa, Oryza sativa, Physcomitrella patens,* and *Cyanidioschyzon merolae*	http://plant-pras.riken.jp
Planteome	For plant and species-specific crop ontologies, a resource for popular reference ontologies. It also provides ontology-based rice gene annotation, QTLs, phenotypes, and germplasms	*Oryza and plant species*	http://www.planteome.org
PlantRNA	Assembles transfer RNA (tRNA) gene sequences obtained from fully annotated plant nuclear, plastid, and mitochondrial genomes	Five flowering plants (*Oryza sativa*, *Arabidopsis thaliana*, *Medicago truncatula*, *Populus trichocarpa*, and *Brachypodium distachyon*), a moss (*Physcomitrella patens*), two green algae (*Ostreococcus tauri and Chlamydomonas reinhardtii*), a pennate diatom (*Phaeodactylum tricornutum*), one glaucophyte (*Cyanophora paradoxa*), and one brown alga	http://plantrna.ibmp.cnrs.fr

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
		(*Ectocarpus siliculosus*).	
PLEXdb	A single resource of gene expression for plants and plant pathogens. It is a phenotype genotype, hypothesis building knowledge warehouse, leveraging highly parallel expression data to associated genetic, physical, and pathway data with seamless portals	*Oryza, Vitis*, maize, *Fusarium graminearum*, *Arabidopsis*, soybean/ *Phytopthora*/soybean cyst nematode, *Brachypodium*, cotton, poplar, Citrus, tomato, and *Medicago*	http://www.plexdb.org
PlnTFDB (plant transcription factor database)	A web interface to navigate various plant species' broad sets of transcription factors. Information is given for each family, including protein sequences, coding regions, genomic sequences, expressed sequence tags (ESTs), domain architecture, and scientific literature	*O. sativa ssp. indica* and *japonica* and other plant species	http://plntfdb.bio.uni-potsdam.de/v3.0
PmiRKB (plant miRNA Knowledge Base)	Information available for four major functional modules- "SNPs", "Pri-miRNAs", "MiR—Tar", and "self-reg"	*21 O. sativa* and Arabidopsis	http://bis.zju.edu.cn/pmirkb
PMRD (plant MicroRNA database)	A plant miRNA data repository containing associated sequence information, secondary structure, target genes, miRNA expression profiles, and their mapping to the browser of the species-specific genome	*O. sativa* and 120 plant species	http://bioinformatics.cau.edu.cn/PMRD
PO (plant ontology)	Robust and flexible controlled vocabulary that accurately represents the biology	*O. sativa* and other plant species	www.plantontology.org

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
	of plant structures and stages of development		
PODC (plant omics data center)	A repository for expression data of annotated gene and omics data analysis tools	*Arabidopsis thaliana*, *Glycine max*, *Medicago truncatula*, *Nicotiana tabacum*, *Oryza sativa*, spreading earthmoss (*Physcomitrella patens*), tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), sorghum (*Sorghum bicolor*), grape (*Vitis vinifera*), corn (*Zea mays*)	http://bioinf.mind.meiji.ac.jp/podc
POGs (putative orthologous groups 2) database	A database that combines data from, Arabidopsis, rice and maize into "putative orthologous groups" (POGs) and permits comparisons among orthologs and extrapolation of annotations among species.	*Arabidopsis thaliana*, *Oryza sativa,* and *Zea mays*	http://pogs.uoregon.edu
Ppdb (plant promoter database)	Information available on Y patches, regulatory element groups (REGs), transcription start sites (TSSs), and core promoter structure (TATA boxes, initiators, GA and CA elements)	*Arabidopsis thaliana,* poplar*, Physcomitrella patens,* and *Oryza sativa.*	http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi
STIFDB2 (stress responsive transcription factor database)	Group of responsive genes for biotic and abiotic stress with options to detect possible transcription factor binding sites in their promoters. The data have been characterized by an integrated biocuration and genomic data mining approach	*O. sativa ssp. japonica* and *Indica* and Arabidopsis	http://caps.ncbs.res.in/stifdb2

(continued)

Table 6.1 (continued)

Database Name	Description	Species	Link
UniProtKB	A central annotated protein resource consisting of two sections: UniProtKB/ Swiss-Prot for annotated manual entries and UniProtKB/ TrEMBL for annotated computer entries	*O. sativa* and other organism species	http://www.uniprot.org

and methodologies that permit a great leap forward in plant breeding, including the genetic dissection and breeding for complex traits and super domestication of crops.

The algorithms and methods used to store and process genomic data created by various technical platforms will rely on what kind of data is being used and what outcome is predicted. The 3000 genome project (http://iric.irri.org/resources/3000-genomes-project) and 1001 Arabidopsis genomes (http://1001genomes.org/) are good examples of why a genetic interface is needed to help breeders with the information they need. Once the information is obtained, results are made available to the breeders [9]. Attracted by the fame and glory of the invention of the global web page, a common and often successful approach to providing the information is through a web page that can be easily browsed. Several extensive bioinformatics resources exist to help scientists study plant and human genetics, like GenBank (http://www.ncbi.nlm.nih.gov/genbank/), the European Bioinformatics Institute (http://www.ebi.ac.uk/), and the Swiss-Prot database (http://expasy.org/sprot/). These above databases are dedicated to storing knowledge for all organisms, although many other more basic databases based on species of importance to the breeders still exist, including the Gramene (http://www.gramene.org/), Sgn (http://solgenomics.net/), Phytozome (http://www.phytozome.net/), which contain information that may have more specific usage for breeding programs. For instance, the "MSU Rice Genome Annotation Project" (http://rice.plantbiology.msu.edu/), the International Rice Genome Sequencing Project (IRGSP) [10], RAPdb [11], and the Oryza Genome Evaluation project [12] are primarily providing assembly, annotation, and related information of rice genome. These genomes are provided by constructing a built-in web resource for rice, including a rice species-specific genome explorer, whole-genome alignment, synteny, genetic and physical maps with genes, gene trees, ESTs and QTL positions, genetic diversity data including SNPs, and advises them on their genome sequence [13].

6.5.2 Databases for Gene Expression Datasets

With the invention of the microarray in the 1980s, it became possible to measure the abundance of all transcripts at the genomic scale. This is now known as the transcriptome. To this date, several gene expression data from such experiments have been stored in public repositories, like the EBI ArrayExpress (AE; https://www.ebi.ac.uk/arrayexpress/) and NCBI Gene Expression Omnibus (GEO; https://www.ncbi.nlm.nih.gov/geo/), after the implementation of the "Minimum Information About a Microarray Experiment" (MIAME) standard [14]. Unlike the International Nucleotide Sequence Databases (http://www.insdc.org/), these two databases, namely AE and GEO, for gene expression have not been sharing data with each other. There have been several instances where AE had started importing GEO data in the past but have recently stopped doing so. Though we still have access to all archive data of AE in GEO, all the new data are not available for us anymore [14]. Thus, at present, researchers operating on a specific topic would need to scan both of the databases since these databases have been independently maintained. Besides that, the DNA DataBank of Japan (DDBJ) recently started another repository for the investigation of gene expression called Genomic Expression Archive (GEA; https://www.ddbj.nig.ac.jp/gea/). The GEA is a repository of functional genomics data such as genotyping SNP arrays, epigenetics, and gene expression. Genomic or DNA microarray data and sequence-based data are acceptable in the MAGE-TAB format, in strict compliance with MIAME and MINSEQE guidelines, respectively [15]. As a consequence, there is a need for the integration of these public gene expression databases. Recently, Bonon, therefore, developed an index of public gene expression databases called All Of the gene Expression (AOE). Thus, he used a database of all gene expressions (All Of the gene Expression) to get a clearer idea of the average amount of genes in a community of employees (AOE). The aim of AOE is to compile and bring together all of the gene expression results and make them all searchable. He has been maintaining the AOE website for 5 years, and it has been helpful for pursuing functional genomics studies [14].

6.5.3 Database for Gene-Interactomes, Pathways, and Ontologies

A gene interaction network is the collection of genes, each linked by an edge indicating a functional relationship between these genes. These edges are called interactions because the two genes are assumed to have either a physical connection with their gene products, e.g., proteins, or one of the genes changes or influences the function of another gene of interest [16]. "The functional products of genes, e.g., proteins, work together to achieve a particular task, and they often physically associate with each other to function or to form a more complex structure. These interactions can be long-lasting, such as forming protein complexes, or brief, when proteins modify each other such as the phosphorylation of a target protein by a protein kinase. Since these interactions are important to carry out most biological processes, knowledge about interacting proteins is crucial for understanding these biological functions, which can be easily done via studying networks of these interactions" [16].

There are other, more complicated genetic variations. So, when all of these gene variants work together, the resulting influence does not manifest itself in just one gene alone. Moreover, it does not manifest itself at all. At high throughput, we can also measure the gene combinations to help further understand this disease. There are two general categories of such interactions: synthetic lethal (Synths) interactions and suppressor (Syns) interactions. The effect is lethal as a result of two nonessential genes combining to form lethal effects, and suppressive effects occur when a lethal variance within one gene "cancels out" or is "negated" by that of another gene. Much more research needs to be done on how drugs act in the human body. This way, we can understand how they work and use this knowledge to prevent or treat diseases [17, 18].

With the involvement of high-throughput methodologies like co-immunoprecipitation followed by mass spectrometry, yeast two-hybrid (Y2H), or tandem affinity purification, studies have been performed, which help to classify physical protein-protein associations for a wide variety of species. The fundamental genetic mechanism of drug action was mapped, and its influence on molecular pathways important in many biological systems, both in humans and for organisms [16]. The expansion of the number of proteins and how they "interact" has continued over the last decade. This has led to the creation of public databases that can be shared among scientists. As in predicting enzyme-specific interactions, computational techniques are used to forecast protein-protein interactions. At some point, the use of genomic data will help us understand the complicated relationship between protein pairs [19, 20] or help us predict novel interactions we have not yet experienced [21].

All the interactions performed in the lab are recorded so that they can be made accessible to the public at large. Scholars continue to be able to utilize various databases as multiple organisms' protein-protein interactions connect these new organisms with many other organisms. A database first gathered their samples from multiple sources. Nevertheless, biomolecular interaction databases like the International Molecular Exchange Consortium now allow researchers to compare protein-protein interactions from a wide taxonomic spectrum of species. Further, these databases agree to create publicly accessible datasets in standardized formats such as MITAB or PSI-MI XML 2.5. Currently, the databases recorded in MPIDB (http://www.jcvi.org/mpidb), DIP (http://dip.doe-mbi.ucla.edu), IntAct (http://www.ebi.ac.uk/intact), MINT (http://mint.bio.uniroma2.it/mint), Pact (http://mips.gsf.de/genre/proj/mpact), MatrixDB (http://matrixdb.ibcp.fr), BioGRID (http://www.thebiogrid.org), InnateDB (http://www.innatedb.com), and BIND (http://www.blueprint.org) are actively generating relevantly large numbers of relevant documents, and provide these through the "Proteomics Standard Initiative Common Query Interface" (PSICQUIC) service. This database can contain and hold interactions all intended towards a specific organism, such as the BioGrid (https://thebiogrid.org/) database, or it can contain and hold interactions specifically targeted at a specific biological domain, such as the MatrixDB (http://matrixdb.univ-lyon1.fr/) database. However, regardless of the file format that is used, this data is available in a standards-compliant, tab-delimited, and XML format. Presently, these databases share a lot of the same documents. We still have a lot of work to do to track the complicated relationships between patients and interactions. However, once the new data entry pipelines for each healthcare system are set in place, accurate reporting across an organization will be easier to track [16].

Understanding and visualizing the networks of these connections are also essential to researchers. Recent advances have been made in the types of software. There is software for different platforms, like Cytoscape (https://cytoscape.org/), Osprey [22], Pajek [23], etc. This software can display the network by employing a graph layout algorithm and will display the network layout attributes as nodes and are visual representations in each node (e.g., protein images, coloring). Moreover, through the usage of a number of various items, such as a plug-in and a filter, it analyzes these interactions and aids in the incorporation of external data sources like gene ontology [24, 25].

6.5.4 Databases for Gene Ontology

The Gene Ontology (GO) resource is the most straightforward and commonly utilized method available in terms of identifying the roles of genes. In GO, all functional knowledge is arranged as well as represented in a form amenable to computational analysis, which is essential for modern biological research. The GO database is organized using a formal ontology by specifying groups of genes and the connections between them. GO words (such as “GO:00086467”, “GO:00093381”, and “GO:00093385”) contain meanings that are sometimes stated as “equivalence axioms” (axioms saying that two terms are identical if they are both closely connected to the same things), since they can be computationally inferred utilizing rational reasoning. The GO framework has been carefully built over the span of 20 years by a small team of ontology developers; it is continually changing in reaction to recent scientific findings and consistently refined to reflect the most current state of biological understanding. The members of the ontology creation team include specialists on biological knowledge representation. They read the literature to validate the correctness of the representation and involve biocurators (those who study and curate biological knowledge) to collaborate alongside them to establish this representation of biological knowledge [26]. The Planteome seems to be a current database (www.planteome.org) which has provided a common collection of ontologies for use in genome, expression, and phenomic projects. Additionally, it provides ontology-based annotations for approximately 85 plant species, including a number of *O.sativa* subspecies *indica* and *japonica* rice as well as wild *Oryza* species.

6.5.5 Databases for Pathway

The word “pathway” is poorly described and may arguably, be used to characterize any sequence of action between biomolecules until a specific product is created. The Reactome project [27] is online, open access, database of biological pathways that has been curated. The reactions are ordered hierarchically, with a series of single reactions required at the lowest level and a succession of interconnected pathways at the top-level [28]. The data in Reactome are obtained from scientific literature, with information being collected by researchers, editors, reviewers, and curators. Subsequently, Reactome contains links to other databases such as Ensembl, UniProt, and KEGG [29].

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a database linking genetic, biochemical, as well as phenotypic details from various sources [30]. It provides knowledge regarding chromosomes and metabolic processes. The proteins and enzymes that belong to such pathways, along with information regarding genetic, molecular, and environmental mechanisms, diseases, and drug mechanisms. Many connections are given to various databases, including the UniProt database and the NCBI Entrez Gene database. Unfortunately, access to KEGG was discontinued in 2011, and we no longer have the ability to access KEGG via FTP. Instead, they only have the feature to access KEGG through their API (application program interfaces). It also inhibits system developers' ability to incorporate the KEGG pathway in the software.

WikiPathways is an open access initiative that is different from the other pathway databases [31]. The tool is part of MediaWiki and enables anyone to contribute to and manage biochemical pathways on the Wikipedia website [32]. WikiPathways encompasses several various signaling pathways involved in multiple biochemical processes over several organisms. WikiPathways is now a novel approach to preserving and processing vast amounts of genetic information in response to the public's desire to ensure and organize the data, thus ensuring its ultimate performance.

Ultimately, MetaCyc [33] is a massive systematic repository of pathways and enzymes among all aspects of existence, with the majority of evidence derived from current literature asserts that it is the most exhaustive set of metabolic pathways accessible. MetaCyc is purported to be the biggest array of curated metabolic pathways. No pathway database can ever be accurate, and lesser of around 10% of predicted genes or proteins can be mapped to a given pathway or reaction in some environmental samples [34]. Thus, it is common to use multiple complex databases and algorithms to get the conclusions that best fit the available data. Gene ontology and pathway enrichment analysis have been discussed in detail in Chap. 12.

6.6 Bioinformatics Tools in Data Mining

To date, several bioinformatics tools have also been developed which are used in various bioinformatics analysis, including sequence alignment (Chap. 7), gene identification and structure annotation (Chap. 8), phylogenetic analysis (Chap. 9), RNA structure prediction (Chap. 10), structural proteomics (Chap. 11), and gene ontology & pathway enrichment analysis (Chap. 12), high-throughput sequencing technologies (Chap. 13), DNA–Protein Interaction Analysis (Chap. 15), RNA–Protein Interaction Analysis (Chap. 16), SNP identification and discovery (Chap. 17), microsatellite markers discovery (Chap. 18), genome-wide association study (Chap. 19), expression profiling and discovery of microRNA (Chap. 20), identifying long noncoding RNA (Chap. 21), metagenomics (Chap. 23), and single-cell RNA sequencing (Chap. 25). Detailed information about each tool and its utility is described in detail in each chapter later.

6.7 Conclusion and Future Perspective

In conclusion, biological databases are life science knowledge collections collected from experimental observations, written literature, technologies for high-throughput experiments, and quantitative analysis. They provide information from study areas such as genomics, proteomics, metabolomics, microarray genes, and phylogenetics. While numerous databases and online resources for protein bioinformatics have been established to assemble and store numerous biological details, there are challenges as well as opportunities to build Next-Generation databases, including resources that facilitate the integration of data, generation of data-driven hypotheses, as well as exploration of biological information [35]. Effective storage and handling of vast quantities of data is the first obstacle that machine biologists would meet. Huge parallel disk technologies (file systems that are distributed, clustered, or parallel) were investigated, in addition to stronger hardware support. Lustre (http://lustre.opensfs.org) and "Hadoop Distributed File System (HDFS)" (http://hadoop.apache.org) are the best examples.

The collection and handling of information is only one side of the same coin. The goal of high-throughput omics studies is to translate clinical data into expertise in biomedical science and healthcare systems. We need accessible computing facilities and an effective data processing system to achieve precision medicine and improved therapies. Cloud computing appears like an inexpensive option for large-scale data processing relative to conventional HPC cluster computing. Bioinformatics research is also altering how the analysis is carried out by hosting cloud-based data storage with massive amounts of high-throughput data. Code is instead going to the data instead of transferring data to the application code. In addition, the performance of converting data into information often involves modern and powerful machine learning and data mining techniques, and analytical architectures. Apache Spark (http://spark.apache.org), for large-scale lightning-quick in-memory clustering computation, is a newly developed fast and general-purpose computing engine. It enabled a wide variety of higher-level software, along with data collection and organization, GraphX for graph processing, MLlib for machine learning, Spark SQL for SQL, and Spark Streaming for apps with scalable streaming. In Big Data analysis, the most difficult challenge is to cope with the data's variability, variety, and uncertainty and to find a better way to incorporate them. Along with analyzing the versatility of NoSQL technology, the implementation of ontology and Semantic Web technology is another exciting area. Ontology plays a perfect function in solving the issues of the variability of data sources as a systematic, a precise description of a commonly accepted conceptualization of a topic of concern. The rapid growth and acceptance of ontologies have helped the scientific community use structured ontologies to annotate and incorporate biological and biomedical data and automate the discovery and design of web resources and workflows for bioinformatics. Linked Data infrastructure offers a means to publish and interconnect organized knowledge on the internet. Bio2RDF [36] and the EBI RDF platform [37] are active Linked Data ventures in the area of bioinformatics. Through identifying a series of basic conventions to construct RDF(s) compliant Linked Data from a diverse collection of heterogeneously structured resources derived from multiple network providers, they utilize Semantic Web technology to develop and provide the largest network of Linked Data for Life Sciences. The task of Linked Data integration is to create software that can ingest such data, extract, and display meaningful biological information in a user-friendly manner.

Sensitive site design that makes the web page appear nice on all platforms is becoming more relevant with the pervasiveness of mobile devices (tablets and phones). Protein bioinformatics databases of the next decade can provide consumers with an optimized viewing and interaction interface through a wide variety of devices utilizing technologies such as Bootstrap (http://www.getbootstrap.com), JQuery (https://www.jquery.com), and Dojo Toolkit (https://dojotoolkit.org), etc. The creation of NoSQL technology and a high-performance index and search framework such as Lucene/Solr (http://lucene.apache.org) for rapid information retrieval has also been motivated by the need for pace, particularly for web-based applications.

Conflict of Interest None.

Additional Information Fig. 6.1 (CC0 1.0) [7] have been reused under Creative Commons Attribution licenses.

References

1. Mitra S, Acharya T. Data mining: multimedia, soft computing, and bioinformatics. 1st ed. Hoboken: Wiley-Interscience; 2003. 424 p.
2. Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011. 740 p.
3. Mittal S, Zaman M. A review of data mining literature. Int J Comput Sci Inform Sec. 2016;14 (11):437.
4. Ramez E, Shamkant N. Fundamentals of database system. London: 7th ed., Pearson Education; 2017. 1272 p.
5. Reeder MM. Reeder and Felson's Gamuts in radiology: comprehensive lists of roentgen differential diagnosis. New York: Springer Science & Business Media; 2013. 691 p.
6. Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AIMag. 1996 Mar 15;17(3):37.
7. Holzinger A, Dehmer M, Jurisica I. Knowledge discovery and interactive data Mining in Bioinformatics - state-of-the-art, future challenges and research directions. BMC Bioinformatics. 2014 May 16;15(6):I1.
8. Pérez-de-Castro AM, Vilanova S, Cañizares J, Pascual L, Blanca JM, Díez MJ, et al. Application of genomic tools in plant breeding. Curr Genomics. 2012 May;13(3):179–95.

9. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet. 2008 Mar;24(3):142–9.
10. Sasaki T, Burr B. International Rice genome sequencing project: the effort to completely sequence the rice genome. Curr Opin Plant Biol. 2000 Apr;3(2):138–41.
11. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 2013 Feb;54(2):e6.
12. Song S, Tian D, Zhang Z, Hu S, Yu J. Rice genomics: over the past two decades and into the future. Genomics Proteomics Bioinformatics. 2018 Dec 1;16(6):397–404.
13. Garg P, Jaiswal P. Databases and bioinformatics tools for rice research. Curr Plant Biol. 2016 Nov 1;7–8:39–52.
14. Bono H. All of gene expression (AOE): an integrated index for public gene expression databases. PLoS One. 2020 Jan 24;15(1):e0227076.
15. Kodama Y, Mashima J, Kosuge T, Ogasawara O. DDBJ update: the genomic expression archive (GEA) for functional genomics data. Nucleic Acids Res. 2019 Jan 8;47(D1):D69–73.
16. Bebek G. Identifying gene interaction networks. Methods Mol Biol. 2012;850:483–94.
17. Avery L, Wasserman S. Ordering gene function: the interpretation of epistasis in regulatory hierarchies. Trends Genet. 1992 Sep;8(9):312–6.
18. Dolma S, Lessnick SL, Hahn WC, Stockwell BR. Identification of genotype-selective antitumor agents using synthetic lethal chemical screening in engineered human tumor cells. Cancer Cell. 2003 Mar;3(3):285–96.
19. Overbeek R, Fonstein M, D'Souza M, Pusch GD, Maltsev N. The use of gene clusters to infer functional coupling. Proc Natl Acad Sci U S A. 1999 Mar 16;96(6):2896–901.
20. Goh C-S, Cohen FE. Co-evolutionary analysis reveals insights into protein-protein interactions. J Mol Biol. 2002 Nov 15;324(1):177–92.
21. Bebek G, Yang J. PathFinder: mining signal transduction pathway segments from protein-protein interaction networks. BMC Bioinformatics. 2007 Sep 13;8:335.
22. Breitkreutz B-J, Stark C, Tyers M. Osprey: a network visualization system. Genome Biol. 2003;4(3):R22.
23. Mrvar A, Batagelj V. Analysis and visualization of large networks with program package Pajek. Complex Adap Syst Model. 2016 Apr 6;4(1):6.
24. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000 May;25(1):25–9.
25. Maere S, Heymans K, Kuiper M. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics. 2005 Aug 15;21(16):3448–9.
26. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. Nucleic Acids Res. 2019 Jan 8;47(D1):D330–8.
27. Fabregat A, Sidiropoulos K, Garapati P, Gillespie M, Hausmann K, Haw R, et al. The Reactome pathway knowledgebase. Nucleic Acids Res. 2016 Jan 4;44(D1):D481–7.
28. Haw R, Stein L. Using the Reactome database. Curr Protoc Bioinformatics. 2012;38 (1):8.7.1–8.7.23.
29. Roumpeka DD, Wallace RJ, Escalettes F, Fotheringham I, Watson M. A Review of Bioinformatics Tools for Bio-Prospecting from Metagenomic Sequence Data. Front Genet [Internet]. 2017;8. [cited 2020 Dec 26]; Available from: https://www.frontiersin.org/articles/10.3389/fgene.2017.00023/full#B23.
30. Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, et al. KEGG for linking genomes to life and the environment. Nucl Acids Res. 2008 Jan 1;36(suppl_1):D480–4.
31. Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo CT, et al. WikiPathways: building research communities on biological pathways. Nucleic Acids Res. 2012 Jan 1;40(D1): D1301–7.
32. Pico AR, Kelder T, van Iersel MP, Hanspers K, Conklin BR, Evelo C. WikiPathways: pathway editing for the people. PLoS Biol. 2008 Jul 22;6(7):e184.

33. Caspi R, Billington R, Ferrer L, Foerster H, Fulcher CA, Keseler IM, et al. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res. 2016 Jan 4;44(D1):D471–80.
34. Wallace RJ, Rooke JA, McKain N, Duthie C-A, Hyslop JJ, Ross DW, et al. The rumen microbial metagenome associated with high methane production in cattle. BMC Genomics. 2015 Oct 23;16(1):839.
35. Chen C, Huang H, Wu CH. Protein bioinformatics databases and resources. Methods Mol Biol. 2017;1558:3–39.
36. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. 2008 Oct;41(5):706–16.
37. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF platform: linked open data for the life sciences. Bioinformatics. 2014 May 1;30(9):1338–9.

Sequence Alignment

7

Manoj Kumar Gupta, Gayatri Gouda, N. Rajesh, Ravindra Donde, S. Sabarinathan, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

The sequence analysis is one of the most effective and commonly applied methods (explicity or implicitly) in biological research. Thus, in this chapter, author attempted to understand the basics of sequence analysis and how researchers implement various computational tools to achieve them. Information obtained revealed that alignment can be either global and local or pairwise sequence alignment and multiple sequence alignment. For performing these alignment, various algorithms like dynamic programming, heuristic algorithms, or probabilistic methods have been developed. Sequence analysis helps us to detect evolutionary relationship as well as scan motifs by taking into consideration of various events, such as mutations, insertions, deletions, and reordering under some circumstances. Thus, sequence alignment serves as an essential requirement for the most of the biological research ranging from genomics to proteomics. However, our perception of alignment biases remains primitive.

M. K. Gupta (✉) · G. Gouda · R. Donde · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_7

Thus, there is an urgent requirement to explore the effect of alignment bias on broad comparative genomics accuracy. In the near future, information present in this chapter will be useful for retriving information biological sequence.

Keywords

Dynamic programming · BLAST · FASTA · Needleman–Wunsch algorithm · gap penalties · Smith-Waterman algorithm

Abbreviations

BLAST	Basic local alignment search tool
DP	Dynamic programming
HMM	Hidden Markov Model
MAFFT	Multiple sequence alignment based on Fast Fourier Transform
MSA	Multisequence alignments
MUSCLE	Multiple sequence comparison by log-expectation
PA	Progressive alignment
PSA	Pairwise sequences alignment
UPGMA	Unweighted pair group method with arithmetic mean

7.1 Introduction

More than 12 million organisms reside on the earth. This biodiversity is mainly due to distinct genomic and proteomic sequences contained in these organisms. These sequences store unique information that modulates various processes required for the survival of these organisms [1]. DNA sequence comparison is a unique approach to evaluate gene-level variations amongst these organisms and to study their differences and similarities [1]. What "similarities" are identified to rely on the alignment process's objectives. The easiest way for comparing two same-length sequences is to identify the number of matching characters. The attribute that calculates sequence similarity is known as the alignment value of two sequences. On the contrary, the degree of dissimilarity between sequences is known as the sequence distance. The amount of characters that do not align is known as the hamming distance. However, while estimating similarity, this approach does not take into consideration of normal biological activities like insertion or deletions.

The classic definition of sequence alignment includes estimating the so-called "edit distance," which normally equals the minimum number of insertions, substitution, and deletion that are necessary for transforming one sequence into another [2]. Earlier several algorithms, like Smith & Waterman and Needleman & Wunsch have been developed for computing "edit distance" [3, 4]. These algorithms were originally developed for protein-protein alignment and subsequently employed for

DNA sequence alignment. In the majority of the real-life scenarios, nevertheless, these algorithms seem inefficient for DNA alignment owing to their runtime as well as memory requirements [2].

To date, several kinds of alignment approaches, like prediction-based methods, pairwise sequences alignment (PSA), profile-based methods, multisequence alignments (MSA), and the structure-based methods have been proposed [5]. The most frequently used are PSA and MSA. In PSA, per sequence is aligned once a time. It is the easiest method of aligning and can be achieved with two strategies: local and global. The MSA approach could also be implemented using local or global strategies but is much more complex. During MSA, many protein sequences are organized into a rectangular array, and residues that are either homologous or identical are placed in one column. MSA is generally employed for detecting conserved regions in protein sequence and for designing protein's secondary and tertiary structures. Homology, as well as evolutionary relationships between sequences, may also be derived via MSA approaches because MSA has an underlying postulation, i.e., all matching sequences would share evolutionary homology [5]. Alignment results are also a requirement for many other downstream analyses, like drug design. Nevertheless, results generated by different methods can be quite diverse [6]. Thus, there is an urgent requirement for the development of systematic metrics that may provide explicit guidance on the strengths as well as shortcomings of the different sequence alignment algorithms. This, in turn, will help us to deduce a more significant relationship between sequences. Considering the above, in this chapter, the author attempted to provide an overview of sequence alignment with a summary of popular specific algorithms, methods, and approaches which underlie the most current method of sequence alignment.

7.2 Basic Terminology

A sequence alignment is a basic analysis in almost every biological study (implicit or explicit). The main objective of sequence alignment is to detect the homologous sites in sequences [7]. Homology is a qualitative argument and identifies shared ancestral relations between sequences. Two distinct types of homology exist, i.e., paralogs (shared ancestry due to a duplication event) and ortholog (shared ancestry due to a speciation event) [8]. *"By definition, orthologs are genes that are related by vertical descent from a common ancestor and encode proteins with the same function in different species. By contrast, paralogs are homologous genes that have evolved by duplication and code for a protein with similar, but not identical functions"* [9]. Other terms that are commonly used during sequences analysis are similarity and identity [10]. Unlike homology, similarity denotes the percentage of aligned residues with the same physicochemical properties that are easier to replace each other. It is pertinent to note that two sequences can be 70% similar but cannot share 70% homology. They are either nonhomologous or homologous [10]. In general, a shared ancestral relationship could be inferred if the sequence similarity level is very high. However, it is not really obvious at what similarity degree one should assume homologous relationships. The solution depends on the sequence type and lengths under consideration [10]. For instance, proteins having high sequence identity and high structural similarity have similar functional and evolutionary relationships [11]. Identity corresponds to the proportion of matches between the two aligned sequences with the same amino acid residue [10].

Another term, namely gap, is common during sequence analysis. A gap can be defined as the absence of a segment in a certain sequence. Gaps are natural feature of biological sequences. A single mutational event can result in the addition or deletion of certain regions of sequences (predominantly in DNA), and thus the effective identification of gaps is an important step toward understanding the various biological phenomenon [12]. A variety of biological processes may lead to the formation of gaps in DNA sequences, like, large pieces of DNA may be replicated and inserted through a single mutational occurrence, and slippage during the replication of the DNA can allow the same region to be replicated many times as replication machine lose its position on the template [12]. Earlier it has been reported that instead of penalizing all editing operations individually, one must penalize the formation of a longer gap more severely than others [13].

7.3 Alignment Methods

To date, different alignment approaches like dynamic programming (DP), heuristic algorithms, or probabilistic methods have been developed [14].

7.3.1 Dynamic Programming

DP is an effective computing strategy implemented to a problem class that can be addressed recursively [15]. When Richard Bellman first developed the DP algorithm in 1953 for researching “multi-stage decision problems,” he certainly did not expect its extensive usage within modern computer programming. Indeed, as Bellman has described in his comical autobiography [16], he wanted to employ the word “dynamic programming” as “an umbrella” for the mathematical research he carried out at RAND Corporation for protecting his boss, who was the Secretary of Defense Wilson and “had a pathological fear of word research.” Since it is one of the first algorithms that were used in bioinformatics research and has since been widely applied [17], DP has become an inevitable algorithmic subject.

DP is indeed a normal preference for evaluating sequences. Needleman & Wunsch initially illustrated the use of bottom-up DP for calculating an optimal pairing amongst two protein sequences [3]. While this algorithm offers a comparative evaluation of sequences pair, it estimates the similarity throughout the complete sequences (a “global alignment algorithm”). Hence, this approach is time-consuming and computationally exhaustive [18]. To overcome this, Smith and Waterman adapted DP for performing local alignments in which alignment was made between similar parts of the input sequences [4]. DP provides an ideal approach for PSA [18]. It is also widely employed to assembling DNA sequence data from fragments obtained from automated sequencing machines and for determining the exon/intron structure within eukaryotic genes [19]. It is also utilized for inferring proteins' function through homology study with other proteins having a known function [3, 4], and for predicting the secondary structure of functional RNA genes or regulatory elements [19].

7.3.2 Heuristic Algorithms

Though DP gives a more accurate result, it is slow [14]. Other efficient approaches, like heuristic algorithms or probabilistic methods, have been developed for large-scale database searching. The term "heuristic" means that the developed algorithm is faster than the classical method but may not be the optimum method [20]. Heuristic algorithms can be categorized into three subgroups, namely, progressive alignment (PA) approach, iterative alignment type, and block-based alignment type [10]. PA approach is the incremental strategy that generates a final MSA through conducting a set of PSA on successively less closely associated sequences. In this approach, we align the two closest-related sequences first and then align the closest-related sequence in the questionnaire to the alignment generated in the previous step. Although success is particularly dependent on the consistency of the initial alignment and dramatically deteriorates when all sequences in the set are related distantly, PA methods are enough to be implemented on a broad scale for several sequences [21]. The most commonly used PA methods are ClustalW (https://www.genome.jp/tools-bin/clustalw) and T-Coffee (https://www.ebi.ac.uk/Tools/msa/tcoffee/). However, it is not possible that the progressive approaches converge to optimal global alignment, and efficiency can be difficult to approximate. Additionally, its true biological importance may be unclear [21].

The iterative method is based on the premise that an ideal solution could be sought by adjusting current suboptimal solutions on a repeated basis. The process begins with a low-quality alignment and gradually improves it through well-defined procedures until no more improvement can be achieved on the alignment scores. Since the sequence order in each iteration is different, this method could mitigate the "greedy" problem of progressive strategy. Nevertheless, this approach is also heuristic in nature and has no promises for optimum alignment [10]. PRRN (https://www.genome.jp/tools-bin/prrn) is a web-based program that utilizes a double-nested iterative strategic plan for multiple alignments. The progressive as well as iterative alignment techniques are primarily global and thus cannot detect conserve motifs and domains amongst strongly diverging sequences of various lengths. A local alignment strategy must be employed for those divergent sequences that share only local similarities. This technique detects the ungapped alignment block that is present in all sequences, and hence this is called the block-based local alignment technique [10]. DIALIGN2 (http://dialign.gobics.de/) web-tools that employ block-based alignment for detecting local alignment.

7.3.3 Probabilistic Methods

Introduction of probabilistic modeling approaches, like profile secret Markov models (profile HMMs) as well as pair-HMMs [22] have advanced sequence similarity search. When variables are probabilities instead of random scores, objective statistical parameters refine them more readily. This helps to create more detailed, biologically relevant models with many parameters. For instance, profile HMMs employ position-specific deletion/insertion probabilities instead of the random, position-invariant gap expense of more conventional approaches like BLAST or PSI-BLAST [23], enabling profile HMMs to model the possibility that indels occur more frequently in certain sections of a protein than others (e.g., in surface loops than submerged core) [24].

The probability method has three primary benefits: (i) Any kind of analogy may be adjusted to the probabilities [e.g., The DNA error-prone reads against the genome]. The comparisons are supposed to be more precise. (ii) We may approximate the reliability, for instance, each column of every alignment part. This is helpful because alignments also have unknown sections owing to high inconsistencies or repeating sequences. (iii) A similarity between two integrated sequences over potential alignments may be calculated. This can more powerfully detect subtle connexions than single ideal alignments [25]. The probabilistic approach, however, also has significant disadvantages. Aside from a moderate computational drawback, the probabilistic method suffers from uncharacterized score statistics - unlike the local alignment of Smith-Waterman, for which at least the form of the ideal score distribution is defined from the null model, relatively little is known about the distribution of the log-like score in the local probabilistic random alignment. It is proven empirically that random usage of the z-score would not deliver really strong results [26].

7.4 Global and Local Alignment

Sequence alignment approaches typically fell into two categories: global and local alignments. While global alignment compares all character of query sequences, local alignments define similarity regions within long sequences that are typically divergent. The Needleman-Wunsch algorithm is a well-known global alignment algorithm designed on the basis of DP. Local alignments are always preferred but more challenging to quantify considering the additional difficulty of recognizing similarities regions. The Smith-Waterman algorithm is a general local alignments method based on the DP system, with added features for beginning and finishing in either place [14]. Most biologists think that local alignment is what really matters when we are looking for functional conservation. Local alignment is more important since certain proteins have roles that are controlled by their capability to attach to some other molecule (protein's ligand); therefore, the role would be maintained if this short portion becomes sustained via evolution, even if there is significant divergence in many other protein regions. As proteins are folded within their natural form, these retained regions need not be continuous protein segments. Indeed, several researchers researching on lymphocyte antigen recognition specifically account for these discontinuities within binding domains (known as “non-linear” epitopes, where an epitope is the ligand of a lymphocyte) [12, 14].

In few cases of the global alignment mode, adding a distance in the leftmost location of the alignment might be needed, but we are not aware of the length of the next reference sequence factor to be already aligned. It is obvious from this scenario that an intermediate alignment is required between the local and global alignment (i.e., semiglobal alignment) [12]. A semiglobal alignment does not penalize starting or ending gaps in any global alignment so that the resultant alignment continues to overlap one end of a sequence with the end of the other [27]. A Parasail is a stand-alone tool that can be employed for performing global, local, and semi-global alignment [27]. Recently, Suzuki & Kasahara developed a semi-global alignment algorithm, namely, “difference recurrence relationships,” that perform better than other available tools by 2.1 factor [28].

7.5 Pairwise Alignments

The most frequently employed mean of collecting information from protein and DNA sequences is a PSA. It is generally used to detect protein homolog, which diverged more than 2 billion years ago. For proteins that share statistically significant sequence similitudes, homology can be accurately inferred. If statistically meaningful similarities to a known sequence are observed, inferences may be made regarding the unknown sequence’s function, structure, and biologically significant residues. Although the homology assumption [29] is very robust (i.e., proteins which share significant similarities within PSA often have similar features), a few of the more detailed preassumptions critically rely on the consistency of the alignment between the two sequences. For instance, functional inferences for protein sequences having more than 60% identity are typically very reliable. However, uncertainty in the alignment of badly conserved areas can lead to errors for more distantly linked proteins [30, 31].

The fundamental law for sequence alignment is the structural alignment amongst two proteins known to have a 3D structure. The 3D-structure comprises more information relative to the 1-D sequence as well as diverges at a very slow rate. Thus, distant evolutionary correlations may also be established amongst sequences which do not display statistically significant similarities. Even directly relevant proteins with major sequence similarities may elicit sequence alignments that differ from the most accurate structural alignments. Since it is not possible to identify the three-dimensional structure of each protein, researchers are continually seeking for strategies for producing structurally correct homology models for sequences with unknown structure. The most common as well as successful methods, are to find a template for constructing the model within the set of established structures. This feature is relatively trivial in the case of high sequence similitude (i.e., $> 60\%$ identity) because both sequences, as well as structural alignments, are typically very near to this range. However, in this zone, there are just a few sequences; in the so-called “twilight zone,” there are several more sequences (i.e., ~20–40 percent sequence identity) where divergent yet clearly homologous protein may be hard to match. Although the precision of the end 3D model is dependent on the degree of alignment of the unspecified sequence to the structural template, researchers are mainly concentrating on enhancing the quality of alignment between proteins that share statistically relevant similarities and have 20% to 40% sequence identity [49, 50]. Dot-matrix techniques, DP, and Word techniques are the most widely used methods for PSA.

7.5.1 DOT Matrix Plot

Since visualization of alignment of character of hundreds or more sequences can be troublesome, scientists created a more visually understandable approach called the dot matrix approach. This sequence alignment process, which was first carried out manually and then computationally, allows the more apparent mapping of similarities for visual inspection. In this process, a sequence is shown on the top and one on the side of the matrix and a mark on the crossroads of the corresponding character pairs [51]. A dot matrix pattern will have a continuous array of dots running along the middle diagonal of the matrix for a pair of exactly matched sequences (Fig. 7.1). However, this trend is hardly used. Sometimes, without further processing, diagonal patterns are hard to recognize. Thus, a number of filters are also added to the results, as well as the use of color and other methods to highlight matching sequences. For instance, typical filtering is a stringency/window combination. The window represents the number of points evaluated at a time, while the minimum number of matches needed in each window is the stringency [51].

The study of the dot matrix is extremely valuable in recognizing recurring characters or short sequences within one sequence, as is the case for the mapping the recurrent regions of entire chromosomes. Repeats of the same character produce artificially high scores and complicate sequence alignment. Methods of dot matrix are most appropriate for single PSA problems, particularly for relatively high similitudes. Sequences with a lower similarity and MSA need more efficient methods [51]. Even though window stringency values are always heuristically determined, they could be dependent on dynamic averages, matched scores in aligned protein groups, or different methods for calculating the amino acid similarity. For example, score matrices establish alignment scores in the aligned protein families depending on their statistical frequency. These matrices may be used to construct a sliding window, where only scores above an average scoring may appear in the matrix, as defined in the following section [51].

To date, various algorithms and computer software tools were created for performing the dot-matrix plot. While several of these tools accommodate 100 kb of sequences, the study of the genome sequences above 10 Mb on a microcomputer remains to be inoperative considering the length of time needed for execution as well as computer memory [53]. In 2004, Huang and Zhang created two dot matrix

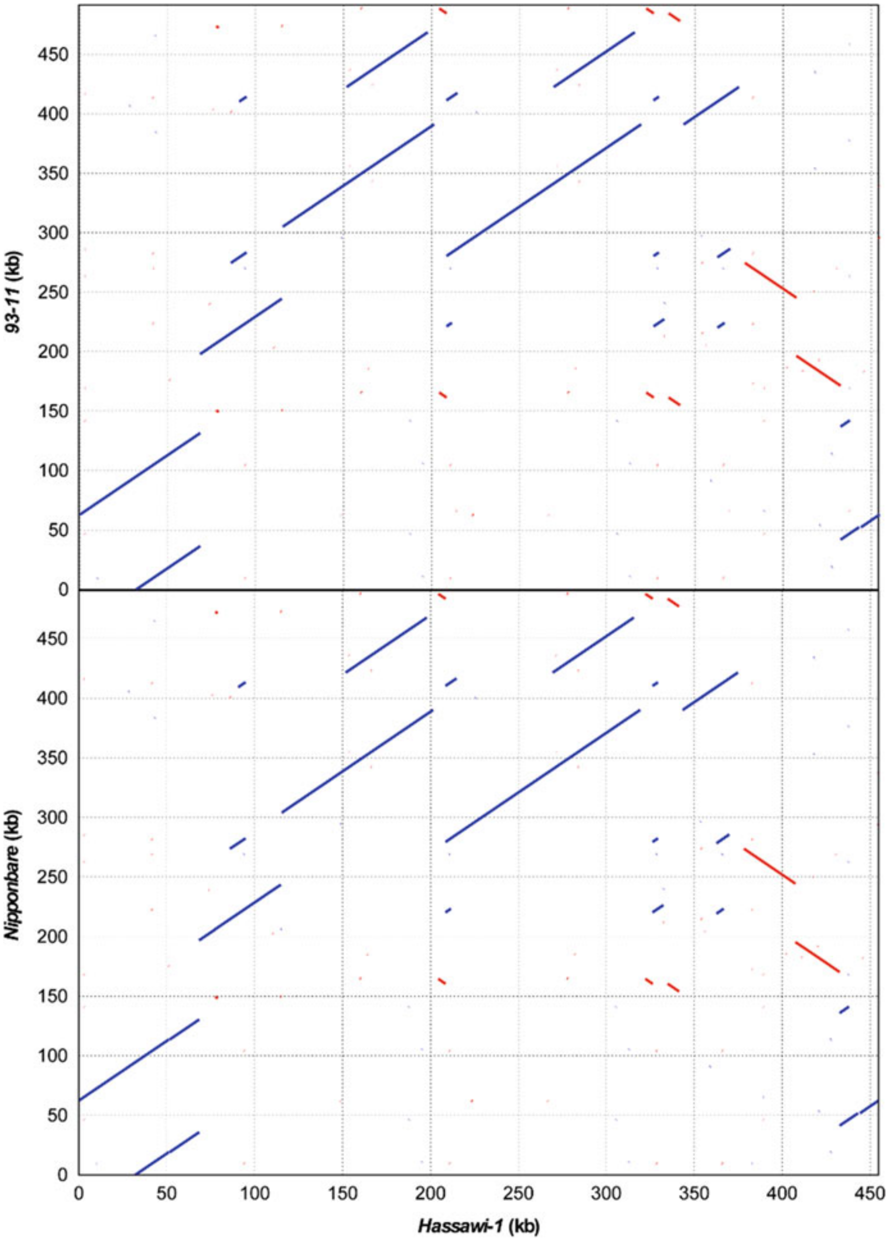


Fig. 7.1 The dot-plot of the alignment for human chromosomes 2, 7, and 14 and mouse chromosome 12. The x-axis indicates the positions of mouse chromosome 12, and y-axis indicates the positions of human chromosomes 2, 7, and 14. The orthologous landmarks are plotted based on the pairwise alignments between the three human chromosomes and mouse chromosome 12 (Adapted from [52]).

Table 7.1 Softwares and tools used for PSA (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Alignment type[a]	Sequence type[b]	References
ACANA	Fast heuristic anchor-dependent PSA	Both	Both	[32]
AlignMe	Membrane PT sequences alignment	Both	PT	[33]
Bioconductor biostrings:: pairwiseAlignment	DP	Both + ends-free	Both	[34]
BioPerl dpAlign	DP	Both + ends-free	Both	https://metacpan.org/pod/release/CJFIELDS/BioPerl-1.6.924/Bio/Tools/dpAlign.pm
BLASTZ, LASTZ	Seeded pattern-matching	LL	Nucleotide	[35]
DNASTAR Lasergene molecular biology suite	Align RNA, DNA, PT, or PT + DNA sequences	Both	Both	https://www.dnastar.com/
FEAST	Posterior-dependent LL extension having descriptive evolution model	LL	Nucleotide	[36]
G-PAS	GPU-based DP with backtracking	LL, SemiGL, GL	Both	http://gpualign.cs.put.poznan.pl/gpas20.html
GapMis	Does PSA with one gap	SemiGL	Both	[37]
Genome magician	Software for ultra-fast LL DNA sequence motif scan as well as PSA of high-throughput data in both FASTA and FASTQ format.	LL, SemiGL, GL	DNA	https://science.do-mix.de/software_genomemagician.php
GGSEARCH, GLSEARCH	GL:LL (GL) and GL: GL (GG) alignment with statistics	GL in query	PT	[38]
JAligner	Java-based techniques of Smith-Waterman	LL	Both	http://jaligner.sourceforge.net/

(continued)

Table 7.1 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	References
K*sync	PT sequence to structure alignment that comprises of secondary structure, structure-derived sequence profiles, structural conservation, and consensus alignment scores	Both	PT	[39]
LALIGN	Multiple, nonoverlapping, LL similarity	LL nonoverlapping	Both	https://www.ebi.ac.uk/Tools/psa/lalign/
mAlign	Modeling alignment; models the information content of the sequences	Both	Nucleotide	[40]
Matcher	Waterman-Eggert LL alignment (dependent on LALIGN)	LL	Both	https://www.ebi.ac.uk/Tools/psa/emboss_matcher/
MCALIGN2	Explicit models of indel evolution	GL	DNA	[41]
MUMmer	Suffix tree-dependent	GL	Nucleotide	[42]
NW-align	Standard Needleman-Wunsch DP algorithm	GL	PT	https://zhanglab.ccmb.med.umich.edu/NW-align/
Needle	Needleman-Wunsch DP	SemiGL	Both	https://www.ebi.ac.uk/Tools/psa/emboss_needle/
Ngila	Logarithmic as well as affine gap costs and explicit models of indel evolution	GL	Both	[43]
Parasail	C/C++/python/Java SIMD DP library for SSE, AVX2	GL, ends-free, LL	Both	[27]
Path	Smith-Waterman on PT back-translation graph (detects	LL	PT	[44]

(continued)

Table 7.1 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	References
	frameshifts at PT level)			
PatternHunter	Seeded pattern-matching	LL	Nucleotide	[45]
SABERTOOTH	Alignment employing predicted "connectivity profiles"	GL	PT	[46]
Satsuma	Parallel whole-genome synteny alignments	LL	DNA	Genome-wide synteny through highly sensitive sequence alignment
SPA: Super pairwise alignment	Fast pairwise GL alignment	GL	Nucleotide	[47]
SWIFOLD	Smith-Waterman acceleration on Intel's FPGA with OpenCL for long DNA sequences	LL	Nucleotide	[48]
UGENE	Opensource Smith-Waterman for SSE/CUDA, suffix array-based repeats finder and dotplot	Both	Both	http://ugene.net/

[a] Alignment type: Global(GL)/Local(LL)
[b] Sequence type: Nucleotide (NT)/Protein(PT)

comparison methods for studying large sequences. Initially, the methods identify similarity regions amongst two sequences using a rapid word search algorithm and explicitly compare these regions. Because several random matches are omitted from the initial sampling, the estimation duration is decreased dramatically. These approaches yield good quality plots of the dot matrix with low background noise. Spatial criteria are linear, so genome scaling sequences can be compared by algorithms. Highly repetitive sequence structures of eukaryote genomes may impact the computational speed. In the 80s, with a 1GHz personalized machine, a dot matrix complot was developed for the yeast genome (12 Mb) for both strands [53].

7.5.2 Dynamic Programming

The most widely employed algorithm of PSA is DP, initially introduced by Needleman and Wunsch [3]. The DP ensures an optimum algorithmic alignment with unique parameters and sequences. However, an optimum sequence alignment score would not assure the structural consistency of the alignment. Additionally, there are no natural mechanisms under which two proteins align together. Therefore "optimum" alignments of the sequence may vary greatly from ideal structural alignments [31]. Moreover, distant-related proteins also have several optimal alignments and a significant number of sub-optimal alignments with scores quite similar to the optimal score [50, 54, 55]. If one moves further from the desired score, the number of alternatives alignment also keeps increasing. Therefore, one must sample the suboptimal alignment space for holding the number of alignments computationally trackable [50, 54, 55].

While a structure-based alignment is the "gold standard" against which sequence alignments are measured, structural alignment may vary, and no optimum structural alignment algorithm is possible [56]. As the number of structures appear to be smaller than the number of sequences, the structural alignment variations are minimal relative to the sequence-structural alignment variations. Although this definitely refers to quite distantly linked proteins that have no meaningful similitude (and therefore cannot be substantially aligned with sequence data alone), the structural and sequence alignment precision of proteins that share statistically significant similarities has not been closely studied [56]. Given that structurally correct alignments frequently include suboptimal alignment scores, researchers have been researching the alternate alignments and wondering whether they include details about precise structural alignments. Jaroszewski et al. [50] have studied alternate alignments, both based on an almost ideal algorithm for alignment generation and by combining score parameters (i.e., substitution matrix and gap penalties), and have found that alignment in the sets is much similar to the structural alignment. Their inference was that the two alternate alignment methods, namely, alternatives and sub-optimizing alignments, had complementary information (in contrast to redundant information) because the combination of the two sets created much higher alignments than any of the sets. The exactness of the optimal sequence alignment was also investigated by Holmes and Durbin [57]. They developed a technique for calculating the expected accuracy. In an algebraic approach, Zhang and Marr [58] used alternate alignments with maximal alignments in the neighborhood.

Various scholars also took the help of a probabilistic approach for producing alternate alignment sets. In 1995, Miyazawa [59] measured alignment likelihoods relying on alignment score exponent and, subsequently, compared the resulting likelihoods of matched amino acids throughout alignment with the respective protein structure alignments. Yu and Hwa investigated the statistically significant of alignments made using a pairwise Hidden Markov Model (HMM) [26]. Knudsen and Miyamoto [60] designed a pairwise HMM alignment approach that provided an explicit indel evolutionary model. Eventually, Mückstein and the team [61] constructed a sampling alignment procedure on the basis of statistical weighting employing partition function overall plausible two-sequence alignments.

Although it is of theoretical interest to compare individual sequence and structure sets in the absence of any structural information, it is only of practical use if the alignment of the sequence can be determined correctly. One approach to resolving this issue is to calculate the accuracy of a certain aligned residual pair (that we term an edge, using the norm for determining the optimum score in the dynamic programming path graph, aligned residues, insertions, and deletions along the edge) [31]. Cline and the team examined four strategies for forecasting the accuracy of a particular pair of aligned residues [62] and concluded that the most improved alignment quality was the method proposed by Yu and Smith [63] for retrieving near-optimal alignments from the HMM profile. The association between both the edge probabilities and structural alignment was studied by Knudsen & Miyamoto [60] and Mückstein et al. [61] and Miyazawa [59]. However, in the former two cases, only in the context of a limited number of protein pairs, usually considered a strong correspondence amongst them. In another study, Mevissen and Vingron [64] have evaluated the feasibility of an edge reliability index known as robustness that Chao and the team had previously defined [65]. They found that an edge's robustness predicted correctly if the edge was still aligned in structural alignment. In another study, Sierka and the team improvised the robustness analysis by adding extra details on alignment consistency and creating a logistic regression model that returns the likelihood that a given edge is embedded in a structural alignment [31].

7.5.3 The Word or K-Tuple (Ktup) Method

It is the heuristic process, which offers greater alignment than DP. Currently, with massive datasets, DP cannot be used. This is why we use the K-tuple approach when searching for a specific question along with a large database. K Tuple corresponds to a series of k words. For instance, for nucleotide and protein, K is defined as 11 and 3, respectively. The K system has been introduced in the family of FASTA and BLAST.

7.5.3.1 FASTA

FASTA is a rapid alignment application for protein and DNA sequence pairs. Rather than comparing individual residues in both sequences, FASTA looks for matching sequence patterns or terms called k-tuples. In both sequences, these patterns contain k consecutive matches of letters. Based on these word matches, the algorithm then tries to establish a local alignment. FASTA is useful for regular database searches of this kind because of the ability of the algorithm to locate similar sequences in a sequence database with high-speed. FASTA programs offer a detailed range of simple similarity search resources (fasta36, fastx36, tfastx36, fasty36, and tfasty36), comparable to those offered by the BLAST tool, as well as programs for local, slower, optimal, as well as global similarity searches (search36, ggsearch36) and oligonucleotide and short peptide searches (fasts36, fastm36). fasta36 employs the FASTA algorithm developed by Pearson alone and Pearson & Lipman and compare protein (or nucleotide) sequence to protein (or nucleotide) sequence database [66, 67]. With the ktup (word size) parameter, search speed and selectivity are regulated. By default, ktup = 2 for protein comparisons; ktup = 1 is more sensitive but slower. By default, ktup = 6 for DNA comparisons; ktup = 3 or ktup = 4 allows maximum sensitivity. fastx36/fasty36 compares the translated nucleotide sequence into three frames and allowing gaps and changes, fastx36 compares a nucleotide sequence to a protein sequence base. Fastx36 uses a faster and simplified alignment algorithm, which only allows the frameshift between codons. However, fasty36 is slower, but better alignments are possible because frame shifts inside codons are permitted [68]. tfastx36/ tfasty36 compares a protein sequence with a nucleotide sequence database and measures comparisons for forward and reverse directed frames-shifts [68]. ssearch36 employs the Smith-Waterman algorithm [4] for comparing a nucleotide (or protein) sequence against a nucleotide (or protein) sequence database. The Fasta36 is just 2–5 times faster than Farrar SSE2 [69]. ggsearch36/ glsearch36 compares a protein (or nucleotide) sequence to a protein (or nucleotide) sequence database, employing an optimal global algorithm: global: local (glsearch36) or global (ggsearch36). fasts36/ tfasts36 compares collection of small peptide fragments as collected from mass-spec, protein research, against nucleotide (tfasts) or protein (fasts) databases [70]. fastm36 compares ordered short nucleotide sequences (or peptides) to a nucleotide (or peptides) database.

The FASTA systems employ an empiric approach for approximating statistical importance that is consistent with a variety of similarities in scores and gap penalties and increases alignment of boundary precision as well as search sensitivity. FASTA systems can generate "BLAST-like" alignment as well as tabular results for ease of integrating analytics pipelines and can scan for small, descriptive datasets and afterward report findings for larger sequences employing small dataset connexions. FASTA systems operate in a wide range of database formats, like PostgreSQL and MySQL databases. Recently, Pearson has developed programs that lay out a strategy for incorporating domain as well as active site annotations into alignments and emphasizing the mutation status of functionally important residues. These protocols also explain how FASTA systems can classify protein and nucleotide sequences through protein: DNA, protein: protein, and DNA: DNA comparative study [71].

7.5.3.2 BLAST

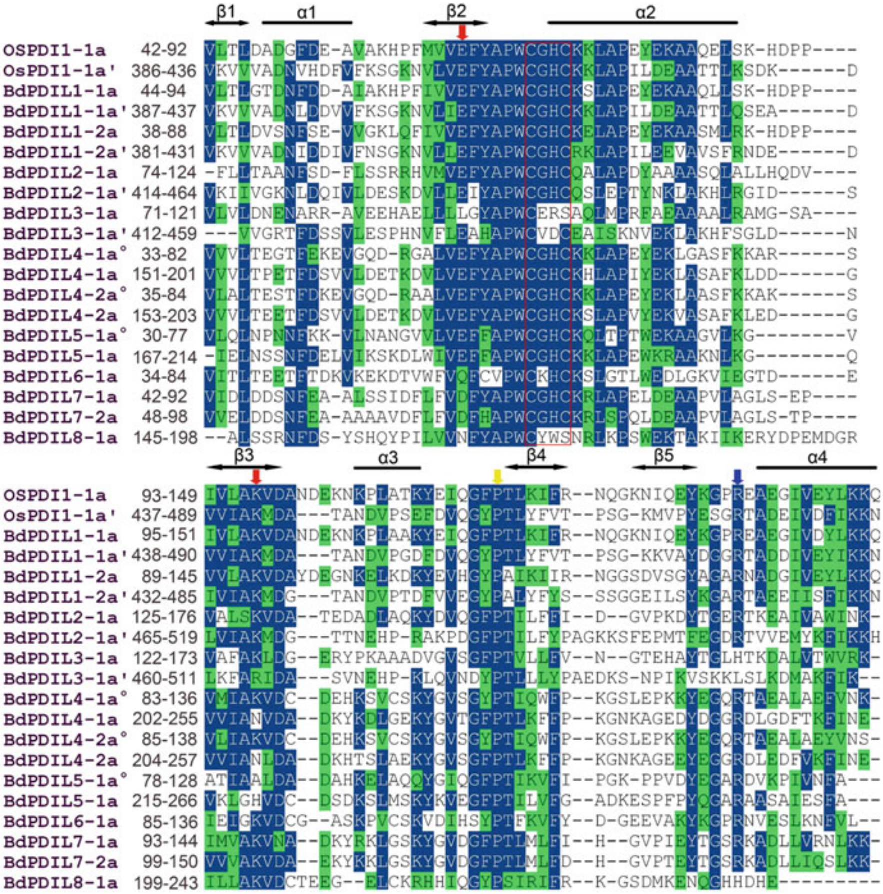
The "Basic local alignment search tool" (BLAST) is a sequence similarity search software which could be employed either as a stand-alone tool or through a web interface for comparing all combinations of protein (or nucleotide) sequence to a protein (or nucleotide) sequence database [72]. BLAST is a heuristic approach that finds short matches between two sequences and tries to initiate alignment from these "hot spots." BLAST also offers statistical details about alignment in addition to executing alignments [72]. The E-value contains details on the probability of a sequence being matched by sheer chance. The smaller the E-value, the less probable the database match is to be attributed to random chance, and thus the more important the match. If $E < 1e^{-\ 50}$ (or 1×10^{-50}), there should be an exceptionally strong conviction that matching the database is the product of a homologous partnership. If E is between 0.01 and $1e - 50$, matching can be viewed as a consequence of homology. If E is between 0.01 and 10, the match is assumed to be nonsignificant but could suggest a possible remote homology relationship. Additional proof is required to validate the partnership. If $E > 10$, the sequences within evaluation are either unrelated or associated with incredibly remote relationships that fall far below the detection limit of the current system [10]. Although the E-value is proportionally influenced by the size of the database, an apparent concern is that as the database expands, the E-value often increases for a given sequence match. Since the true evolutionary relationship between the two sequences remains unchanged, as the database expands, the decline in the sequence match's credibility means that one will "lose" homologs previously observed as the database enlarges. Consequently, an alternative to E-value calculations is needed [10].

BLAST is a family of services that comprises BLASTN, BLASTX, BLASTTP, TBLASTX, and TBLASTN. BLASTN searches nucleotide sequences in the nucleotide sequence database. BLASTP employs protein sequences as requests to scan a database of protein sequences. BLASTX employs nucleotide sequences as inputs and converts them into all six reading frames to generate translated protein sequences that are used to query the protein sequence database. TBLASTN requests protein sequences to a nucleotide sequence database, with sequences encoded into all six reading frames. TBLASTX employs nucleotide sequences that are interpreted into all six frames to scan a nucleotide sequence database that has all the sequences interpreted into six frames. In addition, also there is a bl2seq program that executes a local alignment of two user-provided input sequences. The graphic production involves horizontal bars as well as a diagonal in a two-dimensional diagram displaying the total degree of the matching between the two sequences [10].

7.6 Multiple Sequence Alignment

MSA is an alignment between more than two biological sequences. In most scenarios, the input sequences are believed to have a shared ancestor. Sequence homology can be derived from the subsequent MSA, and a phylogenetic study can be carried out to determine the common ancestral roots of the sequences. Visual alignment representations, as seen in the Fig. 7.2, demonstrate mutation occurrences like point mutations (single nucleotide or amino acid changes) that occur as distinct symbols within a single alignment column and insertion/deletion of mutations (indels or gaps) that occur as hyphens in one or more alignment sequences. MSA can also be used to determine sequence conservation of protein domains, tertiary as well as secondary structures, as well as specific amino acids or nucleotides [73–75].

Since MSA of three or more lengthy sequences may be complicated and are often time-consuming to be aligned by hand, statistical algorithms are often used for generating and evaluating alignments. MSAs need more advanced approaches than PSA since they are more computationally complicated. Many MSA programs use heuristic approaches rather than global optimization since it is prohibitively costly to determine the optimum alignment amongst more than a few sequences of moderate length. On the other side, heuristic approaches usually refuse to guarantee the consistency of the answer, with heuristic strategies sometimes found to be well below the ideal solution in the case of benchmarks [73–75].



```
OSPDI1-1a     42-92   VLTLDADGFDE-AVAKHPFMVVEFYAPWCGHCKKLAPEYEKAAQELSK-HDPP----
OsPDI1-1a'   386-436  VKVVVADNVHDFVFKSGKNVLVEFYAPWCGHCKKLAPILDEAATTLKSDK------D
BdPDIL1-1a    44-94   VLTLGTDNFDD-AIAKHPFIVVEFYAPWCGHCKSLAPEYEKAAQLLSK-HDPP----
BdPDIL1-1a'  387-437  VKVVVADNLDDVVFKSGKNVLIEFYAPWCGHCKKLAPILDEAATTLQSEA------D
BdPDIL1-2a    38-88   VLTLDVSNFSE-VVGKLQFIVVEFYAPWCGHCKELAPEYEKAASMLRK-HDPP----
BdPDIL1-2a'  381-431  VKVVVADNIDDIVFNSGKNVLLEFYAPWCGHCRKLAPILEEVAVSFRNDE------D
BdPDIL2-1a    74-124  -FLLTAANFSD-FLSSRRHVMVEFYAPWCGHCQALAPDYAAAASQLALLHQDV----
BdPDIL2-1a'  414-464  VKIIVGKNLDQIVLDESKDVLLEIYAPWCGHCQSLEPTYNKLAKHLRGID------S
BdPDIL3-1a    71-121  VLVLDNENARR-AVEEHAELLLLGYAPWCERSAQLMPRFAEAAAALRAMG-SA----
BdPDIL3-1a'  412-459  ---VVGRTFDSSVLESPHNVFLEAHAPWCVDCEAISKNVEKLAKHFSGLD------N
BdPDIL4-1a°   33-82   VVVLTEGTFEKEVGQD-RGALVEFYAPWCGHCKKLAPEYEKLGASFKKAR------S
BdPDIL4-1a   151-201  VVVLTPETFDSVVLDETKDVLVEFYAPWCGHCKHLAPIYEKLASAFKLDD------G
BdPDIL4-2a°   35-84   VLALTESTFDKEVGQD-RAALVEFYAPWCGHCKKLAPEYEKLAASFKKAK------S
BdPDIL4-2a   153-203  VVVLTEETFDSVVLDETKDVLVEFYAPWCGHCKSLAPVYEKVASAFKLED------G
BdPDIL5-1a°   30-77   VLQLNPNNFKK-VLNANGVVLVEFFAPWCGHCKQLTPTWEKAAGVLKG--------V
BdPDIL5-1a   167-214  -IELNSSNFDELVIKSKDLWIVEFFAPWCGHCKKLAPEWKRAAKNLKG--------Q
BdPDIL6-1a    34-84   VITLTEETFTDKVKEKDTVWFVQFCVPWCKHCKSLGTLWEDLGKVIEGTD------E
BdPDIL7-1a    42-92   VIDLDDSNFEA-ALSSIDFLFVDFYAPWCGHCKRLAPELDEAAPVLAGLS-EP----
BdPDIL7-2a    48-98   VVELDDSNFEA-AAAAVDFLFVDFHAPWCGHCKRLSPQLDEAAPVLAGLS-TP----
BdPDIL8-1a   145-198  --ALSSRNFDS-YSHQYPILVVNFYAPWCYWSNRLKPSWEKTAKIIKERYDPEMDGR

OSPDI1-1a     93-149  IVLAKVDANDEKNKPLATKYEIQGFPTLKIFR--NQGKNIQEYKGPREAEGIVEYLKKQ
OsPDI1-1a'   437-489  VVIAKMDA---TANDVPSEFDVQGYPTLYFVT--PSG-KMVPYESGRTADEIVDFIKKN
BdPDIL1-1a    95-151  IVLAKVDANDEKNKPLAAKYEIQGFPTLKIFR--NQGKNIQEYKGPREAEGIVDYLKKQ
BdPDIL1-1a'  438-490  VVIAKMDA---TANDVPGDFDVQGYPTLYFVT--PSG-KKVAYDGGRTADDIVEYIKKN
BdPDIL1-2a    89-145  VVLAKVDAYDEGNKELKDKYEVHGYPAIKIIR--NGGSDVSGYAGARNADGIVEYLKKQ
BdPDIL1-2a'  432-485  IVIAKMDG---TANDVPTDFVVEGYPALYFYS--SSGGEILSYKGARTAEEIISFIKKN
BdPDIL2-1a   125-176  VALSKVDA--TEDADLAQKYDVQGFPTILFFI--D--GVPKDYTGERTKEAIVAWINK-
BdPDIL2-1a'  465-519  LVIAKMDG---TTNEHP-RAKPDGFPTILFYPAGKKSFEPMTFEGDRTVVEMYKFIKKH
BdPDIL3-1a   122-173  VAFAKLDG--ERYPKAAADVGVSGFPTVLLFV--N--GTEHAYTGLHTKDALVTWVRK-
BdPDIL3-1a'  460-511  LKFARIDA---SVNEHP-KLQVNDYPTLLLYPAEDKS-NPIKVSKKLSLKDMAKFIK--
BdPDIL4-1a°   83-136  VMIAKVDC--DEHKSVCSKYGVSGYPTIQWFP--KGSLEPKKYEGQRTAEALAEFVNK-
BdPDIL4-1a   202-255  VVIANVDA--DKYKDLGEKYGVTGFPTLKFFP--KGNKAGEDYDGGRDLGDFTKFINE-
BdPDIL4-2a°   85-138  VLIAKVDC--DEHKSVCSKYGVSGYPTIQWFP--KGSLEPKKYEGQRTAEALAEYVNS-
BdPDIL4-2a   204-257  VVIANLDA--DKHTSLAEKYGVSGFPTLKFFP--KGNKAGEEYEGGRDLEDFVKFINE-
BdPDIL5-1a°   78-128  ATIAALDA--DAHKELAQQYGIQGFPTIKVFI--PGK-PPVDYEGARDVKPIVNFA---
BdPDIL5-1a   215-266  VKLGHVDC--DSDKSLMSKYKVEGFPTILVFG--ADKESPFPYQGARAASAIESFA---
BdPDIL6-1a    85-136  IEIGKVDCG--ASKPVCSKVDIHSYPTFKVFY--D-GEEVAKYKGPRNVESLKNFVL--
BdPDIL7-1a    93-144  IMVAKVNA--DKYRKLGSKYGVDGFPTLMLFI--H--GVPIEYTGSRKADLLVRNLKK-
BdPDIL7-2a    99-150  VVVAKVDA--EKYKKLGSKYGVDGFPTLMLFD--H--GVPTEYTGSRKADLLIQSLKK-
BdPDIL8-1a   199-243  ILLAKVDCTEEG--ELCKRHHIQGYPSIRIFR--K-GSDMKENQGHHDHE---------
```

Fig. 7.2 "Multiple sequence alignment of *a*-type domains of *B. distachyon* PDI and PDI-like proteins and a typical rice PDI. These thioredoxin-like domains of the *B. distachyon* were annotated in Phytozome database, and comparative analysis used BioEdit software. Residues highlighted in deep blue and green show they were identical and similar, respectively. Open bars and arrowheads represent the α helices and β strands, respectively. The red box indicates the -CxxC- catalytic site, and red arrows indicate the glutamicacid–lysine charged pair. Blue and yellow arrows represent the conserved arginine (R) and the *cis* pralines (P) near the active site, respectively" (Adapted from [76]).

7.6.1 Dynamic Programming

The complex programming algorithms, namely, Smith-Waterman and Needleman-Wunsch, that are employed for a PSA, can also be used for evaluating the optimum alignment of over two sequences. Nevertheless, the difficulty of this algorithm is much shoddier than that of PSA. For performing PSA, the running period of the algorithm is proportionate to m × n, where m and n are the lengths of two aligned sequences. If n ≥ m, the argument is generalized to indicate that the algorithm's

Table 7.2 Softwares and tools used for MSA (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Alignment type[a]	Sequence type[b]	Alignment type[a]	References
ABA	A-Bruijn alignment	GL	PT	GL	[77]
CHAOS, DIALIGN	Iterative alignment	LL (preferred)	Both	LL (preferred)	[78]
ClustalW	PA	LL or GL	Both	LL or GL	[79]
CodonCode aligner	MSA; ClustalW and Phrap support	LL or GL	NT	LL or GL	https://www.codoncode.com/aligner/
Compass	COmparison of multiple PT sequence alignments through statistical assessement	GL	PT	GL	[80]
DECIPHER	Progressive-iterative alignment	GL	Both	GL	[81]
DIALIGN-TX and DIALIGN-T	Segment-based method	LL (preferred) or GL	Both	LL (preferred) or GL	[82]
DNA baser sequence assembler	MSA; full automatic sequence alignment; automatic ambiguity correction; internal base caller; command line seq alignment	LL or GL	NT	LL or GL	https://www.dnabaser.com
DNADynamo	Linked DNA to PT MSA with MUSCLE, Smith-Waterman and Clustal	LL or GL	Both	LL or GL	https://www.bluetractorsoftware.com/
DNASTAR Lasergene molecular biology suite	Software to align RNA, DNA, PT, or DNA + PT sequences via pairwise and MSA algorithms	LL or GL	Both	LL or GL	https://www.dnastar.com/
FAMSA	PA for extremely huge PT families	GL	PT	GL	[83]
FSA	Sequence annealing	GL	Both	GL	http://fsa.sourceforge.net/
Geneious	Progressive-iterative alignment; ClustalW plugin	LL or GL	Both	LL or GL	https://www.geneious.com/
Kalign	PA	GL	Both	GL	[84]

MAFFT	Progressive-iterative alignment	LL or GL	Both	LL or GL	[85]
MARNA	MSA of RNAs	LL	RNA	LL	[86]
MAVID	PA	GL	Both	GL	[87]
MSAProbs	DP	GL	PT	GL	[88]
MULTALIN	DP-clustering	LL or GL	Both	LL or GL	[89]
Multi-LAGAN	Progressive DP alignment	GL	Both	GL	[90]
MUSCLE	Progressive-iterative alignment	LL or GL	Both	LL or GL	[91]
Opal	Progressive-iterative alignment	LL or GL	Both	LL or GL	[92]
Pecan	Probabilistic consistency	GL	DNA	GL	[93]
Phylo	A human computing framework for comparative genomics to solve MSA	LL or GL	NT	LL or GL	[94]
Praline	Progressive-iterative-consistency-homology-extended alignment with preprofiling and secondary structure prediction	GL	PT	GL	[95]
PicXAA	Nonprogressive, maximum expected accuracy alignment	GL	Both	GL	[96]
POA	Partial order/HMM	LL or GL	PT	LL or GL	[97]
Probalign	Probabilistic/consistency with partition function probabilities	GL	PT	GL	[98]
ProbCons	Probabilistic/consistency	LL or GL	PT	LL or GL	[99]
PROMALS3D	PA/HMM/secondary structure/3D structure	GL	PT	GL	[100]
PRRN/PRRP	Iterative alignment (especially refinement)	LL or GL	PT	LL or GL	https://www.genome.jp/tools-bin/prrn
PSAlign	Alignment preserving nonheuristic	LL or GL	Both	LL or GL	[101]
RevTrans	Combines DNA and PT alignment, by back translating the PT alignment to DNA.	LL or GL	DNA/PT (special)	LL or GL	[102]

(continued)

Table 7.2 (continued)

Name	Description	Alignment type[a]	Sequence type[b]	Alignment type[a]	References
StatAlign	Bayesian co-estimation of alignment and phylogeny (MCMC)	GL	Both	GL	[103]
Stemloc	MSA and secondary structure prediction	LL or GL	RNA	LL or GL	[104]
T-coffee	More sensitive PA	LL or GL	Both	LL or GL	[105]
UGENE	Supports MSA with MUSCLE, KAlign, Clustal, and MAFFT plugins	LL or GL	Both	LL or GL	http://ugene.net/
GLProbs	Adaptive pair-HMM based approach	GL	PT	GL	[106]

[a] Alignment type: GL (GL)/LL(LL)

[b] Sequence type: Nucleotide (NT)/proteins(PT)

execution time is n^2. The exponent in the n^2 definition derives from the presumption that, during PSA, if we presume that our sequences length is n, then n × n cells need to be filled within the dynamic programming matrix. If we were to employ either Needleman-Wunsch or Smith-Waterman algorithm to three sequences, we would need to build a 3-dimensional array for measuring and monitoring the alignment. Therefore, for sequences having n length, we will have n × n × n cells for filling in (http://readiab.org/book/0.1.3/2/3). Runtime for MSA employing complete DP algorithms increases dramatically with the sequences number to be aligned. If s and n are the sequence number and sequence length, respectively, then the execution time will be ns. However, in PSA, s = 2, which makes the problem handier (http://readiab.org/book/0.1.3/2/3).

7.6.2 Progressive Alignment

PA is a heuristic approach and does not optimize any obvious alignment score. The aim is to accomplish a series of PSA that begins with aligning nearest identical sequence pairs and subsequently aligning least similar ones [22, 107]. The PA method reduced the overall computational difficulty to polynomial-time by splitting the MSA problem into a set of PSA guided by a tree reflecting the evolutionary sequence relation [108]. Today, most popular alignment programs that employ the progressive approach are ClustalW [79], Mafft ("Multiple sequence alignment based on Fast Fourier Transform") [109], "Multiple sequence comparison by log-expectation" (MUSCLE) [91], and T-Coffee [110].

7.6.2.1 ClustalW

ClustalW is currently the most commonly deployed alignment software, and the oldest of the modules examined. The program conducts a PA, first using PSA through computing the distance matrix that retains the sequence's discrepancy. Just after the matrix is collected, a guided tree is created utilizing Neighbor-Joining algorithms, accompanied by a final stage where the sequences are aligned as per the branching order within the guide tree. In its alignment procedure, the software utilizes two gap penalties: gap expansion and gap opening, during polypeptides availability, a total amino acid weight matrix. These distance penalties rely strongly on variables like sequence length, similarity, and weight matrix. In a simple scenario, Clustal W will exactly match the related domains and sequences of established secondary or tertiary structures but can be seen as a strong starting point for more refinement in more complicated cases (Fig. 7.3a) [73, 79].

7.6.2.2 Mafft

Mafft is a program that can be employed with different alignment methods, either PA alone (with Fast Fourier Transform) or iteratively aligned PA. Mafft's basic run requires up to three stages, but the default procedure performs the first two steps. The first stage is to create a PA centered on each sequence pair's rough distance, on the basis of the mutual 6-tuples. The unweighted pair group method with arithmetic

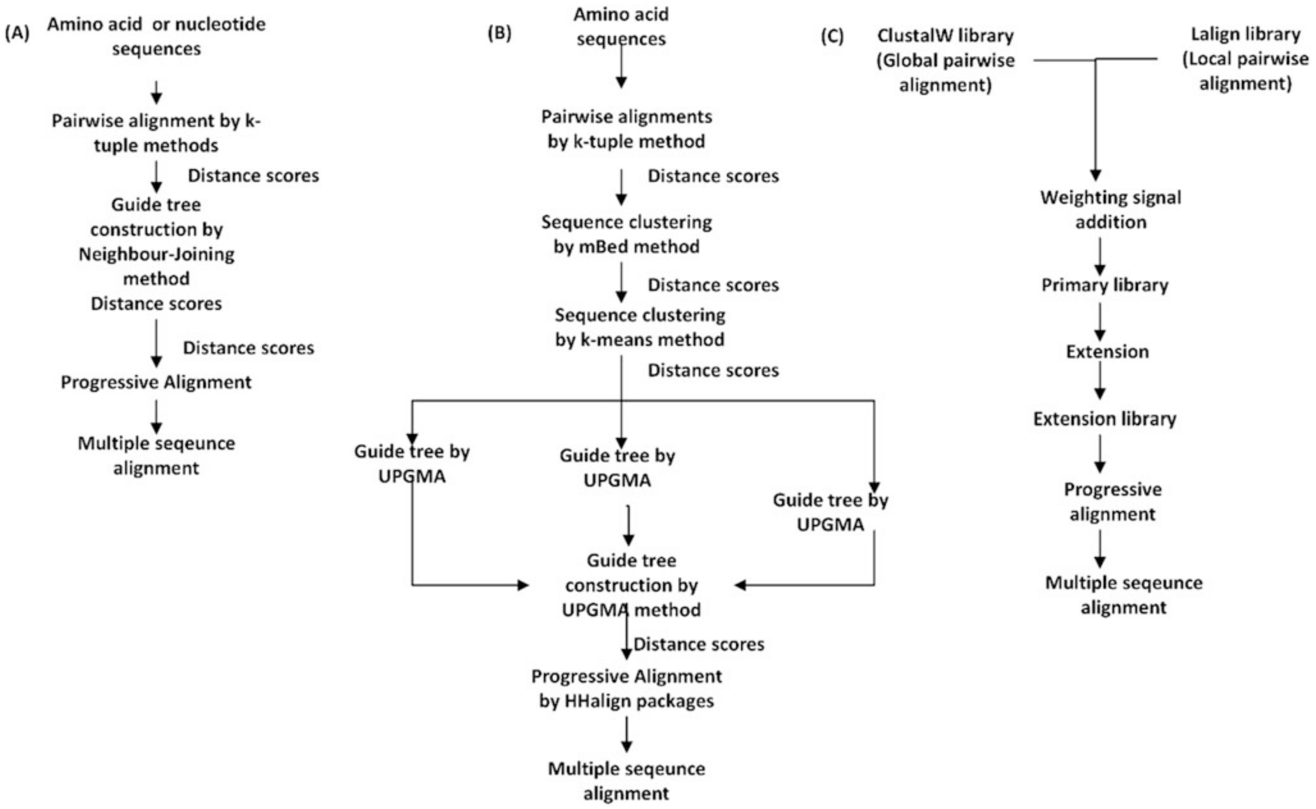

(A)
Amino acid or nucleotide sequences
Pairwise alignment by k-tuple methods
Distance scores
Guide tree construction by Neighbour-Joining method
Distance scores
Distance scores
Progressive Alignment
Multiple seqeunce alignment
(B)
Amino acid sequences
Pairwise alignments by k-tuple method
Distance scores
Sequence clustering by mBed method
Distance scores
Sequence clustering by k-means method
Distance scores
Guide tree by UPGMA
Guide tree by UPGMA
Guide tree by UPGMA
Guide tree construction by UPGMA method
Distance scores
Progressive Alignment by HHalign packages
Multiple seqeunce alignment
(C)
ClustalW library (Global pairwise alignment)
Lalign library (Local pairwise alignment)
Weighting signal addition
Primary library
Extension
Extension library
Progressive alignment
Multiple seqeunce alignment

Fig. 7.3 Steps for generating MSA via (a) ClustalW, (b) Clustal omega, and (c) T-Coffee (Adapted from [75])

Table 7.3 Softwares and tools used for motif scanning (Adapted from https://en.wikipedia.org/wiki/List_of_sequence_alignment_software)

Name	Description	Sequence type[a]	Reference
BASALT	Multiple motif and regular expression scan	Both	http://www.proteinguru.com/toolbox/basalt/
BLOCKS	Ungapped motif prediction from BLOCKS database	Both	https://www.genome.jp/tools/motif/
CUDA-MEME	GPU accelerated MEME (v4.4.0) algorithm for GPU clusters	Both	https://cuda-meme.sourceforge.io/homepage.htm#latest
eMOTIF	Extraction and prediction of shorter motifs	Both	http://motif.stanford.edu/distributions/emotif/
FMM	Motif scan and prediction (can get also positive and negative sequences as input for enriched motif scan)	NT	[128]
Gibbs motif sampler	Stochastic motif extraction by statistical likelihood	Both	[129]
HMMTOP	Prediction of transmembrane helices and topology of PTs	PT	[130]
MEME/MAST	Motif prediction and scan	Both	[125]
MERCI	Discriminative motif prediction and scan	Both	[131]
PHI-blast	Motif scan and alignment tool	Both	[132]
Phyloscan	Motif scan tool	NT	[133]
PMS	Motif scan and prediction	Both	[134]
PRATT	Pattern production for use with ScanProsite	PT	https://www.ebi.ac.uk/Tools/pfa/pratt/
ScanProsite	Motif database scan tool	PT	https://prosite.expasy.org/scanprosite/
TEIRESIAS	Motif extraction and database scan	Both	[135]

[a] Sequence type: Protein (PT) or Nucelotide (NT)

mean (UPGMA) guide tree is then generated with the changed linkage, and the sequences are then aligned with the tree branch order (the so-called FFT-NS-1 strategy). In the second phase, the distance matrix is recalculated based on the knowledge obtained from the previous stage, and the PA is reassessed using a tree from the existing matrix as the starting point (till this process, the technique is known as FFT-NS-2 and is the preferred approach used by the software). The final step is the iterative refinement, which optimizes the "Gotoh weighted pair sum" (WSP) score [111], the "group-to-group alignment" [85], and "the tree-dependent constraint partition technique" [112]. The method is referred to as FFT-NS-i, where all three steps are used, which indicates that it employs the FFT method to conveniently distinguish the homologous regions throughout the sequences followed by the refining iterative process. The FFT converts an amino acid inside a sequence into a vector describing volume and polarity that is key to replacement instances, allowing the software to accurately predict these events [73].

Three additional refining algorithms are also provided by Mafft: L-INS-i, G-INS-i, and E-INS-I [113]. These strategies improve the number of steps required to align the MSA to five. In such instances, the first step would also entail the formation of a distance matrix, not employing six-fold. In comparison to the FFT-NS- * solution, the UPGMA tree is not rebuilt, and the program continues into the second step, splitting gap-free segments and store the scoring arrays from sequence to sequence for each gap-free segment. Mafft subsequently calculates the “importance” value of the segment score and stores the residue in other segments. All “importance” values are then obtained in step three of the “importance” matrix, which is rapidly followed by a group-to-group alignment of scores and a weighting scheme based on the Needleman-Wunsch algorithm [79]. The final stage refines the alignments obtained, increases the WSP score, and the fixed “importance” values. All “importance” values are then obtained in step three of the “importance” matrix, which is rapidly accompanied by a group-to-group alignment of scores and a weighting scheme centered on the Needleman-Wunsch algorithm [79]. The final stage refines the alignments obtained, strengthens the WSP score, and the prescribed “importance” values.

7.6.2.3 Muscle

The muscle uses a pairwise alignment technique to the profile. First, the program establishes a progressive alignment, which is then refined and configured in two following stages. After the similarity of the sequence, the PA is produced, the distance estimation and the UPGMA tree are calculated. Muscle utilizes two distance measurements: a km distance for unaligned series pairs and a Kimura distance for ordered pairs [91]. A new tree with the already defined Kimura distance matrix is generated by the optimization stage of PA, which guarantees a stronger alignment centered on this improved tree. The last step of refinement uses the restricted partition variant tree-dependent [112]. This approach eliminates one of the tree edges, splits the orientation, and eliminates the profiles of the two partitions, which would then be re-aligned with the profile-profile alignment. Each tree edge will be iteratively visited and the alignment with the updated description score of each sequence pair will be preserved. The edges are inspected to minimise the gap from the root by reshaping each sequence and moving to similarly associated sequence classes [91].

7.6.2.4 Clustal Omega

Clustal Omega is the Clustal family’s new MSA algorithm [75]. This algorithm is used only for aligning protein sequences (though nucleotide sequences are likely to be introduced in time). The precision of Clustal Omega is comparable to other high-quality aligners on limited numbers of sequences; moreover, Clustal Omega surpasses other MSA algorithms in terms of completion time as well as overall quality of alignment on large sequence sets. In a few hours, Clustal Omega is able to align 190,000 sequences on a single process. By firstly generating pairwise alignments using the k-tuple form, the Clustal Omega algorithm generates a multiple sequence alignment. Then, employing the mBed method, the sequences are clustered. This is accompanied by the clustering process of k-means. Next, the guide tree is built using the UPGMA method. Finally, using the HHalign module, which aligns two profile hidden Markov models (HMM) as seen in Fig. 7.3b, the multiple sequence alignment is made.

7.6.2.5 T-Coffee

T-Coffee has a radical approach to match sequences. The software first builds a library from two separate sources: Clustal W's global alignment and Lalign's local alignment [114]. Global alignments and pairwise local alignments for each pair of sequences are generated from the top ten nonoverlapping segments. The software processes global and local information and assigns weights to all PSA according to sequence identity [115]. This is accompanied by a mixture of groups that converge into a single repository. This consolidated library has an extension phase, such that the final weight of any pair of residues constitutes part of the information contained in the library. The ultimate step involves calculating the distance matrix and the neighboring joint tree by aligning the two nearest weight sequences on the tree with the stored weight of the consolidated library with a PA. The initial pair is then fixed, and no other gap can be consequently transmitted. The PA will proceed until all sequences fit [73].

Irrespective of their uses, earlier researchers have detected that the majority of PA programs employ the Neighbor-Joining algorithm for inferring a guided tree. Neighbor-Joining's O(N 3) time complexity renders it a bottleneck when large data sets are aligned. The Relaxed Neighbor-Joining algorithm relaxes the joining nodes and decreases standard time complexity to O(N 2 log N) without any major qualitative results [47]. In 2008, Sheneman explored the relationship between the topology of the guide tree and the alignment reliability. He developed two different genetic algorithms, each of which enhances the population of tree guide topologies utilizing stochastic crossover and mutation operators. One genetic algorithm, EVALYN, generates highly accurate scores when evaluated against established reference samples. Nevertheless, we find that the disruptive crossover of EVALYN restricts the genetic algorithm to a stochastic hill climb (Fig. 7.3c).

7.6.3 Probabilistic Alignment

7.6.3.1 PRANK

PRANK [116] is one of the best examples of a probabilistic MSA tool. In comparison to other alignment systems, PRANK uses phylogenetic knowledge to identify alignment differences created through deletions or insertions and then treats the two forms of events differently. As a by-product of the proper handling of inserts and deletions, PRANK will also have assumed ancestral sequences as part of the production and label the alignment gaps differently based on their origin in the insertion or deletion incident. As the algorithm infers the ancestral history of the sequences, PRANK could be vulnerable to errors in the phylogeny guide as well as a violation of basic assumptions about the origin as well as the pattern of the gaps [116].

7.6.3.2 PSAR

In 2014, Kim and Ma developed a new metric, known as PSAR [117], that can metric the reliability of the MSA by agreeing to probabilistically sample Suboptimal Alignments (SAs). The SAs offer extra information which cannot be obtained by optimizing alignment on its own, particularly when the ideal alignment is not too far preferable to the SAs [117].

7.6.3.3 ProbPFP

Recently, Zhan and the team developed ProbPFP that incorporates HMM configured with partition function by particle swarm. The PSO algorithm was used to refine the parameters of the HMM. Subsequently, the posterior likelihood obtained by the HMM was compared with that retrieved through the partition function, and hence the integrated substitution score for the alignment was determined. To test the effectiveness of ProbPFP, 13 excellent or classical MSA methods were compared. The results show that the alignments obtained by ProbPFP have the highest mean SP and TC values for both SABmark and OXBench data sets, as well as the second highest mean TC scores and mean SP scores for BAliBASE. ProbPFP is also compared with four other excellent approaches by restoring phylogenetic trees spanning six protein families in the TreeFam database based on alignments achieved across these five approaches. The results show that the reference trees are like the phylogenetic trees rebuilt from the ProbPFP alignments compared with other approaches [118].

7.6.3.4 ProbCons

ProbCons is a modification of the regular pair-score approach and also provides a secret PA algorithm based on the pair-hidden Markov model. The alignment method is divided into the following steps, starting with the calculation of the reverse likelihood matrices for each pair of sequences. The alignment method is split into the following steps, starting with the calculation of the posterior-probability matrices for each pair of sequences. This is accompanied by a complex software calculation of each PSA's expected accuracy. The probabilistic quality transition is then used to reassess the match's accuracy. A hierarchical clustering determines the guiding tree by the similarities defined by the weighted average of the values between the sequences of every cluster. The guidance tree is employed for matching sequences with a progressive strategy. There is also a postprocessing phase in which random bipartitions of the generated alignment are realigned to find better regions for alignment. ProbCons varies from other alignment systems because it does not implement biological principles like evolutionary tree construction, role-specific gap score, and other features typically utilized with other packages [99].

7.7 Motif Search

Motif exploration is an application layer sequence analysis problem and one of the main obstacles while developing bioinformatics applications. Sequence motifs are constant in size, frequently repetitive and conserved, but at the same moment are small (approximately 6–12 Bp) and very long and are also highly variable in intergenic regions that make the motif discovery a difficult task. A motif is also known as regulatory elements in eukaryotic genes and occurs in the Regulatory Region (RR). These patterns play a crucial role in the identification of the Transcription Factor Binding Sites (TF-BSs), which aid in the understanding of gene expression regulation mechanisms [119, 120]. Motifs are broadly categorized into various forms, namely, sequence motifs, planted motifs, gapped motifs, structured motifs, and network motifs [119]. There are two major forms of algorithms for motif discovery, i.e., enumeration approach probabilistic technique. Enumeration method looks for consensus sequences; motifs are projected dependent on word counts and word similitudes; thus, this method is often named as word enumeration approach to solving Motif problem with panted Motif Problem with motif length and a maximum number of mismatches [120]. The algorithms focused on the word enumeration method extensively scan the entire search field for classifying the ones with potential substitutes, and then normally locate the global optimum. This implies, though, that they are exponential time algorithms that take long for detecting the larger one and inefficient to accommodate hundreds of sequences, and are thus only appropriate for the short motif. Additionally, these algorithms require several user-defined parameters, including the length of the motif, the number of mismatches permitted, and a minimum of sequences the motif requires to appear in [121]. The method to word enumeration can be accelerated by utilizing various data structures, like parallel processing or suffix trees. CisFinder (https://lgsun.grc.nia.nih.gov/CisFinder/), DREME [122], Weeder [123], and MCES [124] are common algorithms based on this method. A second group is a probabilistic method. This constructs a probabilistic model known as Position-Specified Weight Matrix (PSWM) or Motif Matrix, which describes a base distribution to differentiate motifs from nonmotifs for each position of TFBS and needs few search parameters [124]. MEME [125], EXTREME [126], and BioProspector [127] are the most common methods focused on probabilistic approaches. The third form, the nature-inspired approach, incorporates the core attributes of the first two approaches. This method is a basic idea and a global scan but can work with large data and long motifs concurrently. It has a dynamic intention representation, contributing to an infinite range of degenerated positions. The final form is the combinatory method, which depends on the hybrid algorithms which shape the appropriate algorithm.

7.8 Conclusion and Future Perspective

In conclusion, sequence alignment serves as a basic requirement for most of the biological research ranging from phylogenetics construction to protein design. Sequence alignment also employed for motif search in biological sequence, which in turn plays a key role in understanding the regulation of various biological phenomenon. However, because of the continuous increase of sequence amount, there is an urgent requirement of developing novel tools and techniques which can improvise the accuracy of the sequence analysis, including motif search, result obtained. Earlier several researchers have suggested that a successful tool for motif discovery can be constructed from different suggested motif discovery methods. The tool should be fitted with these features: (1) all models should be identified, (2) the overall search feature should be optimized, (3) the parallel processing abilities are needed, (4) optimized data structures should be accessible, (5) the overall search function should be able to locate both long and short motifs, (6) several motif discovery capabilities at the same time, i.e., without elimination of the discovered motif to find another motif. This research would then establish a new algorithm for motif discovery, which incorporates the key characteristics of enumerative and probabilistic approaches and utilizes them as a seed to a naturally inspired algorithm, taking into account the above-noted variables [120].

Conflict of Interest None.

Additional Information Figure 7.1 (CC BY 4.0) [52], Fig. 7.2 (CC BY 4.0) [76], Fig. 7.3 (CC BY 3.0) [75], and Tables 7.1, 7.2, and 7.3 (CC BY-SA 3.0) have been reused under Creative Commons Attribution licenses

References

1. Saeed U, Usman Z. Biological Sequence Analysis. In: Husi H, editor. Computational Biology [Internet]. Brisbane: Codon Publications; 2019. [cited 2020 Oct 13]. Available from: http://www.ncbi.nlm.nih.gov/books/NBK550342/.
2. Prjibelski AD, Korobeynikov AI, Lapidus AL. Sequence Analysis. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology [Internet], Academic Press. Oxford; 2019. p. 292–322. [cited 2020 Oct 11]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128096338201064.
3. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol. 1970;48:443–53.
4. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147:195–7.
5. Wang Y, Wu H, Cai Y. A benchmark study of sequence alignment methods for protein clustering. BMC Bioinformatics. 2018;19:529.
6. Wong KM, Suchard MA, Huelsenbeck JP. Alignment uncertainty and genomic analysis. Sci Am Assoc Adv Sci. 2008;319:473–6.
7. Rosenberg MS. Sequence alignment: Concepts and history. Sequence Alignment: Methods, Models, Concepts, and Strategies. California: University of California Press; 2009. p. 1–22.

8. Koonin EV. Orthologs, paralogs, and evolutionary genomics. Ann Rev Genet. 2005;39:309–38.
9. Koonin EV, Mushegian AR, Bork P. Non-orthologous gene displacement. Trends Genet. 1996;12:334–6.
10. Xiong J. Essential bioinformatics. Cambridge: Cambridge University Press; 2006.
11. Hark Gan H, Perlow RA, Roy S, Ko J, Wu M, Huang J, et al. Analysis of protein sequence/ structure similarity relationships. Biophys J. 2002;83:2781–91.
12. Barton C, Flouri T, Iliopoulos CS, Pissis SP. Global and local sequence alignment with a bounded number of gaps. Theor Comput Sci. 2015;582:1–16.
13. Gotoh O. An improved algorithm for matching biological sequences. J Mol Biol. 1982;162:705–8.
14. Polyanovsky VO, Roytberg MA, Tumanyan VG. Comparative analysis of the quality of a global algorithm and a local algorithm for alignment of two sequences. Algorithms Mol Biol. 2011;6:25.
15. Ye Y, Tang H. Dynamic Programming Algorithms for Biological Sequence and Structure Comparison. Bioinform Algorithms [Internet]. 2007:7–28. [cited 2020 Oct 15]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470253441.ch2
16. Bellman R. Eye of the hurricane. Singapore: World Scientific Publishing Co Pte Ltd; 1984.
17. Sankoff D. The early introduction of dynamic programming into computational biology. Bioinformatics. 2000;16:41–7.
18. Nalbantoğlu ÖU. Dynamic programming. In: Russell DJ, editor. Multiple sequence alignment methods [internet]. Totowa: Humana Press; 2014. p. 3–27. . [cited 2020 Oct 15]. https://doi.org/10.1007/978-1-62703-646-7_1.
19. Giegerich R. A systematic approach to dynamic programming in bioinformatics. Bioinformatics. 2000;16:665–77.
20. Mukhopadhyay CS, Choudhary RK, Iquebal MA. Basic Applied Bioinformatics. Wiley-Blackwell, Hoboken; 2017.
21. Saeed F, Khokhar A. An Overview of Multiple Sequence Alignment Systems. arXiv:09012747 [cs, q-bio] [Internet]. 2009. [cited 2020 Oct 15]; Available from: http://arxiv.org/abs/0901.2747
22. Durbin R, Eddy SR, Krogh A, Mitchison G. Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge: Cambridge University Press; 1998.
23. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25:3389–402.
24. Eddy SR. A Probabilistic Model of Local Sequence Alignment That Simplifies Statistical Significance Estimation. PLOS Comput Biol. 2008;4:e1000069.
25. Frith MC. How sequence alignment scores correspond to probability models. Bioinformatics. 2020;36:408–15.
26. Yu YK, Hwa T. Statistical significance of probabilistic sequence alignment and related local hidden Markov models. J Comput Biol. 2001;8:249–82.
27. Daily J. Parasail: SIMD C library for global, semi-global, and local pairwise sequence alignments. BMC Bioinformatics [Internet]. 2016;17. [cited 2020 Oct 16], Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4748600/
28. Suzuki H, Kasahara M. Introducing difference recurrence relations for faster semi-global alignment of long sequences. BMC Bioinformatics. 2018;19:45.
29. Brenner SE, Chothia C, Hubbard TJ. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci U S A. 1998;95:6073–8.
30. Venclovas C. Comparative modeling in CASP5: progress is evident, but alignment errors remain a significant hindrance. Proteins. 2003;53(Suppl 6):380–8.
31. Sierk ML, Smoot ME, Bass EJ, Pearson WR. Improving pairwise sequence alignment accuracy using near-optimal protein sequence alignments. BMC Bioinformatics. 2010;11:146.

32. Huang W, Umbach DM, Li L. Accurate anchoring alignment of divergent sequences. Bioinformatics. 2006;22:29–34.
33. Stamm M, Staritzbichler R, Khafizov K, Forrest LR. AlignMe—a membrane protein sequence alignment web server. Nucleic Acids Res. 2014;42:W246–51.
34. Aboyoun P. Pairwise Sequence Alignments. p. 34.
35. Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, et al. Human–mouse alignments with BLASTZ. Genome Res. 2003;13:103–7.
36. Hudek AK, Brown DG. FEAST: sensitive local alignment with multiple rates of evolution. IEEE/ACM Trans Comput Biol Bioinform. 2011;8:698–709.
37. Flouri T, Frousios K, Iliopoulos CS, Park K, Pissis SP, Tischler G. GapMis: a tool for pairwise sequence alignment with a single gap. Recent Pat DNA Gene Seq. 2013;7:84–95.
38. Pearson WR. FASTA Search Programs. eLS [Internet]. American Cancer Society; 2014 . [cited 2020 Dec 12]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0005255.pub2
39. Chivian D, Baker D. Homology modeling using parametric alignment ensemble generation with consensus and energy-based model selection. Nucleic Acids Res. 2006;34:e112.
40. Wheeler WC, Gladstein DS. MALIGN: A Multiple Sequence Alignment Program. J Hered. 1994;85:417–8.
41. Wang J, Keightley PD, Johnson T. MCALIGN2: faster, accurate global pairwise alignment of non-coding DNA sequences based on explicit models of indel evolution. BMC Bioinformatics. 2006;7:292.
42. Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. MUMmer4: A fast and versatile genome alignment system. PLOS Computational Biology. 2018;14:e1005944.
43. Cartwright RA. Ngila: global pairwise alignments with logarithmic and affine gap costs. Bioinformatics. 2007;23:1427–8.
44. Girdea M, Noe L, Kucherov G. Back-translation for discovering distant protein homologies in the presence of frameshift mutations. Algorithms Mol Biol. 2010;5:6.
45. Ma B, Tromp J, Li M. PatternHunter: faster and more sensitive homology search. Bioinformatics. 2002;18:440–5.
46. Teichert F, Bastolla U, Porto M. SABERTOOTH: protein structural alignment based on a vectorial structure representation. BMC Bioinformatics. 2007;8:425.
47. Sheneman LJ. The limits of progressive multiple sequence alignment [phd]. [USA]: University of Idaho; 2008.
48. Rucci E, Garcia C, Botella G, De Giusti A, Naiouf M, Prieto-Matias M. SWIFOLD: Smith-Waterman implementation on FPGA with OpenCL for long DNA sequences. BMC Syst Biol. 2018;12:96.
49. Vitkup D, Melamud E, Moult J, Sander C. Completeness in structural genomics. Nat Struct Biol. 2001;8:559–66.
50. Jaroszewski L, Li W, Godzik A. In search for more accurate alignments in the twilight zone. Protein Sci. 2002;11:1702–13.
51. Bergeron BP. Bioinformatics computing. Prentice Hall Professional; 2003.
52. Lin H-N, Hsu W-L. GSAlign: an efficient sequence alignment tool for intra-species genomes. BMC Genomics. 2020;21:182.
53. Huang Y, Zhang L. Rapid and sensitive dot-matrix methods for genome analysis. Bioinformatics. 2004;20:460–6.
54. Waterman MS, Byers TH. A dynamic programming algorithm to find all solutions in a neighborhood of the optimum. Math Biosci. 1985;77:179–88.
55. Zuker M. Suboptimal sequence alignment in molecular biology. Alignment with error analysis. J Mol Biol. 1991;221:403–20.
56. Lathrop RH. The protein threading problem with sequence amino acid interaction preferences is NP-complete. Protein Eng. 1994;7:1059–68.
57. Holmes I, Durbin R. Dynamic Programming Alignment Accuracy. J Comput Biol. 1998;5:493–504.

58. Zhang MQ, Marr TG. Alignment of molecular sequences seen as random path analysis. J Theor Biol. 1995;174:119–29.
59. Miyazawa S. A reliable sequence alignment method based on probabilities of residue correspondences. Protein Eng Des Sel. 1995;8:999–1009.
60. Knudsen B, Miyamoto MM. Sequence alignments and pair hidden Markov models using evolutionary history. J Mol Biol. 2003;333:453–60.
61. Mückstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics. 2002;18(Suppl 2):S153–60.
62. Cline M, Hughey R, Karplus K. Predicting reliable regions in protein sequence alignments. Bioinformatics. 2002;18:306–14.
63. Yu L, Smith TF. Positional statistical significance in sequence alignment. J Comput Biol. 1999;6:253–9.
64. Mevissen HT, Vingron M. Quantifying the local reliability of a sequence alignment. Protein Eng. 1996;9:127–32.
65. Chao KM, Hardison RC, Miller W. Locating well-conserved regions within a pairwise alignment. Comput Appl Biosci. 1993;9:387–96.
66. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A. 1988;85:2444–8.
67. Pearson WR. Effective protein sequence comparison. Methods Enzymol. 1996;266:227–58.
68. Zhang Z, Pearson WR, Miller W. Aligning a DNA sequence with a protein sequence. J Comput Biol. 1997;4:339–49.
69. Farrar M. Striped Smith-Waterman speeds database searches six times over other SIMD implementations. Bioinformatics. 2007;23:156–61.
70. Mackey AJ, Haystead TAJ, Pearson WR. Getting more from less: algorithms for rapid protein identification with multiple short peptide sequences. Mol Cell Proteomics. 2002;1:139–47.
71. Pearson WR. Finding protein and nucleotide similarities with FASTA. Curr Protoc Bioinformatics. 2016;53:3.9.1–25.
72. Ye J, McGinnis S, Madden TL. BLAST: improvements for better sequence analysis. Nucleic Acids Res. 2006;34:W6–9.
73. Nuin PA, Wang Z, Tillier ER. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics. 2006;7:471.
74. Thompson JD, Linard B, Lecompte O, Poch O. A Comprehensive Benchmark Study of Multiple Sequence Alignment Methods: Current Challenges and Future Perspectives. Plos One. 2011;6:e18093.
75. Daugelaite J, O' Driscoll A, Sleator RD. An Overview of Multiple Sequence Alignments and Cloud Computing in Bioinformatics [Internet]. Hindawi: ISRN Biomathematics; 2013. p. e615630. [cited 2020 Oct 17]. Available from: https://www.hindawi.com/journals/isrn/2013/615630/?utm_source=google&utm_medium=cpc&utm_campaign=HDW_MRKT_GBL_SUB_ADWO_PAI_DYNA_JOUR_X_PCUPS&gclid=CjwKCAjwiaX8BRBZEiwAQQxGx2v_vI4i9kMbWescOdwJwv8fn0RGzfe3dBlNeNp-D_OfmWBKpzMnNhoCQ28QAvD_BwE
76. Zhu C, Luo N, He M, Chen G, Zhu J, Yin G, et al. Molecular Characterization and Expression Profiling of the Protein Disulfide Isomerase Gene Family in Brachypodium distachyon L. Plos One. 2014;9:e94704.
77. Raphael B, Zhi D, Tang H, Pevzner P. A novel method for multiple alignment of sequences with repeated and shuffled elements. Genome Res. 2004;14:2336–46.
78. Brudno M, Steinkamp R, Morgenstern B. The CHAOS/DIALIGN WWW server for multiple alignment of genomic sequences. Nucleic Acids Res. 2004;32:W41–4.
79. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
80. Low A, Rodrigue N, Wong A. COMPASS: the COMPletely arbitrary sequence simulator. Bioinformatics. 2017;33:3101–3.

81. Wright ES. DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. BMC Bioinformatics. 2015;16:322.
82. Subramanian AR, Weyer-Menkhoff J, Kaufmann M, Morgenstern B. DIALIGN-T: an improved algorithm for segment-based multiple sequence alignment. BMC Bioinformatics. 2005;6:66.
83. Deorowicz S, Debudaj-Grabysz A, Gudyś A. FAMSA: Fast and accurate multiple sequence alignment of huge protein families. Sci Rep. 2016;6:33964.
84. Lassmann T, Sonnhammer EL. Kalign – an accurate and fast multiple sequence alignment algorithm. BMC Bioinformatics. 2005;6:298.
85. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
86. Siebert S, Backofen R. MARNA: multiple alignment and consensus structure prediction of RNAs based on sequence structure comparisons. Bioinformatics. 2005;21:3352–9.
87. Bray N, Pachter L. MAVID: constrained ancestral alignment of multiple sequences. Genome Res. 2004;14:693–9.
88. González-Domínguez J, Liu Y, Touriño J, Schmidt B. MSAProbs-MPI: parallel multiple sequence aligner for distributed-memory systems. Bioinformatics. 2016;32:3826–8.
89. Mitchell C. MultAlin–multiple sequence alignment. Bioinformatics. 1993;9:614.
90. Brudno M, Do CB, Cooper GM, Kim MF, Davydov E, Program NCS, et al. LAGAN and multi-LAGAN: efficient tools for large-scale multiple alignment of genomic DNA. Genome Res. 2003;13:721–31.
91. Edgar RC. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics. 2004;5:113.
92. Wheeler TJ, Kececioglu JD. Multiple alignment by aligning alignments. Bioinformatics. 2007;23:i559–68.
93. Paten B, Herrero J, Beal K, Fitzgerald S, Birney E. Enredo and pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. Genome Res. 2008;18:1814.
94. Kawrykow A, Roumanis G, Kam A, Kwak D, Leung C, Wu C, et al. Phylo: A Citizen Science Approach for Improving Multiple Sequence Alignment. PLOS ONE. 2012;7:e31362.
95. Simossis VA, Heringa J. PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information. Nucleic Acids Res. 2005;33: W289–94.
96. Sahraeian SME, Yoon B-J. PicXAA-web: a web-based platform for non-progressive maximum expected accuracy alignment of multiple biological sequences. Nucleic Acids Res. 2011;39:W8–12.
97. Lee C, Grasso C, Sharlow MF. Multiple sequence alignment using partial order graphs. Bioinformatics. 2002;18:452–64.
98. Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics. 2006;22:2715–21.
99. Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. ProbCons: probabilistic consistency-based multiple sequence alignment. Genome Res. 2005;15:330–40.
100. Pei J, Kim B-H, Grishin NV. PROMALS3D: a tool for multiple protein sequence and structure alignments. Nucleic Acids Res. 2008;36:2295–300.
101. Sze S-H, Lu Y, Yang Q. A polynomial time solvable formulation of multiple sequence alignment. J Comput Biol. 2006;13:309–19.
102. Wernersson R, Pedersen AG. RevTrans: multiple alignment of coding DNA from aligned amino acid sequences. Nucleic Acids Res. 2003;31:3537–9.
103. Arunapuram P, Edvardsson I, Golden M, Anderson JWJ, Novák Á, Sükösd Z, et al. StatAlign 2.0: combining statistical alignment with RNA secondary structure prediction. Bioinformatics. 2013;29:654–5.
104. Bradley RK, Pachter L, Holmes I. Specific alignment of structured RNA: stochastic grammars and sequence annealing. Bioinformatics. 2008;24:2677–83.

105. Di Tommaso P, Moretti S, Xenarios I, Orobitg M, Montanyola A, Chang J-M, et al. T-coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 2011;39:W13–7.
106. Ye Y, Cheung DW, Wang Y, Yiu S-M, Zhang Q, Lam T-W, et al. GLProbs: aligning multiple sequences adaptively. IEEE/ACM Trans Comput Biol Bioinformatics. 2015;12:67–78.
107. Feng D-F, Doolittle RF. [21] Progressive alignment of amino acid sequences and construction of phylogenetic trees from them. In: Methods in Enzymology [Internet]. London: Academic Press; 1996. p. 368–82. [cited 2020 Oct 17]. Available from: http://www.sciencedirect.com/science/article/pii/S0076687996660236.
108. Maiolo M, Zhang X, Gil M, Anisimova M. Progressive multiple sequence alignment with indel evolution. BMC Bioinformatics. 2018;19:331.
109. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
110. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment1Edited by J Thornton. J Mol Biol. 2000;302:205–17.
111. Gotoh O. A weighting system and algorithm for aligning many phylogenetically related sequences. Comput Appl Biosci. 1995;11:543–51.
112. Hirosawa M, Totoki Y, Hoshida M, Ishikawa M. Comprehensive study on iterative algorithms of multiple sequence alignment. Comput Appl Biosci. 1995;11:13–8.
113. Katoh K, Kuma K, Toh H, Miyata T. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res. 2005;33:511–8.
114. Huang XQ, Hardison RC, Miller W. A space-efficient algorithm for local similarities. Comput Appl Biosci. 1990;6:373–81.
115. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins. 1991;9:56–68.
116. Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 2008;320:1632–5.
117. Kim J, Ma J. PSAR-align: improving multiple sequence alignment using probabilistic sampling. Bioinformatics. 2014;30:1010–2.
118. Zhan Q, Wang N, Jin S, Tan R, Jiang Q, Wang Y. ProbPFP: a multiple sequence alignment algorithm combining hidden Markov model optimized by particle swarm optimization with partition function. BMC Bioinformatics. 2019;20:573.
119. Bataineh MA, Al-qudah Z, Al-Zaben A. Iterative sequential Monte Carlo algorithm for motif discovery. IET Signal Proc. 2016;10:504–13.
120. Hashim FA, Mabrouk MS, Al-Atabany W. Review of different sequence motif finding algorithms. Avicenna J Med Biotechnol. 2019;11:130–48.
121. Zhang Y, Wang P, Yan M. An Entropy-Based Position Projection Algorithm for Motif Discovery. Biomed Res Int [Internet]. 2016. [cited 2020 Oct 19]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5110948/
122. Bailey TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics. 2011;27:1653–9.
123. Pavesi G, Mereghetti P, Mauri G, Pesole G. Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. Nucleic Acids Res. 2004;32:W199–203.
124. Yu Q, Huo H, Chen X, Guo H, Vitter JS, Huan J. An efficient algorithm for discovering motifs in large DNA data sets. IEEE Trans Nanobioscience. 2015;14:535–44.
125. Bailey TL, Johnson J, Grant CE, Noble WS. The MEME suite. Nucleic Acids Res. 2015;43:W39–49.
126. Quang D, Xie X. EXTREME: an online EM algorithm for motif discovery. Bioinformatics. 2014;30:1667–73.
127. Liu X, Brutlag DL, Liu JS. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput. 2001:127–38.

128. Sharon E, Lubliner S, Segal E. A Feature-Based Approach to Modeling Protein–DNA Interactions. PLoS Comput Biol [Internet]. 2008;4. [cited 2020 Dec 13], Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2516605/
129. Thompson W, McCue LA, Lawrence CE. Using the Gibbs motif sampler to find conserved domains in DNA and protein sequences. Curr Protoc Bioinformatics. 2005. Chapter 2:Unit 2.8.
130. Tusnády GE, Simon I. The HMMTOP transmembrane topology prediction server. Bioinformatics. 2001;17:849–50.
131. Vens C, Rosso M-N, Danchin EGJ. Identifying discriminative classification-based motifs in biological sequences. Bioinformatics. 2011;27:1231–8.
132. Zhang Z, Miller W, Schäffer AA, Madden TL, Lipman DJ, Koonin EV, et al. Protein sequence similarity searches using patterns as seeds. Nucleic Acids Res. 1998;26:3986–90.
133. Carmack CS, McCue LA, Newberg LA, Lawrence CE. PhyloScan: identification of transcription factor binding sites using cross-species evidence. Algorithms Mol Biol. 2007;2:1.
134. Dinh H, Rajasekaran S, Kundeti VK. PMS5: an efficient exact algorithm for the (ℓ, d)-motif finding problem. BMC Bioinformatics. 2011;12:410.
135. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics. 1998;14:55–67.

Gene Identification and Structure Annotation

8

Puja Sashankar, Santhosh N Hegde, and N. Sathyanarayana

Abstract

Rice (*O. sativa* L.) is one among the necessary food crops worldwide. Due to ever-increasing demand, many are undertaking several efforts to enhance its productivity - the latest being the sequel of Rice genome sequencing projects. The accelerated developments in next-generation sequencing (NGS) has bolstered these efforts in hundreds to thousands of rice varieties, which has enabled researchers to unpack the hidden potential of vast and diverse rice germplasm. One of the important objectives of these projects is to accurately characterize the gene models, which has a major significance for the in-depth study of gene function and, thus, various applications. Bioinformatics plays a major role in gene structure identification and its biological function through various algorithms and software. Hence, this chapter aims to elucidate the approach of identifying, characterizing, and finding the function of different types of rice genes.

Keywords

Rice · Annotations · Gene Prediction · Coding genes · Exon

P. Sashankar · N. Sathyanarayana (✉)
Molecular Biology and Biotechnology Laboratory, Department of Botany, Sikkim University, Gangtok, Sikkim, India
e-mail: nsathyanarayana@cus.ac.in

S. N. Hegde
Center for Functional Genomics and Bioinformatics, University of Trans-Disciplinary Health Sciences and Technology (TDU), Bengaluru, Karnataka, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_8

Abbreviations

CDS	Coding DNA Sequence
CGSNL	Committee on Gene Symbolization Nomenclature and Linkage
ESTs	Expressed Sequence Tags
FLcDNA	Full-length complementary DNA
IRGSP	The International Rice Genome Sequencing Project
miRNA	microRNA
MSU	Michigan State University
NGS	Next-Generation Sequencing
NIAS	National Institute of Agrobiological Sciences
OMAP	Oryza Map Alignment Project
ORF	Open Reading Frame
RAP	Rice Annotation Project
RGAP	Rice Genome Annotation Project
RGKbase	Rice Genome Knowledgebase
RMD	Rice Mutant Database
rRNA	ribosomal RNA
snoRNA	small nucleolar RNA
TE	Transposon Element
tRNA	transfer RNA

8.1 Introduction

Rice (*Oryza sativa L.*) is the foremost vital staple food crop worldwide, which provides food for 3/4th of the world human population [1]. There are 24 species of *Oryza*, among which two major species, i.e., *O. sativa* and *O. glaberrima,* are cultivated worldwide. The rice cultivated in Asia, *O. sativa,* is grown in almost all parts of the world, and, *O. glaberrima,* considered to be the rice cultivated and grows on a small scale in the Western part of Africa [2]. The species *O. sativa* originated from the southern and eastern parts of Asia and is divided into two other subspecies, such as japonica and indica [3]. Genome-wide research on the diversity of two varieties of *Oryza sativa* (indica and japonica) showed that the rice originated genetically from different gene pools, having single wild ancestor species, *Oryza rufipogon*, indicating various domestications of *O. sativa* [4, 5]. Genetic studies revealed a profound population structure among domesticated species across several approaches. However, indica and japonica are two major subspecies that were found to have various subpopulations among each group [6].

Rice is the smallest of the genomes among the cereal crops in terms of genome size. Due to this, it is considered to be a model organism for several genomic studies on monocotyledonous plants [7]. “The International Rice Genome Sequencing Project (IRGSP)” started in the year 1997 and with the completion of sequencing of Nipponbare, a japonica type of rice cultivar, and generated high-quality genome sequences in 2005, i.e., “International Rice Genome Sequencing Project 2005”, and these data were further filtered and improved in the year 2013 [8]. This effort is so far precise and till today it aids as an important base for cereal crop genomics study as well as evaluation [9]. Afterward, another two rice genome assemblies were generated separately. The first by the “Rice Genome Annotation Project” primarily situated at “The Institute for Genomic Research” and currently in the “Michigan State University (MSU)” and the other one by the “Rice Annotation Project (RAP)” [8, 10]. The size of the finalized IRGSP genome when published was 370 Mb in length, covering 95% of the sequenced genome of rice altogether. Though, later, a relative study of the two other reported draft genomes and IRGSP revealed that the overall coverage was 69% for indica and 78% for japonica [11]. However, to gain detailed insight into the genetic composition, genome annotation is highly required. The genome annotation is the process of identifying the functional elements within a genome using different bioinformatics algorithms and software. The genome annotation process is very important in identification of candidate genes and will enhance the genetic improvement of the crops. Thus, in this chapter, we described the process of in silico gene identification, aspects of gene identification, and structure annotation in rice through different software and some of the major databases of rice gene information.

8.2 Gene Annotation

The major step in genome analysis is the annotation of the genome. In several scientific researches, genome annotation seems to be essential for retrieving any information about the genome. Annotation is the identification of genes, its structure, and other miscellaneous features in the genome as well as finding biological functions of identified features. The genome annotation is classified into structural and functional annotation. Structural annotation involves finding the locations of genes in the chromosomes, gene structure such as exons, introns, UTR, etc., and finding other genomic features such as repeats, transposable elements, etc. Structural annotation of the genome can be done through two methods, namely, ab-initio gene prediction and homology-based gene prediction. The ab-initio gene prediction method predicts the gene structure based on the different patterns or features of the gene. The homology-based methods involve the prediction of genes based on the significant similarity between query sequences and already identified gene sequences. Most of the gene prediction algorithms or tools make use of both methods.

8.2.1 Ab-initio Gene Prediction

The ab-initio gene prediction method involves identification of genes using gene features including start and stop codons, intron–exon junctions, binding sites for a transcription factor, etc. In addition, the ab-initio gene prediction algorithm uses gene content, which is a statistical overview of the coding part of a gene. The gene features will be identified by ab-initio methods using mathematical and statistical methods. Many algorithms are applied for in ab-initio gene prediction, such as Dynamic Programming, linear discriminant analysis, Linguist methods, Hidden Markov Model (HMM), Support Vector Machines (SVMs), and Neural Network to help differentiate between coding and noncoding regions of genes. Based on these models, a great number of ab-initio gene prediction programmes have been developed such as Genscan [12], Gene Locator and Interpolated Markov ModelER (GLIMMER) [13], FGENESH (http://linux1.softberry.com/), Augustus [14], and MAKER [15]. GLIMMER is one of the gene predictions tools for the microbial genome. To classify the coding regions and separate them from noncoding DNA regions, interpolated Markov models (IMMs) are used in GLIMMER [13]. FGENESH is the efficient and most precise gene finder available, and it uses HMM for gene structure prediction. It also predicts genes using homology-based methods. GENSCAN is a programme for the prediction of complete structures of genes including gene location and their intron–exon boundaries in genomic DNA in a variety of species [12]. AUGUSTUS is one of the best eukaryotic gene prediction pipelines, which uses both de novo and homology-based methods to predict accurate gene models [14]. MAKER is one of the easiest automated genome annotation workflows for identifying genes and repeats. Firstly, it uses ab-initio gene prediction method to identify the genes in the genome, and after that, it uses EST and protein evidence to increase the predicted gene models with the ab-initio method [15]. However, the most successful programmes so far are based on the Hidden Markov Model. The ab-initio gene prediction programmes used in the benchmark analysis are focused on mathematical models that are qualified and usually perform well in identifying conserved or well-studied genes [16].

Ab-initio gene prediction method mainly depends on two types of sequence information: signal sensors and content sensors. Signal sensors involve identifying gene-specific sites and patterns such as alternative splicing sites, promoter sequences, and transcription start sites. Content sensors involve identifying coding and noncoding patterns such as exon or intron lengths or nucleotide composition. Gene prediction algorithms often make serious errors such as in the resulting gene models include missing exons, noncoding sequence retention in exons, fragmenting genes and merging neighboring genes and can affect downstream analyses, including functional annotations, gene recognition, evolutionary studies, etc. This is particularly true in the case of large “draft” genomes, where an imperfect genome assembly, poor coverage, low consistency, and high complexity of the gene structures are usually encountered by the researcher [17].

8.2.2 Homology-Based Methods

The homology-based approach includes sequence similarity search, which is a conceptually straightforward approach based on identifying similarities between gene sequences of input genome and ESTs (expressed sequence tags), proteins, or other genomes. This strategy is based on the assumption that evolutionarily coding regions (exons) are more conserved than noncoding regions (intergenic or intronic regions). The similarity information between certain genomic regions of the input genome and an EST, DNA, or protein can be used to infer gene structure or function of that region. The similarity to known proteins of closely related organisms is used to predict the coding potential of the gene. The similarity of EST sequences to the input genome is to detect the alternative splicing. The similarity of EST sequences to the input genome typically has limitations in that ESTs often refer to small portions of the gene which code, which means that the full gene structure of a given region is often difficult to predict. The greatest drawback of this method is that only about half of the genes being discovered have significant homology to genes in the databases. And in some cases where EST sequences or closely related genomes are not available, it is difficult to predict genes. Since this approach is based on similarity to the known genes, proteins, or ESTs, the probability of predicting novel genes is very less.

The major goal of genome annotation is to identify genes and their function computationally with near 100% accuracy. In several scientific researches, genome annotation seems to be essential for any information on the genome. However, particularly for eukaryotes, the gene prediction is highly difficult because of many computational problems. The major principle of gene prediction is pattern matching to find the gene features. Normally, there are no retained motifs for coding regions. Detection of the coding potential of genomic regions must depend on subtle gene-related features that may be very difficult to identify. In the case of homology-based methods, there are a large number of genes, whose functions were not identified yet. Hence, the combination of both HMMs (ab-initio) and homology-based algorithms can produce improved accuracy.

Once the gene models are predicted, the next step is to identify the biological functions of these gene models, called a functional annotation. This involves finding significant matches to the predicted genes using database search using various alignment programs. For instance, if a translated DNA sequence of a predicted gene is found to have significant similarity with a known protein using a database search, then predicted gene function is also similar to that known protein. Approaches for function annotation of genes and its products, i.e., protein has been discussed in detail in Chaps. 7 and 11 of this book.

8.3 Gene Annotations in Rice

The rice genome annotation was conducted using prediction-based and homology-based searches. The "Institute for Genomic Research" (TIGR), which again is part of the IRGSP, started annotation project on rice gene, known commonly as the "Rice Genome Annotation Project" (RGAP), prior to the completion of the IRGSP genome sequencing project, and subsequently reported the results to the scientific world [18]. The official project on genome annotation, the "Rice Annotation Project"

(RAP), an initiative of the “National Institute of Agrobiological Sciences” (NIAS) in Japan, has been initiated by IRGSP representatives, and data are released by RAP-DB. The National Agriculture and Food Research Organization currently manages the RAP-DB. Experimental data from closely related sequences, for example, full-length rice cDNA sequences, ESTs (expressed sequence tags) of rice, and RNA seq/protein sequences have been used in order to improve the gene models in the rice annotations [11].

8.3.1 Gene Structure and Function Prediction

The steps involved in the rice genome annotation, which is published in the Rice Genome Annotation Project, are as follows. The first steps involve assembling BAC/PAC clones into 12 pseudomolecules. The second step is to annotate the 12 pseudomolecules, which involves the prediction of gene models and other miscellaneous features using bioinformatic algorithms and software such as FGENESH, Genemark.hmm, Genscan, GeneSplicer (to predict exons and introns), tRNAscan-SE (to predict tRNA) with Maize and *Arabidopsis thaliana* as reference gene models. The next step is to add biological information to the predicted genes. Rice ESTs and FL-cDNAs and transcript assemblies (PUTs) from the PlantGDB were aligned to the pseudomolecules using gmap. Only the EST and FL-cDNA alignments were used for gene model improvement by PASA. The repetitive elements and transposons (DNA transposons, retroelements, MITEs, etc.) were identified by searching the sequence of each pseudomolecule. RepeatMasker was used to identify the Simple repeats (Fig. 8.1) [19, 20].

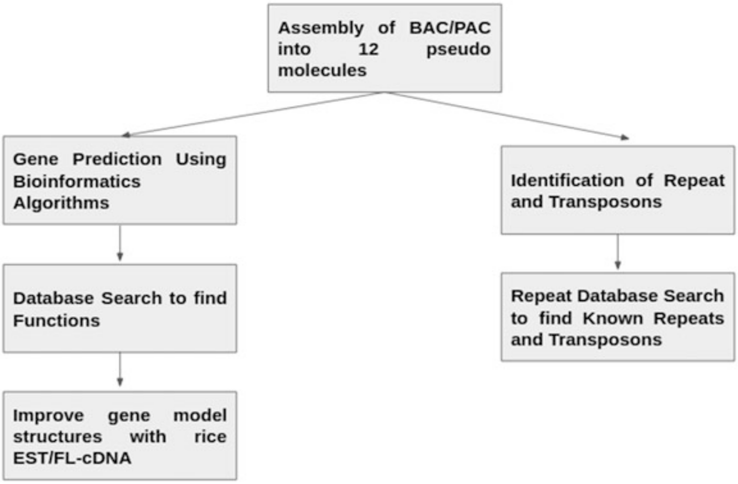

Assembly of BAC/PAC into 12 pseudo molecules
Gene Prediction Using Bioinformatics Algorithms
Identification of Repeat and Transposons
Database Search to find Functions
Repeat Database Search to find Known Repeats and Transposons
Improve gene model structures with rice EST/FL-cDNA

Fig. 8.1 Flow chart showing annotation in rice explained in the Rice Genome Annotation Project

8.3.2 Criteria for the Definition of Genes [19, 20]

- The models of genes are named after matching to proteins in the database to indicate similarity.
- In particular, gene models are annotated as "xxxx, putative" with more than 30 percent identification and greater than 50 percent coverage.
- Genes with transcript evidence are additionally annotated as "expressed."
- Genes were annotated as "conserved hypothetical protein," if they were not validated by transcript evidence but correlate with protein sequences from a recognized gene.
- Predicted genes are classified as "hypothetical proteins" when they are not aligned with known genes, transcripts, or protein sequences.
- Gene models that do not have homology to known genes or proteins but that are supported by rice transcript evidence are labeled as "expressed protein."

The system of gene nomenclature for rice was done by the "Committee on Gene Symbolization, Nomenclature and Linkage (CGSNL) of the Rice Genetics Cooperative" to enhance the methods of identifying, characterizing, and describing the genes of rice. This explains the classification of rice genes into different groups built on their sequence resemblance to earlier annotated genes. The predicted genes in the rice, based on similarity of the sequence, have been classified into five major groups. Only if there is significant experimental evidence that a gene has a sequence similarity to a formerly annotated rice gene with an identified feature, genes are then given a gene name and a gene symbol. Gene functions were assigned categories II–V for the genes with insufficient evidence [21].

With the above-explained approach by The Rice Genome Annotation Project, the genes were annotated with Transposon Element loci and without Transposon Element Loci (Table 8.1).

8.4 Types of Genes in Rice

8.4.1 Protein-Coding Genes

A gene consists of four major components: 5'-UTR, coding regions (cds), intron, and 3'-UTR [22]. An international consortium for rice genome annotation, the Rice Annotation Project (RAP), has been created to standardize the annotated genome

Table 8.1 Summary of annotated genes in rice genome from the "Rice Genome Annotation Project" [8]

Class	Number	Gene models	Gene size	Exons/gene	Introns/gene
Non-TE loci	39,045	49,066	2853 bp	4.9	3.9
TE loci	16,941	17,272	3223 bp	4.2	3.2

[a] *TE* Transposon Element

data for Oryza sativa subspecies Nipponbare [23]. *Oryza sativa* haploid genome (chromosome number 12) contains about 36,000 protein-coding genes. Over the past decade, genomes of cultivated rice and their wild varieties have been sequenced, but the most efforts are focused on genome assembly and gene prediction of two primary rice cultivars, 93–11 (indica) and Nipponbare (japonica) [24]. The Rice Annotation Project has taken the annotation of the Rice genome after completion of rice genome assembly from IRGSP. Using an ab-initio gene prediction method, 37,544 protein-coding genes associated with non-transposable elements were identified, and 2859 rice genes were not formally reported in the genome of Arabidopsis [11]. The average gene size in rice is about 2853 bp. Table 8.2 shows the statistics of genes in each chromosome of rice.

There is a total of 37,860 gene locus and 44,924 transcripts; which were supported by FL-cDNAs, Expressed Sequence Tags (ESTs), or proteins in the rice genome sequenced and assembled by IRGSP. Figure 8.2 shows gene distribution in all 12 rice chromosomes of both japonica and indica based on the recent proteomes submitted to the Uniprot database (https://ebi16.uniprot.org/).

8.4.2 Exons

An exon is the portion of a gene that codes for amino acids. They are that part of the mRNA that code for proteins. From the sequence analysis, it is observed that the average length of exon in rice genes decreases with an increasing number of introns. Linear correlation is observed between the total exon length and the number of introns, and the length of a gene depends on the number of introns [26]. There is a total of 166,057 exons in the rice genome with an average length of 193 bp, which contains 48.6% of GC content. Table 8.3 shows the statistics of exons in rice.

8.4.3 Introns

Introns are the noncoding regions that disrupt some coding regions in the genome. Rice introns are generally longer and have higher GC content [28]. There is a total of 111,343 introns in the rice genome with an average length of 433 bp, which contains 37.3% of GC content (Table 8.4).

8.4.4 Pseudogenes

The gene that has homology to known protein-coding genes but contains a frame-shift and/or stop codon, which disrupts the ORF is called Pseudogenes. These pseudogenes are thought to have arisen from duplication followed by loss of function. A total of 1439 pseudogenes were predicted in the rice genome, which was predicted based on similarity with (proper) fully supported models of the gene and also the existence of the premature translational stop codons or frameshifts. For

Table 8.2 Chromosome level gene statistics from the Rice Annotation Project Database (October 02, 2012) [25]

Chromosome	Protein-coding loci with rice FLcDNAs	Protein-coding loci without rice FLcDNAs	Nonprotein-coding loci with rice FLcDNAs	Nonprotein-coding loci without rice FLcDNAs	Alternative variants	Ab-initio predicted genes with evidence of expression	Ab-initio predicted genes without any transcripts evidence	Loci with CGSNL annotation
1	3421	1247	93	211	985	301	943	240
2	2737	994	71	162	821	252	754	188
3	3050	1031	76	203	972	190	773	231
4	2135	805	55	145	609	219	752	133
5	1968	719	35	133	547	172	649	145
6	1896	859	64	122	512	218	655	148
7	1830	696	40	101	465	203	664	108
8	1602	666	44	104	449	172	633	113
9	1321	525	39	79	355	147	541	89
10	1230	609	38	95	321	143	494	73
11	1386	653	50	100	294	235	687	61
12	1367	532	48	82	337	148	576	79
Total	**23,943**	**9336**	**653**	**1537**	**6667**	**2400**	**8121**	**1608**

[a] *FLcDNA* Full-length complementary DNA

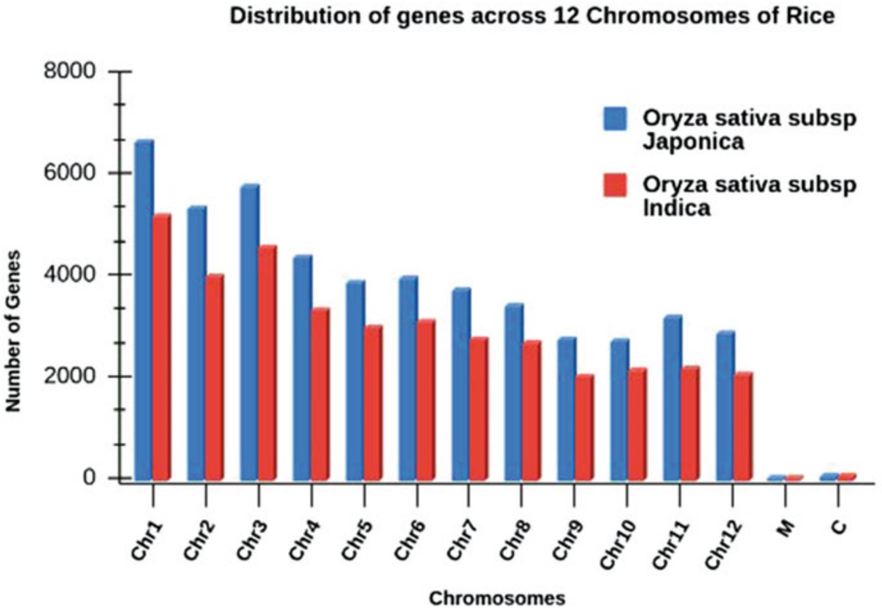


Fig. 8.2 Distribution of genes across the chromosomes of rice (japonica and indica subspecies). *M* Mitochondria, *C* Chloroplast

Table 8.3 Statistics of Exon in rice [27]

Contents	Number
Total Exons	166,057
Internal Exons	92,336
Average length, bp	193
Median length, bp	119
GC content	48.6%

Table 8.4 Statistics of Introns in rice [28]

Contents	Number
Total introns	111,343
Long introns (>1 kb)	11,541 (10.4%)
Average length, bp	433
Median length, bp	160
GC content	37.3%

816 pseudogenes, out of which 75% resulted from events like gene duplication, and the rest 25% were likely to be the product of retrotransposition, the likely origin could be determined. A total of 12% of the pseudogenes have gene expression evidence. F-box, BTB/POZ, terpene synthases, chalcone synthases, and cytochrome P450 protein families were coded by a significant number of pseudogenes [29].

8.4.5 Noncoding RNA Genes

Noncoding RNAs (ncRNAs) are the RNAs which are functional and transcribed from DNA but not translated into protein and which regulate the gene expression. These noncoding RNA genes include microRNA (miRNA), tRNA (transfer RNA), and rRNA (ribosomal RNA) genes. The short arm of chromosome 9 in *O. sativa* ssp. *japonica*, in particular at the telomeric end, consists of ribosomal DNA coding units called17S–5.8S–25S. These 17S–5.8S–25S rDNA locus in *O. sativa* ssp. *indica* was also found at the end of the short arm but in a different chromosome that is at chromosome 10 [25]. In all 12 chromosomes, 763 transfer RNA genes have been found, along with 14 tRNA pseudogenes. There is a single tRNA cluster in chromosome 4 and in chromosome 10, there are two large clusters, originated through inserted chloroplast DNA. The large clusters appear only in the regions of intermediate density on chromosomes like 1, 2, 8, and 12, apart from these they do not exist in the other regions.

MicroRNAs (miRNAs) is a class of eukaryotic noncoding RNAs, known to interact with the target mRNA and regulate gene expression [30]. In total, 158 miRNAs are identified in chromosomes of rice. 93 spliceosomal RNA genes and 215 small nucleolar RNA (snoRNA) were spotted, along with other noncoding RNAs [31].

8.5 List of Rice Gene Databases

Rice gene database provides genome sequence as well as annotation information of all the 12 chromosomes of the rice with genome browsers, providing an integrated display of annotation data [8]. The primary objective of the rice gene databases is to provide an effective and consistent annotation of the genome sequence of rice to the research world. The facilitation of a systematic study of the genomic structure as well as the function of rice, based on annotation, is one of the key goals of the rice gene databases [25]. Information about different rice databases is described in detail in Chap. 3. Some of the major rice gene annotation databases are given below.

8.5.1 Rice Genome Annotation Project

The Michigan State University (MSU) Rice Genome Annotation Project is a National Science Foundation project and is a major database for the rice gene annotation which provides rice genome sequence and annotations, particularly Nipponbare subspecies. All 12 chromosomes of Nipponbare have been annotated and the information can be available on the search page of the database. In the database, the rice genome browser reveals the annotated data across various tracks [8] (http://rice.plantbiology.msu.edu/).

8.5.2 The Rice Annotation Project (RAP-DB)

The “Rice Annotation Project (RAP-DB)” gives the recent and updated annotation of the rice genome using the “Os-Nipponbare-Reference-IRGSP-1.0,” which is developed in association with the MSU Rice Genome Annotation Project and RAP. At first, the IRGSP rice genome sequences were scrutinized using sequencing reads with ~44X coverage obtained by next-generation sequencing platforms, and secondly, the sequencing errors were corrected properly. The key objective of the RAP-DB project is to facilitate an overall study of the annotation-based gene structure and function of rice. The database also contains tools such as BLAST, BLAT, Gbrowse for rice genes and genome, etc. [25] (https://rapdb.dna.affrc.go.jp/).

8.5.3 Oryzabase

Oryzabase is a rice database which integrates the biological and genome data of rice and was created in Japan by a group of rice researchers in the year 2000. Initially, the database aims to collect as much expertise as possible, covering from classical genetics of rice to modern genomics and also from basic details to other relevant information of rice. This database also offers a detailed view of rice by combining molecular genomic knowledge with biological data. Oryzabase includes many genetic, physical, and expression maps that combine biological sequences with complete genome and cDNA sequences [32] (https://shigen.nig.ac.jp/rice/oryzabase/).

8.5.4 Rice Mutant Database (RMD)

The “Rice Mutant Database (RMD)” is a combined national programme developed by the Wuhan Group, China’s “National Special Key Program” on Functional Genomics of rice, and sustained by the Huazhong Agricultural University’s (Wuhan), which comprises thorough information on roughly 129,000 transfer DNA insertion (enhancer trap) lines of rice and comprehensive information on mutant phea traps. You can scan for RMD using keywords and nucleotide or protein sequences. RMD offers mainly three types of functions, such as the novel genes finding, the identification of regulatory elements, and the identification precise growth phases by predicting ectopic expression (misexpression) pattern lines for the targeted gene in particular tissues [33] (http://rmd.ncpgr.cn/).

8.5.5 RiTE DB

Arizona Genomics Institute created Rice TE Database, which gathers repeat DNA sequences and transposable elements (TEs) of various closely related rice species. This database facilitates the annotation of TE and repetitive sequences and genomic analysis of a multiple genome data, established for the Oryza Genome Evolution project, which is a part of the "International Oryza Map Alignment Project (OMAP)" [34] (https://www.genome.arizona.edu/cgi-bin/rite/index.cgi).

8.5.6 RiceFREND

RiceFREND contains gene coexpression data for rice and offers a platform for the prediction of genes which are functionally related in several pathways and/or metabolic processes in rice with a large set of microarray data obtained at various phases of growth and development from different rice tissues [35]. The gene expression data used for coexpression data analysis is accessed from RiceXPro, a repository of gene expression information gathered from microarray analysis [36] (https://ricefrend.dna.affrc.go.jp/).

8.5.7 The Rice Genome Knowledgebase (RGKbase)

RGkbase is a database majorly developed for the relative study of the rice genome which focuses on three key components. Firstly, for rice genomics and molecular biology, integrated data curation. Secondly, user-friendly interface in terms of annotation of genome and evolutionary dynamics to access the data easily and finally, the bioinformatic methods for descriptions of gene ontology, pathway analysis, and classifications of the gene family. RGKbase currently provides genomic data for five varieties of rice [37] (http://rgkbase.big.ac.cn/RGKbase/).

8.6 Conclusion and Future Perspective

Rice is one of the major food crops which most of the world's population feeds. In terms of genome size, rice is the smallest of the cereal crops' genomes with the genome size of 400–430 Mb. The first rice genome sequencing was carried out by "The International Rice Genome Sequencing Project (IRGSP)" which started in the year 1997 and generated high-quality genome sequences of Nipponbare, a japonica type of rice cultivar in 2005. The genome annotation is an important step in any genomic research. The rice genome annotation was conducted using prediction-based and homology-based searches by using various bioinformatics algorithms. Employing gene annotation approach genes and other miscellaneous features of genome such as noncoding RNA genes, transposable elements, etc., have been identified with their biological function. The bioinformatic workflow for annotation of rice involved several steps including gene structure prediction (identification of introns, exons, UTRs, etc.), finding their biological function, and validating the genes predicted based on protein or EST evidence. The availability of rice genome annotation data would improve plant breeders to recognize candidate genes and enhance novel breeding strategies to overcome future challenges.

Conflict of Interest None.

References

1. Eswaran R, Sofiya M, Anbanandan V. Identification of cold tolerant rice genotypes and associated traits at seedling stage. J Pharmacogn Phytochem. 2019;8(2S):774–6.
2. Khush GS. Origin, dispersal, cultivation and variation of rice. Plant Mol Biol. 1997 Sep 1;35 (1):25–34.
3. Fuller DQ, Sato Y-I, Castillo C, Qin L, Weisskopf AR, Kingwell-Banham EJ, et al. Consilience of genetics and archaeobotany in the entangled history of rice. Archaeol Anthropol Sci. 2010 Jun 1;2(2):115–31.
4. Kovach MJ, Sweeney MT, McCouch SR. New insights into the history of rice domestication. Trends Genet. 2007 Nov 1;23(11):578–87.
5. Song S, Tian D, Zhang Z, Hu S, Yu J. Rice genomics: over the past two decades and into the future. Genomics Proteomics Bioinformatics. 2018;16(6):397–404.
6. Sweeny M, McCouch S. The complex history of the domestication of rice. Ann Bot. 2007;100:951–7.
7. Satoh K, Doi K, Nagata T, Kishimoto N, Suzuki K, Otomo Y, et al. Gene Organization in Rice revealed by full-length cDNA mapping and gene expression analysis through microarray. PLoS One. 2007 Nov 28;2(11):e1235.
8. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013 Feb 6;6(1):4.
9. Matsumoto T, Wu J, Itoh T, Numa H, Antonio B, Sasaki T. The Nipponbare genome and the next-generation of rice genomics research in Japan. Rice. 2016 Jul 22;9(1):33.
10. Eckardt NA. Sequencing the Rice genome. Plant Cell. 2000 Nov 1;12(11):2011–7.
11. Kumagai M, Tanaka T, Ohyanagi H, Hsing Y-IC, Itoh T. Genome sequences of Oryza species. In: Sasaki T, Ashikari M, editors. Rice genomics, genetics and breeding [internet]. Singapore: Springer; 2018. p. 1–20. [cited 2021 Jan 16]. Available from: https://doi.org/10.1007/978-981-10-7461-5_1.
12. Burge CB. Identification of Genes in Human Genomic DNA. [PhD Thesis]. Stanford University Stanford, CA; 1997.
13. Delcher AL, Bratke KA, Powers EC, Salzberg SL. Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics. 2007 Mar 15;23(6):673–9.
14. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 19(Suppl 2):ii215–25.
15. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008 Jan 1;18(1):188–96.
16. Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. A benchmark study of ab initio gene prediction methods in diverse eukaryotic organisms. BMC Genomics. 2020;21:1–20.
17. Wang Y, Chen J-Q, Araki H, Jing Z, Jiang K, Shen J, et al. Genome-wide identification of NBS genes in japonica rice reveals significant expansion of divergent non-TIR NBS-LRR genes. Mol Gen Genomics. 2004 May 1;271(4):402–15.
18. Yuan Q, Ouyang S, Liu J, Suh B, Cheung F, Sultana R, et al. The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. Nucleic Acids Res. 2003 Jan 1;31(1):229–33.
19. Yuan Q, Ouyang S, Wang A, Zhu W, Maiti R, Lin H, et al. The institute for genomic research Osa1 Rice genome annotation database. Plant Physiol. 2005 May 1;138(1):18–26.

20. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, et al. The TIGR Rice genome annotation resource: improvements and new features. Nucleic Acids Res. 2007 Jan 3;35 (Database):D883–7.
21. McCouch SR, CGSNL (Committee on Gene Symbolization N and L Rice Genetics Cooperative). Gene Nomenclature System for Rice. Rice. 2008 Sep 1;1(1):72–84.
22. Ressayre A, Glémin S, Montalent P, Serre-Giardi L, Dillmann C, Joets J. Introns structure patterns of variation in nucleotide composition in Arabidopsis thaliana and Rice protein-coding genes. Genome Biol Evol. 2015 Oct 1;7(10):2913–28.
23. Ohyanagi H. The Rice annotation project database (RAP-DB): hub for Oryza sativa ssp. japonica genome information. Nucleic Acids Res. 2006 Jan 1;34(90001):D741–4.
24. Wang D, Xia Y, Li X, Hou L, Yu J. The rice genome knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology. Nucleic Acids Res. 2013 Jan 1;41(D1):D1199–205.
25. Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice annotation project database (RAP-DB): an integrative and interactive database for Rice genomics. Plant Cell Physiol. 2013 Feb 1;54(2):e6.
26. Atambayeva SA, Khailenko VA, Ivashchenko AT. Intron and exon length variation in Arabidopsis, rice, nematode, and human. Mol Biol. 2008 May 25;42(2):312.
27. Wang J, Wan X, Crossa J, Crouch J, Weng J, Zhai H, et al. QTL mapping of grain length in rice (Oryza sativa L.) using chromosome segment substitution lines. Genet Res. 2006 Oct;88 (2):93–104.
28. Wang B-B, Brendel V. Genomewide comparative analysis of alternative splicing in plants. PNAS. 2006 May 2;103(18):7175–80.
29. Thibaud-Nissen F, Ouyang S, Buell CR. Identification and characterization of pseudogenes in the rice gene complement. BMC Genomics. 2009 Jul 16;10(1):317.
30. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004 Jan 23;116 (2):281–97.
31. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005 Aug;436(7052):793–800.
32. Kurata N, Yamazaki Y. Oryzabase. An integrated biological and genome information database for Rice. Plant Physiol. 2006 Jan 1;140(1):12–7.
33. Zhang J, Li C, Wu C, Xiong L, Chen G, Zhang Q, et al. RMD: a rice mutant database for functional analysis of the rice genome. Nucleic Acids Res. 2006 Jan 1;34(suppl_1):D745–8.
34. Copetti D, Zhang J, El Baidouri M, Gao D, Wang J, Barghini E, et al. RiTE database: a resource database for genus-wide rice genomics and evolutionary biology. BMC Genomics. 2015 Jul 22;16(1):538.
35. Sato Y, Takehisa H, Kamatsuki K, Minami H, Namiki N, Ikawa H, et al. RiceXPro version 3.0: expanding the informatics resource for rice transcriptome. Nucleic Acids Res. 2013 Jan 1;41 (D1):D1206–13.
36. Sato Y, Namiki N, Takehisa H, Kamatsuki K, Minami H, Ikawa H, et al. RiceFREND: a platform for retrieving coexpressed gene networks in rice. Nucleic Acids Res. 2013 Jan 1;41 (D1):D1214–21.
37. Wang D, Xia Y, Li X, Hou L, Yu J. The rice genome knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology. Nucleic Acids Res. 2012 Nov 27;41(D1):D1199–205.

9 Phylogenetic Analysis

Manoj Kumar Gupta, Gayatri Gouda, S. Sabarinathan, Ravindra Donde, N. Rajesh, Pallabi Pati, Sushil Kumar Rathore, Lambodar Behera, and Ramakrishna Vadde

Abstract

In this chapter, the authors attempt to understand the underlying phylogeny principle and how researchers implement diverse methods to discover the appropriate phylogeny. Results obtained revealed that phylogenetic trees reflect evolutionary past as a canonical framework. Phylogenetic tree building step essentially comprises of five steps: (a) selecting molecular markers; (b) multiple sequence alignment; (c) determining the best evolutionary model; (d) determination of tree building method; and (e) assessment of tree reliability. Phylogenetic trees have various functional uses in different biological fields, such as conservation biology, epidemiology, forensics, cancer evolution, HIV transmission, gene expression prediction, protein structure prediction, and drug design. However, researchers face different challenges for generating a more accurate tree, like memory efficiency and implementation and optimization of the likelihood function. The authors believe, in the near future, the development of exciting new

M. K. Gupta · G. Gouda · R. Donde · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

N. Rajesh · R. Vadde (✉)
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_9

algorithms, which dramatically reduce the necessary amount of likeliness assessment, combined with enhanced knowledge of previously described high-performance machine problems in the group, is likely to detect more accurate phylogenetic tree that include 10,000–20,000 sequences. Additionally, it will also permit the tree inferences on medium-sized PC.

Keywords

Phylogeny · Molecular markers · Bootstrapping · Character-based · Distance-based · Jackknifing

Abbreviations

BI	Bayesian inference
cpDNA	Chloroplast DNA
dN	Non-synonymous
dS	Synonymous
GBS	Genotyping-by-sequencing
HTU	Hypothetical taxonomic units
ITS	Internal transcribed spacer
JC	Jukes and Cantor
LCA	Last common ancestor
LUCA	Last universal common ancestor
ML	Maximum-like
MSA	Multiple sequence alignment
OTUS	Operational taxonomic units
PCR	Polymerase chain reaction
UCES	Ultra-conserved elements

9.1 Introduction

Before the development of DNA sequence technology, phylogenetic trees have been predominantly employed to identify the connections between various organisms [1]. Ever since the development of sequencing technologies and use of sequence data for phylogenetics, there has been drastic improvement in our understanding of the tree of life, and tremendous advancement has been made in Darwin's dream of "very fairly true genealogical trees of each great kingdom of nature" (https://www.darwinproject.ac.uk/letter/DCP-LETT-2143.xml). Woese and Fox's [2] 1977 paper was an early illustration of what we term molecular phylogenetics today - comparison between macromolecular sequences to predict genealogical associations and thereby evolution. Crick [3] and more formally Zuckerkandl and Pauling put forward the idea of matching sequences with more closer relationships [4]. At this period, protocols developed by Sanger, who had won his first Nobel Prize for this discovery [5], and the identification of the amino acid sequence of insulin in the 1950s, rendered it more tractable to establish the sequences of different proteins. Subsequently, protein biochemists started to develop phylogenetic maps focused on amino acid sequences obtained from different species, primarily animals. Russell Doolittle drafted the evolution of vertebrate with blood-clotting fibrinopeptides [6]. Fitch and Margoliash employed the mitochondrial protein cytochrome C to link animals and certain fungi [7]. Subsequently, phylogenies have been used in almost every biological branch, including characterizing connections between paralogs in the population history [8], gene family [9], evolutionary and epidemiological dynamics of pathogens [10], language evolution [11], and the genealogical association between cancerous somatic cells [12]. In recent years, molecular phylogenetics has also become an utterly important method for comparing various genomes [13–15] and the reconstruction of ancestral genomes [16]. Thus, in this chapter, the authors attempt to understand phylogeny's basic concept and how researchers implement various approaches to reconstruct the best phylogeny tree. Subsequently, the authors also described in brief about problems researchers face while reconstructing phylogenetic tree and how, in the near future, we can overcome them.

9.2 Basic Concept

A fusion of different Greek terms produces the word "Phylogeny." In Greek, phylon means "tribe" or "clan" or "race," and genesis means "source" or "origin." The word may also be used to identify genes originating from a single ancestral gene [17]. Phylogenetic trees are composed of leaves, nodes, and roots similarly like botanical trees (Fig. 9.1) [18]. The leaves of a tree are often called node or tips or operational taxonomic units (**OTUs**) [18, 19]. The OTUs are real objects, such as plants, cultures, genes, or protein sequences - while internal nodes are hypothetical taxonomic units (HTUs). An HTU is a deducted unit that reflects the last common ancestor (LCA) of the nodes from this stage. Sister groups are descendants (taxa) separate from the same node, and a taxon outside of the clade is referred to as an outgroup [19]. A branch that reflects the continuation of a lineage over time will subjugate one or more leaves. Branches link at nodes with other branches reflects the last common predecessors of species at the descendant lines' tips. An external branch is a branch that links a tip to a node, while a branch joining two nodes is considered an internal branch [18]. A node depicts a spot where the ancestral line (a branch below the node) separates and gives rise to a few or more descendant lines (a branch above the node).

Branching within an evolutionary tree is often termed "lineage splitting" or "cladogenesis." After a group breaks into two, evolution occurs in these separately created lineages individually. The lineage sequence divides into a tree creates its form or "topology." Tree topology tells us how the lines are branched over time and give rise to their tips. "Clades" are tree groupings with a node, and all lineages come

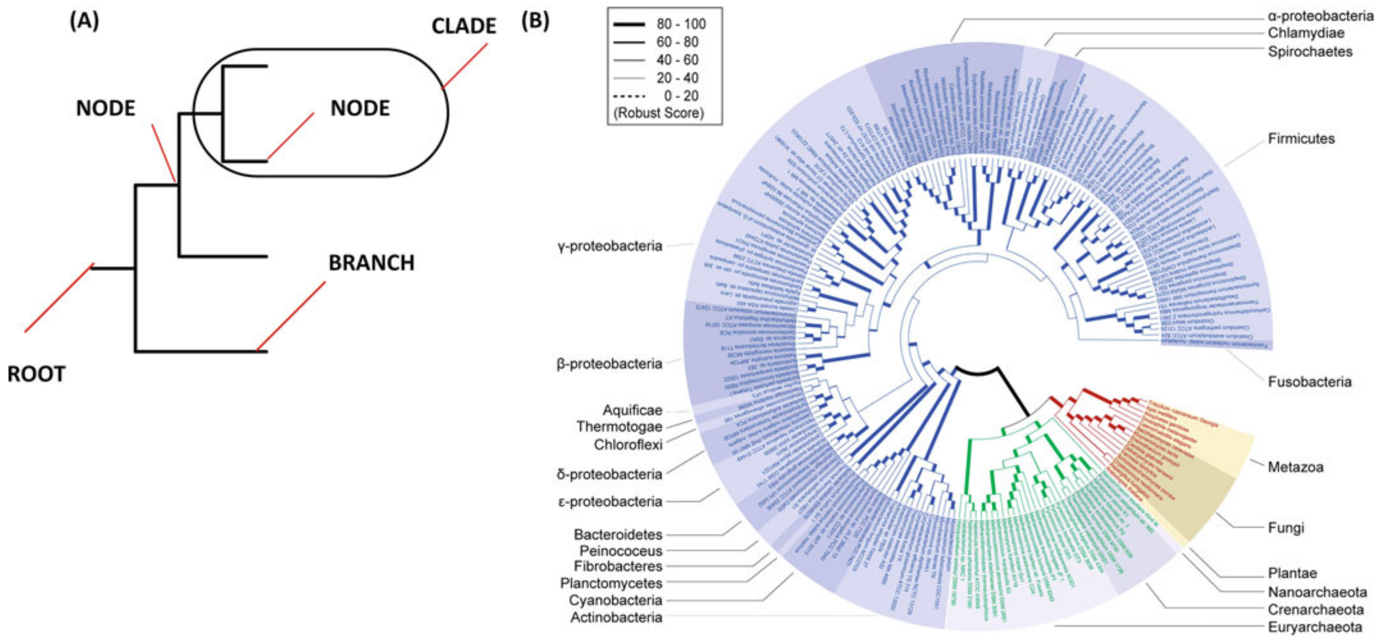


Fig. 9.1 (**a**) Representation of simple phylogenetic tree. (**b**) Phylogenetic tree depicting the tree of life based on genome context networks of 195 representative species (Adapted from [20])

from that node. The compilation of all tips is regarded since "monophyletic," as it contains all the daughters of an ancestral lineage. A topology of a tree may now be more precisely described as the set of claddings within a tree [18].

Phylogenetic trees may be unscaled or scaled. In a scaled tree, the branch's length refers to the sum of evolutionary divergence happening along that branch (e.g., number of the nucleotide substitutions). For an unscaled tree, the length of the branch is not proportionate to the sum of evolutionary variations; however, the exact quantity is typically seen elsewhere in the branch [19]. Phylogenetic trees may be unrooted or rooted. A rooted tree has a node (root) which diverges the rest of the tree. This source is also considered the "last universal common ancestor" (LUCA), from which the other taxonomical classes have come down and diverged over time. LUCA and LCA are described in molecular phylogenetics by DNA or protein sequences. The creation of the rooted tree is optimal, but most phylogenetic reconstruction algorithms generate unrooted trees [19].

Phylogenetic trees may also be phylogram, cladogram, and dendrogram (Fig. 9.2). A phylogram is a scaled phylogenetic tree in which the branch's lengths correspond to the degree of evolutionary differentiation, e.g., the number of nucleotide substitutions between related branch points will decide a branch length. A cladogram is a branched hierarchical tree which shows the connections among clades; cladograms are not scaled. The term dendrogram implies a hierarchical organization of the clusters, where related artifacts are categorized into clusters (based on such specified criteria), and thus the connections between the clusters can be seen in a dendrogram. Dendrograms are also used to demonstrate branching based on the clustering of genes or proteins in computational molecular biology [19].

9.3 Phylogenetic Tree Construction Methods

Phylogenetic tree building essentially takes five steps: (a) selecting molecular markers; (b) multiple sequence alignment (MSA); (c) determining the best evolutionary model; (d) determination of tree building method; and (e) assessment of tree reliability [21–23]. List of softwares as well as tools that are widely used for constructing pjjhylogenetic tree is depicted in Table 9.1.

9.3.1 Selecting Molecular Markers

One may use nucleotide or protein sequence data to create molecular phylogenetic trees. The choice of molecular markers is critical since it can make a big difference in the correct tree generation. The choice to use nucleotide or protein sequences depends on the sequence properties and the research's intent. Nucleotide sequences that are more quickly changing than proteins may be used to research very closely associated animals. However, conserved protein sequences make sense rather than utilizing nucleotide sequences if the phylogenetic relationships to be established are at the deepest stage, for example, between bacteria and eukaryotes. Protein

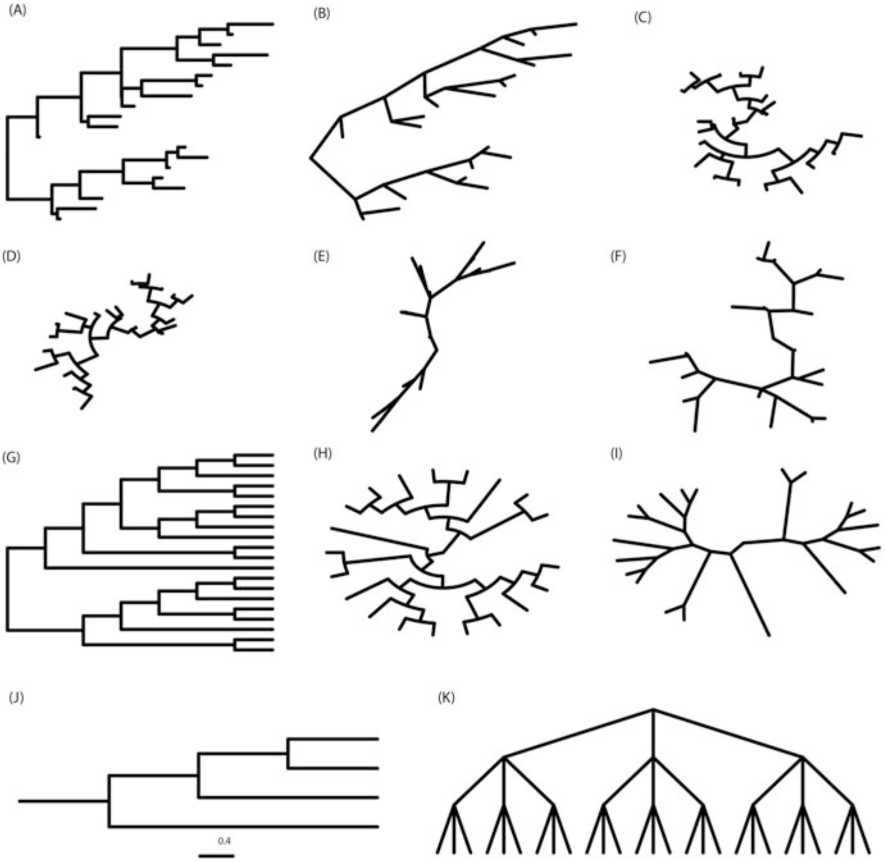


Fig. 9.2 Different types of phylogenetic trees. Phylogram: (**a**) rectangular layout, (**b**) slanted layout, (**c**) circular layout, and (**d**) fan layout. Unrooted: (**e**) equal-angle method and (**f**) daylight method. Cladogram: (**g**) rectangular layout, (**h**) circular layout, and (**i**) unrooted layout. (**j**) Scaled layout and (**k**) Dendrogram layout

sequences are also preferred to nucleotide sequences since the protein sequences are most likely to be preserved owing to the genetic code degeneracy for which 61 codons encode 20 amino acids because a few codon switch cannot modify amino acid. Protein sequences will therefore stay the same as corresponding DNA sequences, even if they have more space for variation, particularly at the third codon site. The major genetic disparity between the three-nucleotide sites often violates one of the assumptions of the tree design. In comparison, even with divergent sequences, the protein sequences do not really suffer from this issue [21]. Additionally, DNA sequences are so much more biased than protein sequences owing to selective usage of a codon in multiple species. In this situation, various codons are used in various frequencies with the same amino acid, resulting in sequence inconsistencies not due to evolution. Moreover, the mitochondria genetic code differs from the normal

Table 9.1 List of software or tools used for phylogeny reconstruction (Adapted from https://en.wikipedia.org/wiki/List_of_phylogenetics_software)

Software/tools	Description	Methods	Reference
AncesTree	Clonal tree reconstruction employing cancerous data	ML	[24]
Ape	R pacakges for phylogenetics and evolution analysis	Employs wide range of phylogenetics functions	[25]
Armadillo workflow platform	General bioinformatic and phylogenetic analysis	Phylogenetic trees reconstruction employing distance, ML, MP, BI, and related workflows.	[26]
BAli-Phy	Simultaneous BI of alignment as well as phylogeny	BI, alignment as well as tree scan.	[27]
BayesPhylogenies	BI of trees employing Markov chain Monte Carlo methods	BI, multiple models, mixture model (auto-partitioning)	http://www.evolution.rdg.ac.uk/BayesPhy.html
BayesTraits	Analyzes trait evolution among groups of species for which a phylogeny or sample of phylogenies is present	BI	http://www.evolution.rdg.ac.uk/BayesTraitsV3.0.2/BayesTraitsV3.0.2.html
BEAST	Bayesian evolutionary analysis sampling trees	BI, relaxed molecular clock, demographic history	[28]
BioNumerics	Universal platform for the management, storage as well as analysis of all forms of biological data, including tree and network inference of sequence data.	NJ, MP, UPGMA, ML, distance matrix methods. Estimation of the consistency of trees/branches employing bootstrapping, error resampling, or permutation resampling	https://www.applied-maths.com/bionumerics
Canopy	Evaluating intratumor heterogeneity as well tracking spatial and longitudinal clonal evolutionary history by next-generation sequencing	ML, Markov chain Monte Carlo (MCMC) methods	[29]
Dendroscope	Tool for visualizing rooted trees and calculating rooted networks	Rooted trees, tanglegrams, consensus networks, hybridization networks	[30]
EzEditor	EzEditor is a java-based sequence alignment	Neighbor joining	[31]

(continued)

Table 9.1 (continued)

Software/tools	Description	Methods	Reference
	editor for rRNA and protein-coding genes. It allows manipulation of both DNA and protein sequence alignments for phylogenetic analysis.		
FastTree 2	Fast phylogenetic inference for alignments having 100 to 1000 sequences	Approximate ML	[32]
Geneious	Have numerous genomes as well as proteome research tools	NJ, UPGMA, GARLi plugin, MrBayes plugin, RAxML plugin, PHYML plugin, FastTree plugin, PAUP* plugin	https://www.geneious.com/
HyPhy	Hypothesis testing employing phylogenies	ML, clustering techniques, NJ, distance matrices	https://www.hyphy.org/
IQ-TREE	An effective phylogenomic software through ML, as successor of TREE-PUZZLE and IQPNNI	ML, model selection, AIC, partitioning scheme finding, BIC, ultrafast bootstrapping, likelihood mapping branch tests, tree topology tests	[33]
jModelTest 2	A high-performance computational tool for carrying out statistical selection of best-fit models of nucleotide substitution	AIC, BIC, DT, dLTR hLTR, ML	https://github.com/ddarriba/jmodeltest2
MEGA	Molecular evolutionary genetics analysis	Distance, parsimony and maximum composite likelihood methods	https://www.megasoftware.net/
Mesquite	Software for evolutionary biology, developed for helping biologists to analyze comparative data of various organisms. It can also be employed for building time trees having a geological timescale, with few optional functionalities.	MP, distance matrix, ML	https://www.mesquiteproject.org/
MetaPIGA2	ML phylogeny inference multicore	ML, stochastic heuristics (genetic	https://www.metapiga.org/

(continued)

Table 9.1 (continued)

Software/tools	Description	Methods	Reference
	program for DNA and protein sequences, and morphological data. It also implements tree visualization tools, ancestral sequences, and automated selection of best substitution model and parameters.	algorithm, simulated annealing, metapopulation genetic algorithm, etc.), ancestral state reconstruction, discrete gamma rate heterogeneity, model testing.	
Modelgenerator	Model selection (nucleotide or protein)	ML	http://mcinerneylab.com/software/modelgenerator/
MOLPHY	Molecular phylogenetics (nucleotide or protein)	ML	https://sbgrid.org/software/titles/molphy
MrBayes	Posterior probability estimation	BI	https://nbisweden.github.io/MrBayes/index.html
PAML	Phylogenetic analysis by ML	ML and BI	[34]
ParaPhylo	Estimation of gene as well as species trees based on event-relations (paralogy and orthology)	Cograph-editing and triple-inference	[35]
PartitionFinder	Combined selection of models of molecular evolution and partitioning schemes for DNA and protein alignments.	ML, AIC, BIC	[36]
PASTIS	R package for phylogenetic assembly	R, two-stage BI employing MrBayes 3.2	[37]
PAUP*	Phylogenetic analysis employing parsimony (*and other methods)	MP, ML, distance matrix	http://paup.phylosolutions.com/
Phangorn	Phylogenetic analysis in R	ML, MP, distance matrix, bootstrap, phylogentic networks, bootstrap, model selection, SH-test, SOWH-test	[38]
Phyclust	Phylogenetic clustering	ML of finite mixture modes	https://snoweye.github.io/phyclust/

(continued)

Table 9.1 (continued)

Software/tools	Description	Methods	Reference
PHYLIP	Phylogenetic inference package	MP, distance matrix, ML	https://evolution.genetics.washington.edu/phylip.html
PhyloWGS	Reconstructing sub clonal composition and evolution from whole-genome sequencing of tumors	MCMC	[39]
PhyML	Fast as well as accurate estimation of phylogenies employing ML	ML	http://www.atgc-montpellier.fr/phyml/
Phyx	Unix/GNU/Linux command line phylogenetic tools	Explore, manipulate, analyze, and simulate phylogenetic objects (alignments, trees, and MCMC logs)	[40]
POY	Supports various data form and can perform alignment as well as phylogeny inference employing numerous heuristic algorithms	MP, ML, continuous characters, chromosome rearrangement, alignment, discreet characters	[41]
ProtTest 3	A high-performance computing program for detecting the model of protein evolution that best fits a given aligned sequences set	ML, AIC, BIC, DT	[42]
PyCogent	Software library for genomic biology	Simulating sequences, alignment, controlling third-party applications, workflows, querying databases, generating graphics and phylogenetic trees	[43]
RAxML-HPC	Randomized Axelerated ML for high performance computing (nucleotides and aminoacids)	ML, simple MP	[44]
RAxML-NG	Randomized Axelerated ML for high performance computing (nucleotides as well asaminoacids) next generation	ML, simple MP	[45]

(continued)

Table 9.1 (continued)

Software/tools	Description	Methods	Reference
SEMPHY	Tree reconstruction employing the combined strengths of maximum-likelihood (accuracy) and NJ (speed). SEMPHY has become outdated.	A hybrid maximum-likelihood/NJ method	http://compbio.cs.huji.ac.il/semphy/
TreeGen	Tree construction given precomputed distance data	Distance matrix	[46]
T-REX (webserver)	Tree inference as well as visualization, MSA, horizontal gene transfer detection	Distance (neighbor joining), parsimony and ML (RAxML PhyML) tree inference, MAFFT, MUSCLE, and ClustalW sequence alignments and related applications	[47]
UGENE	Fast as well as free multiplatform tree editor	Based on Phylip 3.6 package algorithms	http://ugene.net/

genetic code. Therefore, while analyzing mitochondrial data, the DNA sequences are converted into protein sequences by comparing the mitochondrial codon table. Moreover, protein sequences make a more sensitive alignment than DNA sequences, as the former has 20 and later has 4 characters. Two randomly associated DNA sequences were found to be capable of generating up to 50% sequence identification if gaps are tolerated compared with 10% for protein sequences. For moderately divergent sequences, the usage of DNA sequences is virtually difficult to achieve proper alignment. In particular, when gaps are formed to optimize alignment scores, protein-coding DNA sequences almost always trigger frameshift errors, rendering the alignment biologically irrelevant. When it comes to alignment as well as phylogenetic study, protein sequences simply have a better signal-to-noise ratio. However, in certain instances, protein-dependent phylogeny might be more important than phylogeny dependent on DNA [21].

DNA sequences may also be very informative in some situations, especially in coding region, due to their higher rate of evolution. DNA sequences often represent synonymous and non-synonymous substitutions to show proof of favorable or harmful selection events. A distinction must be created between synonymous substitutes and non-synonymous substitutions in order to consider positive or negative selection. Synonymous substitutions are nucleotide variations in the coding sequence which do not alter the encoding protein's amino acid sequence. Non-synonymous substitutions are nucleotide modifications that lead to amino acid sequence modifications. Comparing the two forms of replacement rates helps to explain an evolutionary cascade pattern. For example, if the non-synonymous substitution rate is considerably higher than the synonymous substitution rate, this means that certain parts of the protein experience active mutations that may lead to the creation of new functions. This is defined as a positive or adaptive selection. However, where the synonymous substitution rate increases, the non-synonymous substitution rate still induces neutral modifications at the amino acid level, which means that the protein sequence is critical enough to avoid modifications at the amino acid sequence level. The pattern, in this case, is assumed to be negative or purifying [21].

Over the last few decades, with molecular methods and tools developing, plant phylogenetics has also started to use more nuclear and chloroplast loci for detecting the best plausible relationships between plants. Due to their clear, stable structure as well as ease of primer design and amplification, the chloroplast loci were commonly used [48]. The single-parent heritage of chloroplast loci is disadvantaged by not generally controlling the population tree and by maintaining their evolutionary pace [48–51]. Nuclear genomes give several separate and unrelated positions that evolve at different rates. They usually grow more rapidly than chloroplast loci [52, 53], but because of the scarcity of sufficient genomic data for most taxa, they have a drawback. Further problems occur from gene duplications or gene destruction, which also shows paralogy when a nuclear locus is sequenced in another taxon [54]. In addition, the sluggish pace of DNA evolution in chloroplast DNA (cpDNA) and most nuclear protein-coding loci indicate that access to only a few loci also does not overcome phylogenies at the species stage. Consequently, several plant phylogenies are focused on many cpDNA loci with only the internal transcribed spacer (ITS) regions reflecting the two-parent nuclear loci. The detection and sequencing of several nuclear loci, which may in conjunction with cpDNA, be useful in using differing rates of evolution as well as possibly different evolutionary histories, is needed to obtain well-resolved phylogenies for species-rich plants. The author believes that exons are more rational when opposed to intronic or noncoding regions as aimed toward strongly differentiated taxa since they have fewer differences and can be more securely matched than intronic and noncoding regions with higher substitution rates as well as excessive length variations [55].

However, it remains challenging to choose the best loci for such work and to find alternatives, which advise on various levels of the evolutionary past, entail resolving a balance among facts as well as harmony. Locus is often essential to be orthological, i.e., shared because of common ancestors and not paralogs from duplication events [54, 56]. Nuclear single or low copy number genes are also beneficial since the probability of the existence of paralogs is decreased considerably [57]. A series of approaches were taken to obtain appropriate loci, mostly in the context of reduced methods of representation [58]. They also use selective enzymes to target the genome subset, which uses the genotyping-by-sequencing (GBS) and have the benefit of not having genomic reference data but usually do not work for remotely associated taxas since the restriction sites are less conserved [59].

Further precise methods of reduced representation include target enrichment, including polymerase chain reaction (PCR) produced probes [60], the creation of markers targeting 3’UTR, or exon capture based on transcriptome [61]. More targeted approaches include utilizing current genomic tools to classify candidate genes – Bragg and the team [58] used lizard Anolis's exons to classify the same homologs in more distant associated taxa transcriptomes using the stronger reciprocal BLAST technique, while Li et al. [62] used a comparative approach to compare exons across fish genomes to supposedly filter one-copied orthologous exons. These approaches could be useful for phylogenetic estimates in more distinct taxa [58, 63].

Similarly, targeting ultra-conserved elements (UCEs) by means of hybrid enrichment can be helpful in solving mid to deep phylogenies [64], but these are not always capable of capturing adequate signals to overcome shallower nodes [65]. Marker sets used for anchored phylogenomic may be useful at different phylogenetic stages, as both retained loci and flanked regions comprising variants can be targeted but have a comparatively long period of development [64]. Likewise, exon capture is widely used in multilevel phylogenetics [64] but needs prior genomic tools, such as a reference genome(s). The discovery of numbers of insightful loci with minimal genomic tools remains a problem. The latest methodologies for finding conserved plant phylogenetic inferences include hybrid approaches utilizing transcriptome and genome skim details in Oxalidaceae [66, 67], or transcriptome and the whole genome sequence records in Ericaceae [68].

9.3.2 Multiple Sequence Alignment

To date, sequence data is now the most widely used form of data for phylogenetic reconstruction. Even before sequence may be used for the reconstruction of phylogeny, it must be aligned, and its accuracy has been shown to influence the consistency of the presumed phylogeny [69]. The most common and widely employed MSA method is progressive alignment. Basically, it operates by first aligning the two nearest sequences and one by one, before any sequence is matched. ClustalW [70, 71] is one of the best-known progressive sequence alignment tools. ClustalW's primary concern is that preliminary pair-wise alignments are resolved, and early glitches cannot be changed later, although such alignments interfere with later sequences [72]. T-Coffee is another common technique for sequencing, which can be used as an advanced process variant [73]. Detail description of MSA has been described in Chap. 7 earlier. Yet, detailed MSA reconstruction is a difficult measurement job because of (1) the evolutionary process's stochastic features, (2) modern bioinformatics approaches have computational restrictions, and (3) a lack of specific evolutionary models representing sequences of processes [74]. Earlier studies have reported that large-scale assessment of alignment methods utilizing simulated as well as empirical datasets yield MSAs that are still subject to significant errors. Such alignment errors were found to influence phylogeny inference precision [75]. Earlier studies indicated that filtering unstable MSA regions before phylogeny inference could increase the accuracy of tree reconstruction. Whatever the filtering technique used, eliminating alignment errors is invariably followed by eliminating insightful phylogenetic signal. This precarious compromise among the amount of noise and the amount of signal excluded from MSA renders it impossible to evaluate the benefits of filtering unreliable alignment regions [76].

Alternatively to screening out unreliable columns of alignment, some experiments recommended reducing ambiguity by seeking MSA consensus to collect alternative MSAs or choose the most reliable MSA [77]. Previously, various methods of alignment, or separate parameter choices of a specific method, produce MSAs that vary considerably from each other [78, 79]. This instability in resulting MSAs indicates that MSA consensus approaches, comparable to the column-filtering approaches mentioned above, minimize noise at the expense of discarding significant phylogenetic signal [74].

The weight of the MSA columns based on their reliability is another solution to filtering faulty columns. Early works demonstrate that introducing a different coding scheme to ambiguous columns contributes to the more precise reconstruction of the trees [80]. A similar suggestion of utilizing an unreliable coding scheme for the maximum-like (ML) system was also proposed [81]. In recent years, ML-tree reconstruction has been found to be more reliable as any column in the MSA input has been weighed by its redundancy [76]. Although MSA's possible advantage directly in tree reconstruction is seen by experiments, they only use just one MSA, which can be used to solve the alignment problem and hence disregard most of the phylogenetic signal found in the next segment of the so-called "alternative MSA." Specifically, these methods assign a null weight to entirely MSA columns in alternate MSAs, only in the weight of MSA and not in the original MSA, irrespective of the frequency of these columns [76].

In the maximum parsimony (MP) paradigm, the mixture of alternative MSAs was previously suggested, created by changing the gap penalty parameter. The technique known as Elision uses the broad, concatenated MSA to inference phylogeny [82]. This method is analogous to weighing MSA columns by the number of iterations before tree reconstruction processes in the concatenated MSA. Researchers also developed a GUIDANCE tool that uses those weights (the number of times a column occurs in a number of alternate MSAs) to calculate the reliability of that column. We found that filtering unreliable positions on the basis of this weighting would boost the positive conclusion in selection [83]. GUIDANCE2 is an tool focused on the generation of alternative MSAs by variation of the reference book for MSA generation, the distance opening penalty, and another similar high-ranking MSAs [84].

Basically, merging MSAs is like having a special MSA, in which each column 's weight is the amount of times it occurs between alternate MSAs. This technique is very similar to the reliability estimate of GUIDANCE. However, although only positions given by the alignment method in the best MSA is weighted in GUIDANCE (the "base MSA"), the recent research by Ashkenazy and the team also weights MSA columns not present in the base MSA, that is, all original MSAs and alternate MSA columns are weighted equally. This method can be broken down into three phases: (1) alternative MSAs are generated; (2) all feasible single alignment columns are extracted from the "MSA Base" and alternative MSAs; (3) use a weighted MSA obtained called SuperMSA as well as the column weights to recreate a phylogenetic tree, which is centered on its frequency between the alternate MSAs and base MSA [74].

9.3.3 Determining the Best Evolutionary Model

The number of replacements in an alignment is essentially a calculation of the divergence between two sequences. The proportion of substitutes specifies the distance between the two sequences found. However, the amount of substitutions observed does not reflect true evolutionary events. When a mutation as A substituted with C is detected, the nucleotide might have required some intermediate measures to becoming C, such as A alternatives T →T → G → T → C. A back mutation may even have happened where a mutant nucleotide returned to the initial nucleotide. This indicates that mutations such as G → C →G could have arisen where the same nucleotide is detected. In addition, an identical nucleotide in the alignment could be found in simultaneous mutations as, e.g., the two sequences mutate into T. The calculation of true evolutionary distances between sequences is obscured by the various substitutions and convergence at individual loci. This phenomenon is known as homoplasy, which may lead to incorrect trees if not corrected. In order to correct homoplasy conditions, statistical modeling is needed to evaluate the true evolutionary distances among sequences. Statistical methods used to fix homoplasy are termed substitution or evolutionary models. Based on the sequence type, to date, numerous substitution models of DNA, amino acid and codon, sequences have been developed [85].

9.3.3.1 DNA Substitution Models

Jukes and Cantor (JC) [86] described the first and easiest model to imitate the DNA substitution process. For both nucleotides substitution and nucleotide frequencies, this model assumes one substitution rate. Changes between bases with an identical chemical composition (transitions) are, however, more frequent than changes among bases with various chemical structures (transversions). Genetic coding also allows further transformations than transversions without the substitution of amino acid [87]. Motivated by this information, Kimura [88] proposed a two-parameter model (K80) in which change rates vary among transitions and transversions. Similarly, Felsenstein [89] (F81) generalized the JC model to incorporate multiple nucleotide frequencies, which may often occur as a result of natural selection as well as nucleotide physicochemical properties. A range of models was developed subsequently by adding extensions to the initial models, e.g., SYM [90] and HKY [91]. According to this pattern, the general time-reversible model (GTR) (Tavaré 1986) integrates variable rates for each modification as well as different nucleotide frequencies. Additionally, few other models also integrate a variance rate across sites (+ G) [92] and/or proportion of invariable sites (+ I) [93].

Stationary, reversible, and homogeneous DNA substitution schemes derived from all feasible combinations have already been established and are being applied in several phylogenetic programs [94, 95]. However, considering the high number of DNA substitution models available, the GTR + G + I model usually suits better with real data than the other (simpler) alternative models [96]. Importantly, the GTR model introduces unnecessary mathematical properties [i.e., two GTR matrice multiplications do not return an additional GTR matrix [96] to be used as a phylogenetic [96, 97]]. In particular, the repeated selection of the most complicated substitution model indicates that the fit of real data could be improvised by more complex models [98].

9.3.3.2 Codon Substitution Models

Sites inside codons evolve at varying speeds and should, therefore, not be considered similarly. In reality, considering molecular evolution at the codon stage, we should take more plausible evolution trends into account for each codon location. Evolutionary models focused on the design of codon evolution are actually more rigorous than those drawn from empirical amino acid models [85]. Interestingly, codon-level mutations are known to be synonymous (silent) as well as non-synonymous (amino-acid replacement), which provides the molecular selective pressure (molecular adaptation) measure. The first models considered non-synonymous (dN) and synonymous (dS) substitution rates [99], or the dN/dS ratio [100]. dN/dS $>$ 1 means that protein-coding gene substitutions have been applied to those that have altered the states of amino acids and imply diversifying (positive) range [101, 102]. By comparison, dN/dS $<$ 1 and dN/dS $=$ 1 may be viewed as a purifying (negative) and balanced selection [101, 102].

Additional codon versions have also been suggested to best match certain codon datasets [100]. The codon model for GY94 was expanded to include different nucleotide models, e.g., GTR [103], dN/dS variation across different branches [104] and across sites [103, 105]. Additional codon models provide details on the physical and chemical properties of the amino acids encoded [106]. Besides, a variety of empirical codon models - focused on comprehensive databases [107] as well as codon models that take codon bias into account [108], or impact of various GC content [109] were also suggested. Codon models thus help us to conduct correct evolutionary research and explore molecular adaptation signatures. However, the broad matrices of these models' exchangeability are troublesome in technical terms (61X 61; remember that stop codons are omitted). As a result, large quantities of data are expected to produce statistically successful analytical codon matrices and a large computational burden. Consequently, vast quantities of data are required to produce statistically well-supported analytical codon matrices, which in turn demands high computational facilities. Fortunately, codon-based optimization research now produces new evolutionary methods for modeling and assessing codon evolution, but further research is required in this regard (e.g., integrating parallel likelihood matrix computing) [85].

9.3.3.3 Amino Acid Substitution Models

Substitute models of amino acid evolution are intended to imitate the evolution of protein data, which are crucial for the test of a variety of hypotheses, for instance, phylogenetic tree reconstructions [110], rate of protein evolution [111], or selection in novel proteins [112]. Amino acid substitution models can be categorized into the following two major groups: (i) empirical model-based on large protein databases, e.g., Dayhoff [113], CpRev [114], and DayhoffDCMUT [115]. (ii) Parametric models are based on protein evolutionary parameters [116]. An empirical model for amino acid substitution involves a matrix of 20 X 20 exchangeability ratios and 20 amino acid frequencies. Compared with other amino acid models, including those focused on structural limitations, this simplification contributes to benefits and pits. Empirical models may be integrated by assuming site-independence (all the sites evolve under the same model) into the most widely used probability functions introduced in standard phylogenetic software. The heterogeneous evolution can also be modeled by identifying numerous empirical models for the different protein sequence partitions (i.e., sites or domains) [117]. However, every real data collection could not be accurately interpreted by any of the analytical models currently available. For example, Keane et al. (2006) showed that the strongest empirical model was originally extracted from the retroviral protein Pol for two large proteobacterium and archaea protein datasets.

However, to offer an alternative to empirical models, protein folding constraints were considered to produce parametric amino acid substitution models that resulted in substantial changes (in terms of empirical models) while fitting actual data [118]. Nevertheless, these models are not designed for evolutionary studies of protein data as the commonly employed evolutionary processes are likely to conduct tasks that do not resolve site dependency. The ongoing research of structurally restricted substitution models is, therefore, designing location-specific matrices that can be inserted into common phylogenetic frames [119]. While current research into amino acid substitution models offers more complex models, these models are not typically used by evolutionary biologists as these models are frequently overlooked in evolutionary frameworks. An alternate technique may be an Approximate Bayesian computation (ABC) method for considering complicated substitution models in model selection and also in evolutionary analysis [85].

9.3.4 Determination of Tree Building Method

Two key types of tree-building approaches currently operate, each with advantages and limitations. The first type is focused on distinct characters of molecular sequences of taxa. The fundamental principle is that characters in corresponding positions are homologous between the sequences. Therefore, this dataset will map the character states of the shared ancestor. Another hypothesis here is that every character evolves separately and is therefore viewed as an individual unit of evolution. The second group of phylogenetic approaches is focused on the difference between pairs of sequences, which is determined according to the orientation of the sequences. Methods based on distance presume that all sequences are homologous, and the branches of the tree are additive, so the distance between the two taxa equals the sum of all the related branches [21].

9.3.4.1 Distance-Based Methods

One job with many applications is to build a tree with a matrix of distances whose relative locations in the leaves somehow represent the distances [120]. This is valuable for both evolutionary genetics, where the tree reflects the evolution of a group of organisms, groups or genes, and for cluster studies, where the tree exposes the correlations in an item array. In evolutionary biology, molecular sequence distances are usually calculated utilizing probabilistic sequence evolution models [121, 122]. The resulting distances can be assumed to be essentially additive. In other terms, a (phylogenetic) tree with branch lengths can be formed to approximately equivalent path lengths (sequences) between the leaves. Finding this tree is the aim of many common tree reconstruction methods based on distance. The key benefit of distance-based techniques is their implementation speed, which is greater than that of other (potentially more accurate) approaches. As a result, distance methods are employed if computational utility is critical: to rebuild very large trees or, as in the case of bootstrapping, large sets of trees or also to create initial phylogenies for quest heuristics centered on sophisticated approaches. Indeed, a general pattern in bioinformatics as well as computational biology is increasing demand for methods that can cope with large DNA sequence datasets. Distance-based approaches are a potential solution to this issue, not just for phylogenetic inference but also for related tasks such as sequence recognition (e.g., in metagenomics) and gene orthology inference (e.g., in functional genomics). A evidence of this demand is the continued popularity of neighbor joining (NJ) [123], which remains the most frequently cited phylogenetics algorithm. The benefit of distance methods is offset by a lower precision than methods which take full-sequence knowledge into account [121], like maximum likelihood (ML). Although recent findings indicate that some distance methods are basically as good as ML, in a measure of statistical efficiency [124]. A drawback of distance-based methods resides in the fact that it can be difficult to conclude such parameters common to all sequences if the distances are determined through similarities between them only in parallel series [125]. These parameters can also be calculated by ML on small sequence samples and then used distance approaches for the whole sample.

9.3.4.2 Character-Based Methods

Character-based approaches, more specifically, pick the ideal tree using sequence data. For example, parsimony techniques choose the tree to describe the observed data in terms of the least number of potential substitutions. Thus, maximal parsimony typically uses a basic sequence substitution model (all modifications are similarly probable). Although its simplicity, or maybe because of it, the MP becomes less common. It was demonstrated that it is vulnerable to the recovery of wrong trees, particularly if more divergent sequences are reused [126]. Methods of maximum probability strive to find the most possible single tree (according to the substitution model used). Therefore, the Bayesians approach sample trees with a frequency proportional to the probability of each tree and eventually create a consensus among the best trees with the most frequently observed nodes. Bayesian and probabilistic approaches have something in common with each other in the way that both approaches can integrate complicated models of character evolution like distance methods. However, the parameters relevant to the model may be optimized to individual tree topologies in probabilistic approaches. Therefore, probabilistic approaches aim to restore the topology of tree, branch size, and model parameters that would produce the observed data most possibly. The big benefit of these approaches is that they are focused on solid statistics and are subject to model and topology match statistical checking (relative merit of alternate models and tree topologies may be testing) [127]. Their downside is that they appear to be intensive in a computational application. Several shortcuts were planned to accelerate their output with a limited loss of performance. For example, it is popular to create a starter tree (e.g., a distance algorithm) and use this to approximate model parameters instead of attempting to maximize all model parameters and topology of the tree at once. A check for probabilities and parameters is then carried out on this tree before further searching and optimizing the parameters. An excellent tutorial about using probability programs such as PAUP* is available online (https://paup.phylosolutions.com/), although several comprehensive analyses of the methodologies were published earlier [127–129].

9.3.5 Assessment of Tree Reliability

The tree is produced based on the tree building method, and the input data is known as the inferred tree. This tree may not have to be the true tree with the phylogenetic data. The accuracy of the phylogenetic tree or section of the tree must therefore be checked. The sample size grows for approaches such as MP, ML, and minimum evolution. Under these circumstances, it should be confirmed if the tree is significant/better than another tree. The quality of the phylogenetic tree or section of the tree is checked by the sampling techniques, while statistical analyses validate the substantial discrepancy between the tree and the other [130].

9.3.5.1 Sampling Methods

Sampling strategies such as bootstrapping, jackknifing, and Bayesian simulation measure the reliability of the phenomenon tree or tree section. Bootstrapping is random sampling by data replacement (distance or sequence: nucleotide or protein) that addresses whether there have been any sampling errors for necessary analysis. Bootstraps repeatedly sample the data for phylogenetic tree in molecular phylogeny and provide us with an ability to test the original tree's power. If the data resampling yields separate trees relative to the original tree, the tree's topology is dependent on data with poor phylogenetic signals. If data resampling yields a tree close to a tree, the tree's topology is dependent on data with adequate phylogenetic signals. Thereby bootstrapping (resampling) gives insights into the trust of tree topology. In phylogenetic research, two forms of bootstrapping are used: parametric or nonparametric bootstrapping. If the data are interrupted by random sampling that produces new datasets, the bootstrapping is nonparametric. If the information is interrupted to construct a new dataset, it is parametrical bootstrapping. The other bootstrapping forms include case resampling, smooth bootstrap, Bayesian bootstrap, Gaussian regression phase bootstrap, resampling residuals, wild bootstrap, and block bootstrap (cluster data: block bootstrap; time series: simple block bootstrap, and time series: moving block bootstrap,). All these trees are summarized in a consensus tree that is based on a majority rule. The most supported branching patterns are labeled with bootstrap values at each node. Bootstrap, therefore, gives a measure of trust to approximate the degree of tree topology [130].

Jackknifing is another re-examination technique where half a dataset is arbitrarily removed, and semi-original datasets are produced. Initially, a phylogenetic tree is formed for the original dataset, then a phylogenetic tree with each new dataset generated by jackknifing is formed using the same approach as the original. Sampling creates separate trees when phylogenetic signals are small, but if phylogenetic signals are high, the study develops a similar tree. Thus, jackknifing may also be used to determine the trust of the topology of the tree (Challa and Neelapu, 2019). Markov chain Monte Carlo (MCMC)-based Bayesian approach re-examines thousands and millions of steps or iterations. The data sample sets are used for the reconstruction of phylogenetic trees close to the presumed original tree. At each intersection of the best Bayesian tree, the posterior probabilities calculate the trust of tree topology [130].

9.3.5.2 Statistical Methods

Statistical assessments, such as the Kishino–Hasegawa (KH) Test and Shimodaira–Hasegawa (SH) Test, confirm the significant difference between different trees. KH test contrasts two topologies of the trees in order to discern one tree from another [131]. Although KH tests may distinguish trees generated by methods such as distance, likelihood, and parsimony, Kishino–Hasegawa explicitly created this test for parsimonious trees. The KH test is a paired Student t-test on the basis of a null statement that “two opposing tree topologies are not substantially different.” The typical variation in branch lengths between two trees is measured at each formation site. The t-value obtained is then compared to the t-distribution values to support or deny the null hypothesis at certain essential thresholds (P-value <0.05) [130]. Shimodaira–Hasegawa have developed an ML-dependent statistical test dependent on the χ^2 test to estimate the strength of two opposing trees [132]. The log probability rates for tree A and tree B equal to lnLA and lnLB are first obtained for the two opposing trees. The two scores log ratio is then derived by $d = 2\ (\ln LA - \ln LB) = 2 \ln\ (LA / LB)$ and is then used to evaluate the χ^2 distribution from a table. The corresponding likelihood value (P-value) defines whether there are major or no significant variations between the two trees [130]. Therefore, if the confidence generated by the phylogenetic tree is strong, the review or conclusion of the research would not be misleading.

9.4 Limitation

Phylogenetic trees have various functional uses in different biological domains, such as conservation biology, epidemiology, forensics, cancer evolution, HIV transmission, gene expression prediction, protein structure prediction, and drug design [133]. However, for generating a more accurate tree, researchers face different challenges like memory efficiency. For instance, for the designing and execution of phylogeny programs, a very significant task is to minimize memory usage and improve cache performance for two principal reasons. Firstly, the latest generation of search algorithms has enabled the estimation of very broad trees feasible. Alignments are often adverse due to the vast accumulation of data. Memory use is constrained for major problems in most current programs. Secondly, the performance of central processing units has risen for many years and now it is to a point greater than the memory access performance such that the productivity of the vast variety of science applications is now restricted by their memory access patterns rather than the speed of the Processor. At present, this pattern is unlikely to be changed. Some techniques to address this burden consist of optimizing programs for minimizing memory usage on a technological basis, implementing a divide-and-conquer strategy to reduce the issue, and utilizing efficient common memory processors [134, 135].

Another significant limitation is the optimization of the probability function, which usually takes up more than 90% of the total deployment period in programs like RAxML or PHYML [135]. Some methods [135, 136], instead of being recalculated each time, rely on the identification of equivalent trends in the axis and reuse of previously calculated values. Additionally, a separate implementation of the computer-intensive functions in different nuclear replacement model would become important to the computation of very large trees, as will the introduction of low-level technological optimizations, such as manual loop unrolling, to the probability functions.

Given the advancement in algorithms in the area, only a few appropriate visualization resources for the study of very large trees are usable. Therefore, it is important to design modern tree viewing methods to speed up the observational phase and to derive valuable knowledge from the data, as well as speed up the cognitive process. Phylograms, radial, and slanting cladogram drawings (https://www.cs.ubc.ca/~tmm/papers/tj/) are among the most common depictions. Popular tree display programs such as ATV [137] have such depictions. These formats and systems, however, aim at medium-sized trees with up to 300–400 taxa. The approaches to broad trees allow use of two-dimensional and three-dimensional [138] hyperbolic space to simultaneously have a comprehensive and spatial visual image of the tree. Additional methods, such as SpaceTree [139], only show very broad trees as symbolic pieces. However, biologists typically favor a simultaneous detailed view of phylogenies and a qualitative view. Earlier Treemaps have also been suggested [140], although this definition is restricted to a range of 2000–3000 taxa. There are also methods focused on virtual reality [141], although these are not available to most researchers because of their very high infrastructure expenses. Carrizo [142] gives a reading and detailed analysis of phylogenetic trees from the viewpoint of knowledge visualization. However, since there is currently no genuinely satisfactory alternative, the creation of suitable visualization software becomes an exceedingly essential challenge, as otherwise, the details in broad phylogenies would be useless. Other current concerns involve the creation of more difficult and practical sequence evolution mathematical models, assessments of final tree consistency and precision, modern infer phylogenetic network approaches, and phylogenetic inference centered on gene-order results [133].

9.5 Conclusion and Future Perspective

Molecular phylogeny defines the connection between the artifacts in the sample. The phylogenetic tree is constructed using data from various sequence (majority from DNA or protein). The numerous methods of tree construction can be broadly classified into character or distance based. Molecular phylogeny has a large variety of uses, and the conclusion of the analysis may be deceptive, where the description of evolutionary trends is not sufficient. The meaning of the tree often depends on the confidence of the phylogenetic tree is examined. The phylogenetic tree's trust may be calculated by sampling methods (bootstrapping, jackknifing, and Bayesian simulation) and methodological methods (KH test and SH test). Therefore, if the trust produced by the phylogenetic tree is strong, it will not mislead the study's understanding or inference. In order to build more reliable trees, however, researchers face multiple obstacles, such as memory efficiency and the use and optimization of the probability function. However, in the near future, the creation of exciting new algorithms that dramatically decrease the required level of likeliness evaluation, combined with an enhanced understanding of previously high-performance system problems in the community, is likely to enable parallel conclusions on medium-sized trees PC clusters of 10,000–20,000 sequences.

Conflicts of Interest None.

Additional Information Figure 9.1b has been used under the terms of the Creative Commons Attribution License [20].

References

1. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012 May;13(5):303–14.
2. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. PNAS. 1977 Nov 1;74(11):5088–90.
3. Crick FH. On protein synthesis. Symp Soc Exp Biol. 1958;12:138–63.
4. Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. J Theor Biol. 1965 Mar 1;8(2):357–66.
5. Sanger F. Chemistry of insulin: determination of the structure of insulin opens the way to greater understanding of life processes. Science. 1959 May 15;129(3359):1340–4.
6. Doolittle RF, Feng DF. Reconstructing the evolution of vertebrate blood coagulation from a consideration of the amino acid sequences of clotting proteins. Cold Spring Harb Symp Quant Biol. 1987 Jan 1;52:869–74.
7. Fitch WM, Margoliash E. Construction of phylogenetic trees. Science. 1967 Jan 20;155 (3760):279–84.
8. Edwards SV. Is a new and general theory of molecular systematics emerging? Evolution. 2009 Jan;63(1):1–19.

9. Mäser P, Thomine S, Schroeder JI, Ward JM, Hirschi K, Sze H, et al. Phylogenetic relationships within cation transporter families of Arabidopsis. Plant Physiol. 2001 Aug 1;126(4):1646–67.
10. Marra MA, Jones SJM, Astell CR, Holt RA, Brooks-Wilson A, Butterfield YSN, et al. The genome sequence of the SARS-associated coronavirus. Science. 2003 May 30;300 (5624):1399–404.
11. Gray RD, Drummond AJ, Greenhill SJ. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. Science. 2009 Jan 23;323(5913):479–83.
12. Salipante SJ, Horwitz MS. Phylogenetic fate mapping. Proc Natl Acad Sci U S A. 2006 Apr 4;103(14):5448–53.
13. Baumann J. Use of homeoplastic auditory ossicles for chain defects within the scope of tympanoplasty. Z Laryngol Rhinol Otol. 1971 Feb;50(2):95–102.
14. Kuzuya T, Kimura Y, Hoshida S, Kodama K, Nakamura N, Hamanaka Y, et al. The effect of CV-4151, a selective inhibitor of thromboxane synthetase, on prostanoid formation and platelet aggregation in humans. Cardiovasc Drugs Ther. 1988 Dec;2(5):693–700.
15. Gronau I, Hubisz MJ, Gulko B, Danko CG, Siepel A. Bayesian inference of ancient human demography from individual genome sequences. Nat Genet. 2011 Sep 18;43(10):1031–4.
16. Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, et al. Genome-wide nucleotide-level mammalian ancestor reconstruction. Genome Res. 2008 Nov;18(11):1829–43.
17. Roy SS, Dasgupta R, Bagchi A. A review on phylogenetic analysis: a journey through modern era. Comput Mol Biosci. 2014 Sep 30;4(3):39–45.
18. Scott AD, Baum DA. Phylogenetic tree. In: Kliman RM, editor. Encyclopedia of evolutionary biology [Internet]. Oxford: Academic Press; 2016. p. 270–6. [cited 2020 Oct 21]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128000496002031.
19. Choudhuri S. Chapter 9 - Phylogenetic analysis**The opinions expressed in this chapter are the author's own and they do not necessarily reflect the opinions of the FDA, the DHHS, or the Federal Government. In: Choudhuri S, editor. Bioinformatics for beginners [Internet]. Oxford: Academic Press; 2014. p. 209–18. [cited 2018 Nov 6]. Available from: http://www.sciencedirect.com/science/article/pii/B9780124104716000098.
20. Ding G, Yu Z, Zhao J, Wang Z, Li Y, Xing X, et al. Tree of life based on genome context networks. PLoS One. 2008 Oct 9;3(10):e3357.
21. Xiong J. Essential bioinformatics. Cambridge: Cambridge University Press; 2006. 360 p.
22. Munjal G, Hanmandlu M, Srivastava S. Phylogenetics algorithms and applications. Ambient Communications and Computer Systems. 2018 Dec 10;904:187–94.
23. Kapli P, Yang Z, Telford MJ. Phylogenetic tree building in the genomic age. Nat Rev Genet. 2020 Jul;21(7):428–44.
24. El-Kebir M, Oesper L, Acheson-Field H, Raphael BJ. Reconstruction of clonal trees and tumor composition from multi-sample sequencing data. Bioinformatics. 2015 Jun 15;31(12):i62–70.
25. Paradis E, Claude J, Strimmer K. APE: analyses of Phylogenetics and evolution in R language. Bioinformatics. 2004 Jan 22;20(2):289–90.
26. Lord E, Leclercq M, Boc A, Diallo AB, Makarenkov V. Armadillo 1.1: an original workflow platform for designing and conducting phylogenetic analysis and simulations. PLoS One. 2012;7(1):e29903.
27. Suchard MA, Redelings BD. BAli-Phy: simultaneous Bayesian inference of alignment and phylogeny. Bioinformatics. 2006 Aug 15;22(16):2047–8.
28. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012 Aug;29(8):1969–73.
29. Jiang Y, Qiu Y, Minn AJ, Zhang NR. Assessing intratumor heterogeneity and tracking longitudinal and spatial clonal evolutionary history by next-generation sequencing. Proc Natl Acad Sci U S A. 2016 Sep 13;113(37):E5528–37.
30. Huson DH, Scornavacca C. Dendroscope 3: an interactive tool for rooted phylogenetic trees and networks. Syst Biol. 2012 Dec 1;61(6):1061–7.

31. Jeon Y-S, Lee K, Park S-C, Kim B-S, Cho Y-J, Ha S-M, et al. EzEditor: a versatile sequence alignment editor for both rRNA- and protein-coding genes. Int J Syst Evol Microbiol. 2014 Feb;64(Pt 2):689–91.
32. Price MN, Dehal PS, Arkin AP. FastTree 2--approximately maximum-likelihood trees for large alignments. PLoS One. 2010 Mar 10;5(3):e9490.
33. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015 Jan;32(1):268–74.
34. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007 Aug;24(8):1586–91.
35. Hellmuth M, Wieseke N, Lechner M, Lenhof H-P, Middendorf M, Stadler PF. Phylogenomics with paralogs. Proc Natl Acad Sci U S A. 2015 Feb 17;112(7):2058–63.
36. Lanfear R, Frandsen PB, Wright AM, Senfeld T, Calcott B. PartitionFinder 2: new methods for selecting partitioned models of evolution for molecular and morphological phylogenetic analyses. Mol Biol Evol. 2017 Mar 1;34(3):772–3.
37. Thomas GH, Hartmann K, Jetz W, Joy JB, Mimoto A, Mooers AO. PASTIS: an R package to facilitate phylogenetic assembly with soft taxonomic inferences. Methods Ecol Evol. 2013;4 (11):1011–7.
38. Schliep KP. Phangorn: phylogenetic analysis in R. Bioinformatics. 2011 Feb 15;27(4):592–3.
39. Deshwar AG, Vembu S, Yung CK, Jang GH, Stein L, Morris Q. PhyloWGS: reconstructing subclonal composition and evolution from whole-genome sequencing of tumors. Genome Biol. 2015 Feb 13;16(1):35.
40. Brown JW, Walker JF, Smith SA. Phyx: phylogenetic tools for unix. Bioinformatics. 2017 Jun 15;33(12):1886–8.
41. Wheeler WC, Lucaroni N, Hong L, Crowley LM, Varón A. POY version 5: phylogenetic analysis using dynamic homologies under multiple optimality criteria. Cladistics. 2015;31 (2):189–96.
42. Darriba D, Taboada GL, Doallo R, Posada D. ProtTest 3: fast selection of best-fit models of protein evolution. Bioinformatics. 2011 Apr 15;27(8):1164–5.
43. Knight R, Maxwell P, Birmingham A, Carnes J, Caporaso JG, Easton BC, et al. PyCogent: a toolkit for making sense from sequence. Genome Biol. 2007 Aug 21;8(8):R171.
44. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006 Nov 1;22(21):2688–90.
45. Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics. 2019 Nov 1;35(21):4453–5.
46. Sun Z, Zhu Q, Xiong Y, Sun Y, Mou L, Zhang L. TreeGen: A Tree-Based Transformer Architecture for Code Generation. arXiv:191109983 [cs] [Internet]. 2019 Nov 28. [cited 2020 Dec 14]; Available from: http://arxiv.org/abs/1911.09983.
47. Boc A, Diallo AB, Makarenkov V. T-REX: a web server for inferring, validating and visualizing phylogenetic trees and networks. Nucleic Acids Res. 2012 Jul;40(Web Server issue):W573–9.
48. Dong W, Liu J, Yu J, Wang L, Zhou S. Highly Variable Chloroplast Markers for Evaluating Plant Phylogeny at Low Taxonomic Levels and for DNA Barcoding. PLoS One [Internet]. 2012 Apr 12;7(4). [cited 2020 Oct 22]; Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3325284/.
49. Gielly L, Taberlet P. The use of chloroplast DNA to resolve plant phylogenies: noncoding versus rbcL sequences. Mol Biol Evol. 1994 Sep;11(5):769–77.
50. Zhu W-D, Nie Z-L, Wen J, Sun H. Molecular phylogeny and biogeography of Astilbe (Saxifragaceae) in Asia and eastern North America. Bot J Linn Soc. 2013 Feb 1;171 (2):377–94.
51. Akhani H, Malekmohammadi M, Mahdavi P, Gharibiyan A, Chase MW. Phylogenetics of the Irano-Turanian taxa of Limonium (Plumbaginaceae) based on ITS nrDNA sequences and leaf

anatomy provides evidence for species delimitation and relationships of lineages. Bot J Linn Soc. 2013 Mar 1;171(3):519–50.

52. Townsend TM, Alegre RE, Kelley ST, Wiens JJ, Reeder TW. Rapid development of multiple nuclear loci for phylogenetic analysis using genomic resources: an example from squamate reptiles. Mol Phylogenet Evol. 2008 Apr;47(1):129–42.
53. Smith DR. Mutation rates in plastid genomes: they are lower than you might think. Genome Biol Evol. 2015 Apr 13;7(5):1227–34.
54. Small RL, Cronn RC, Wendel JF. Use of nuclear genes for phylogeny reconstruction in plants. Aust Systematic Bot. 2004;17(2):145–70.
55. Boekhorst J, Snel B. Identification of homologs in insignificant blast hits by exploiting extrinsic gene properties. BMC Bioinformatics. 2007 Sep 21;8(1):356.
56. Tekaia F. Inferring Orthologs: open questions and perspectives. Genomics Insights. 2016;9:17–28.
57. Sang T. Utility of low-copy nuclear gene sequences in plant phylogenetics. Crit Rev Biochem Mol Biol. 2002;37(3):121–47.
58. Bragg JG, Potter S, Bi K, Moritz C. Exon capture phylogenomics: efficacy across scales of divergence. Mol Ecol Resour. 2016 Sep;16(5):1059–68.
59. Rubin BER, Ree RH, Moreau CS. Inferring phylogenies from RAD sequence data. PLoS One. 2012;7(4):e33394.
60. Peñalba JV, Smith LL, Tonione MA, Sass C, Hykin SM, Skipwith PL, et al. Sequence capture using PCR-generated probes: a cost-effective method of targeted high-throughput sequencing for nonmodel organisms. Mol Ecol Resour. 2014 Sep;14(5):1000–10.
61. Bi K, Vanderpool D, Singhal S, Linderoth T, Moritz C, Good JM. Transcriptome-based exon capture enables highly cost-effective comparative genomic data collection at moderate evolutionary scales. BMC Genomics. 2012 Aug 17;13:403.
62. Li C, Ortí G, Zhang G, Lu G. A practical approach to phylogenomics: the phylogeny of ray-finned fish (Actinopterygii) as a case study. BMC Evol Biol. 2007 Mar 20;7:44.
63. Portik DM, Smith LL, Bi K. An evaluation of transcriptome-based exon capture for frog phylogenomics across multiple scales of divergence (class: Amphibia, order: Anura). Mol Ecol Resour. 2016 Sep;16(5):1069–83.
64. Lemmon EM, Lemmon AR. High-throughput genomic data in systematics and Phylogenetics. Annu Rev Ecol Evol Syst. 2013 Nov 23;44(1):99–121.
65. Faircloth BC, McCormack JE, Crawford NG, Harvey MG, Brumfield RT, Glenn TC. Ultraconserved elements anchor thousands of genetic markers spanning multiple evolutionary timescales. Syst Biol. 2012 Oct;61(5):717–26.
66. Weitemier K, SCK S, Cronn RC, Fishbein M, Schmickl R, McDonnell A, et al. Hyb-Seq: Combining target enrichment and genome skimming for plant phylogenomics. Appl Plant Sci. 2014 Sep;2(9):1400042.
67. Schmickl R, Liston A, Zeisek V, Oberlander K, Weitemier K, Straub SCK, et al. Phylogenetic marker development for target enrichment from transcriptome and genome skim data: the pipeline and its application in southern African Oxalis (Oxalidaceae). Mol Ecol Resour. 2016 Sep;16(5):1124–35.
68. Kadlec M, Bellstedt DU, Le Maitre NC, Pirie MD. Targeted NGS for species level phylogenomics: "made to measure" or "one size fits all"? PeerJ. 2017;5:e3569.
69. Yue F, Shi J, Tang J. Simultaneous phylogeny reconstruction and multiple sequence alignment. BMC Bioinformatics. 2009 Jan 30;10(Suppl 1):S11.
70. Higgins DG, Sharp PM. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene. 1988 Dec 15;73(1):237–44.
71. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994 Nov 11;22(22):4673–80.
72. Notredame C. Recent progress in multiple sequence alignment: a survey. Pharmacogenomics. 2002 Jan;3(1):131–44.

73. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment11Edited by J. Thornton Journal of Molecular Biology. 2000 Sep 8;302 (1):205–17.
74. Ashkenazy H, Sela I, Levy Karin E, Landan G, Pupko T. Multiple sequence alignment averaging improves phylogeny reconstruction. Syst Biol. 2019 Jan 1;68(1):117–30.
75. Blackshields G, Wallace IM, Larkin M, Higgins DG. Analysis and comparison of benchmarks for multiple sequence alignment. In Silico Biol. 2006;6(4):321–39.
76. Chang J-M, Di Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol. 2014 Jun;31(6):1625–37.
77. Collingridge PW, Kelly S. MergeAlign: improving multiple sequence alignment performance by dynamic reconstruction of consensus multiple sequence alignments. BMC Bioinformatics. 2012 May 30;13:117.
78. Lake JA. The order of sequence alignment can bias the selection of tree topology. Mol Biol Evol. 1991 May;8(3):378–85.
79. Penn O, Privman E, Landan G, Graur D, Pupko T. An alignment confidence score capturing robustness to guide tree uncertainty. Mol Biol Evol. 2010 Aug;27(8):1759–67.
80. Lutzoni F, Wagner P, Reeb V, Zoller S. Integrating ambiguously aligned regions of DNA sequences in phylogenetic analyses without violating positional homology. Syst Biol. 2000 Dec;49(4):628–51.
81. Lücking R, Hodkinson BP, Stamatakis A, Cartwright RA. PICS-Ord: unlimited coding of ambiguous regions by pairwise identity and cost scores ordination. BMC Bioinformatics. 2011 Jan 7;12:10.
82. Wheeler WC. Sequence alignment, parameter sensitivity, and the phylogenetic analysis of molecular data. Syst Biol. 1995 Sep 1;44(3):321–31.
83. Privman E, Penn O, Pupko T. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol. 2012 Jan;29(1):1–5.
84. Sela I, Ashkenazy H, Katoh K, Pupko T. GUIDANCE2: accurate detection of unreliable alignment regions accounting for the uncertainty of multiple parameters. Nucleic Acids Res. 2015 Jul 1;43(W1):W7–14.
85. Arenas M. Trends in substitution models of molecular evolution. Front Genet [Internet]. 2015;6. [cited 2018 Nov 6]; Available from: https://www.frontiersin.org/articles/10.3389/fgene.2015.00319/full#B99.
86. Jukes TH, Cantor CR. Chapter 24 - Evolution of protein molecules. In: Munro HN, editor. Mammalian protein metabolism [Internet]. New York: Academic Press; 1969. p. 21–132. [cited 2020 Oct 23]. Available from: http://www.sciencedirect.com/science/article/pii/B9781483232119500097.
87. Collins DW, Jukes TH. Rates of transition and Transversion in coding sequences since the human-rodent divergence. Genomics. 1994 Apr 1;20(3):386–96.
88. Kimura M. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol. 1980 Dec;16(2):111–20.
89. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol. 1981 Nov 1;17(6):368–76.
90. Zharkikh A. Estimation of evolutionary distances between nucleotide sequences. J Mol Evol. 1994 Sep 1;39(3):315–29.
91. Hasegawa M, Kishino H, Yano T. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. J Mol Evol. 1985;22(2):160–74.
92. Yang Z. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J Mol Evol. 1994 Sep 1;39(3):306–14.
93. Shoemaker JS, Fitch WM. Evidence from nuclear sequences that invariable sites should be considered when sequence divergence is calculated. Mol Biol Evol. 1989 May;6(3):270–89.
94. Darriba D, Taboada GL, Doallo R, Posada D. jModelTest 2: more models, new heuristics and parallel computing. Nat Methods. 2012 Aug;9(8):772.

95. Arenas M, Posada D. Simulation of genome-wide evolution under heterogeneous substitution models and complex multispecies coalescent histories. Mol Biol Evol. 2014 May;31 (5):1295–301.
96. Sumner JG, Jarvis PD, Fernández-Sánchez J, Kaine BT, Woodhams MD, Holland BR. Is the general time-reversible model bad for molecular Phylogenetics? Syst Biol. 2012 Dec 1;61 (6):1069–74.
97. Gatto L, Catanzaro D, Milinkovitch MC. Assessing the applicability of the GTR nucleotide substitution model through simulations. Evol Bioinformatics Online. 2007 Feb 4;2:145–55.
98. Jayaswal V, Jermiin LS, Poladian L, Robinson J. Two stationary nonhomogeneous Markov models of nucleotide sequence evolution. Syst Biol. 2011 Jan;60(1):74–86.
99. Muse SV, Gaut BS. A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol. 1994 Sep;11(5):715–24.
100. Goldman N, Yang Z. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol. 1994 Sep;11(5):725–36.
101. Gupta MK, Vadde R. Genetic basis of adaptation and maladaptation via balancing selection. Zoology. 2019 Jul;10:125693.
102. Gupta MK, Vadde R. Divergent evolution and purifying selection of the Type 2 diabetes gene sequences in Drosophila: a phylogenomic study. Genetica [Internet]. 2020 Aug 17 . [cited 2020 Aug 29]; https://doi.org/10.1007/s10709-020-00101-7.
103. Kosakovsky Pond SL, Frost SDW. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol. 2005 May 1;22(5):1208–22.
104. Pond SLK, Frost SDW. A genetic algorithm approach to detecting lineage-specific variation in selection pressure. Mol Biol Evol. 2005 Mar;22(3):478–85.
105. Yang Z, Nielsen R, Goldman N, Pedersen AM. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics. 2000 May;155(1):431–49.
106. Wong WSW, Sainudiin R, Nielsen R. Identification of physicochemical selective pressure on protein encoding nucleotide sequences. BMC Bioinformatics. 2006 Mar 16;7:148.
107. Schneider A, Cannarozzi GM, Gonnet GH. Empirical codon substitution matrix. BMC Bioinformatics. 2005 Jun 1;6(1):134.
108. Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 2008 Mar;25(3):568–79.
109. Misawa K. A codon substitution model that incorporates the effect of the GC contents, the gene density and the density of CpG islands of human chromosomes. BMC Genomics. 2011 Aug 6;12:397.
110. Perez-Jimenez R, Inglés-Prieto A, Zhao Z-M, Sanchez-Romero I, Alegre-Cebollada J, Kosuri P, et al. Single-molecule paleoenzymology probes the chemistry of resurrected enzymes. Nat Struct Mol Biol. 2011 May;18(5):592–6.
111. Alvarez-Ponce D, Fares MA. Evolutionary rate and duplicability in the Arabidopsis thaliana protein–protein interaction network. Genome Biol Evol. 2012 Dec 1;4(12):1263–74.
112. Fares MA, Barrio E, Sabater-Muñoz B, Moya A. The evolution of the heat-shock protein GroEL from Buchnera, the primary endosymbiont of aphids, is governed by positive selection. Mol Biol Evol. 2002 Jul 1;19(7):1162–70.
113. Dayhoff MO, Schwartz RM, Orcutt BC. 22 a model of evolutionary change in proteins. In: Atlas of protein sequence and structure. Silver Spring: National Biomedical Research Foundation; 1978. p. 345–52.
114. Adachi J, Waddell PJ, Martin W, Hasegawa M. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. J Mol Evol. 2000;50 (4):348–58.
115. Kosiol C, Goldman N. Different versions of the Dayhoff rate matrix. Mol Biol Evol. 2005 Feb 1;22(2):193–9.

116. Liberles DA, Teichmann SA, Bahar I, Bastolla U, Bloom J, Bornberg-Bauer E, et al. The interface of protein structure, protein biophysics, and molecular evolution. Protein Sci. 2012;21(6):769–85.
117. Halpern AL, Bruno WJ. Evolutionary distances for protein-coding sequences: modeling site-specific residue frequencies. Mol Biol Evol. 1998;15(7):910–7.
118. Taverna DM, Goldstein RA. The distribution of structures in evolving protein populations. Biopolymers. 2000;53(1):1–8.
119. Arenas M, Sánchez-Cobos A, Bastolla U. Maximum-likelihood phylogenetic inference with selection on protein folding stability. Mol Biol Evol. 2015 Aug 1;32(8):2195–207.
120. Pardi F, Gascuel O. Combinatorics of distance-based tree inference. PNAS. 2012 Oct 9;109 (41):16443–8.
121. Felsenstein J, Felenstein J. Inferring phylogenies, vol. 2. Sunderland: Sinauer Associates; 2004.
122. Yang Z. Computational molecular evolution. Oxford: OUP; 2006. 375 p.
123. Saitou N, Nei M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol. 1987;4(4):406–25.
124. Roch S. Toward extracting all phylogenetic information from matrices of evolutionary distances. Science. 2010 Mar 12;327(5971):1376–9.
125. Steel M. A basic limitation on inferring phylogenies by pairwise sequence comparisons. J Theor Biol. 2009 Feb 7;256(3):467–72.
126. Huelsenbeck JP. Is the Felsenstein zone a fly trap? Syst Biol. 1997 Mar;46(1):69–74.
127. Whelan S, Liò P, Goldman N. Molecular phylogenetics: state-of-the-art methods for looking into the past. Trends Genet. 2001 May;17(5):262–72.
128. Huelsenbeck JP, Larget B, Miller RE, Ronquist F. Potential applications and pitfalls of Bayesian inference of phylogeny. Syst Biol. 2002;51(5):673–88.
129. Holder M, Lewis PO. Phylogeny estimation: traditional and Bayesian approaches. Nat Rev Genet. 2003 Apr;4(4):275–84.
130. Challa S, Neelapu NRR. Phylogenetic trees: applications, construction, and assessment. In: Hakeem KR, Shaik NA, Banaganapalli B, Elango R, editors. Essentials of bioinformatics, In silico life sciences: agriculture [Internet], vol. III. Cham: Springer International Publishing; 2019. p. 167–92. https://doi.org/10.1007/978-3-030-19318-8_10. [cited 2020 Oct 24].
131. Kishino H, Hasegawa M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol. 1989 Aug 1;29(2):170–9.
132. Shimodaira H, Hasegawa M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. Mol Biol Evol. 1999 Aug 1;16(8):1114.
133. Stamatakis A. Phylogenetics: applications, Software and Challenges. Cancer Genomics Proteomics. 2005 Sep 1;2(5):301–5.
134. Charalambous M, Trancoso P, Stamatakis A. Initial experiences porting a bioinformatics application to a graphics processor. In: Bozanis P, Houstis EN, editors. Advances in informatics, Lecture notes in computer science. Berlin: Springer; 2005. p. 415–25.
135. Stamatakis A, Ott M, Ludwig T. RAxML-OMP: an efficient program for phylogenetic inference on SMPs. In: Malyshkin V, editor. Parallel computing technologies. Berlin: Springer; 2005. p. 288–302. (Lecture Notes in Computer Science).
136. Kosakovsky Pond SL, Muse SV. Column sorting: rapid calculation of the phylogenetic likelihood function. Syst Biol. 2004 Oct;53(5):685–92.
137. Zmasek CM, Eddy SR. ATV: display and manipulation of annotated phylogenetic trees. Bioinformatics. 2001 Apr 1;17(4):383–4.
138. Hughes T, Hyun Y, Liberles DA. Visualising very large phylogenetic trees in three dimensional hyperbolic space. BMC Bioinformatics. 2004 Apr 29;5(1):48.

139. Plaisant C, Grosjean J, Bederson BB. Spacetree: supporting exploration in large node link tree, design evolution and empirical evaluation. In: IEEE Symposium on Information Visualization, 2002 INFOVIS 2002; 2002. p. 57–64.
140. Arvelakis A, Reczko M, Stamatakis A, Symeonidis A, Tollis IG. Using treemaps to visualize phylogenetic trees. In: Oliveira JL, Maojo V, Martín-Sánchez F, Pereira AS, editors. Biological and medical data analysis. Berlin: Springer; 2005. p. 283–93. (Lecture Notes in Computer Science).
141. Stolk B, Abdoelrahman F, Koning A, Wielinga P, Neefs J-M, Stubbs A, et al. Mining the human genome using virtual reality. In: Proceedings of the Fourth Eurographics Workshop on Parallel Graphics and Visualization. Goslar, DEU: Eurographics Association; 2002. p. 17–21. (EGPGV ‘02).
142. Carrizo SF. Phylogenetic trees: an information visualisation perspective. In: Proceedings of the second conference on Asia-Pacific bioinformatics, vol. 29. Darlinghurst: Australian Computer Society, Inc.; 2004. p. 315–20. (APBC ‘04).

RNA Structure Prediction

10

Manoj Kumar Gupta, Gayatri Gouda, Ravindra Donde,
Piyali Goswami, N. Rajesh, Pallabi Pati, Sushil Kumar Rathore,
Ramakrishna Vadde, and Lambodar Behera

Abstract

One of the significant forms of molecules present in living cells is ribonucleic acid (RNA). RNA structural elements moderate various biological process, including epigenetic function, modify mRNA stability, and alternate splicing. The study of the secondary structures of RNA is, therefore, crucial for interpreting the role as well as the regulatory mechanism of RNA transcripts. But experimental methods are tedious, time-consuming, pricey, requires special equipment, and, thus, cannot often be implemented. Methods for statistical simulation are an option and parallel to experimental approaches. Additionally, the findings from the RNA-Puzzles, joint research on the estimation of RNA structures, suggest that computational methods can be employed for effective RNA modeling. However, there is still space for improvement. Considering this, in the chapter, authors attempted to understand the various forms of RNA and how computational approaches can be employed to predict their structure more precisely. The RNA

M. K. Gupta · G. Gouda · R. Donde · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

P. Goswami
Department of Biotechnology, IIT Kharagpur, Kharagpur, West Bengal, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_10

is classified mainly according to its existence, role, and structure into three groups, messenger RNA, transfer RNA, and ribosomal RNA. To date, numerous algorithms, and tools, have been designed for predicting the secondary structure of RNA. However, since three-dimensional structures are highly required for getting insight into the function of the RNA, few approaches have also been developed for predicting tertiary structures of RNA atoms. However, the authors believe that, in the near future, by combining experimental and computational approaches, we will be able to predict the structure of RNA more accurately, which in turn will enable us to understand its structure and function more precisely.

Keywords

RNA · tRNA · mRNA · rRNA · Structure prediction

Abbreviations

CSA	Comparative sequence analysis
DP	Dynamic programming
dsRNAs	Double-strand RNAs
lncRNAs	Long noncoding RNA
miRNAs	microRNA
MMP	MC-fold/Mc-Sym Pipeline
mRNA	Messenger RNA
PPV	Positive predictive value
pre-rRNA	Precursors-rRNAs
rDNA	Ribosomal DNA
RNA	Ribonucleic acid
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
siRNA	Short interfering RNA
snoRNAs	Small nucleolar RNAs
SSs	Secondary structures
t6A	Threonyl-carbamoyl adenosine
tRNA	Transfer RNA

10.1 Introduction

One of the significant forms of molecules present in living cells is ribonucleic acid (RNA) [1]. After the central dogma was hypothesized in 1950, the key function assigned to RNA was to serve as the intermediary between DNA and protein synthesis [2]. However, out of ~70% of the genome transcribed, only a limited portion encodes for protein sequences [3], which means that most RNAs might have various biological functions. Earlier, several researchers have also suggested that RNAs are transporters for genetic material and are also associated with many biological processes that are incredibly significant [4]. For instance, RNA transcripts fold into structures (SSs) (Fig. 10.1), which have different catalytic, ligand, and scaffolding functions that shape a crucial biological regulatory activity. RNA structural elements moderate epigenetic function, modify mRNA stability and translation, scaffold large macromolecular complexes, transduce signals, and monitor alternate splicing. The study of the SSs of RNA is, therefore, crucial for interpreting the role as well as the regulatory mechanism of RNA transcripts [2].

RNA folds into a 3D system through hydrogen interaction and base-stacking, which in the sequence are not consecutive [7]. The 3D structure of the RNA molecule decides its function, like proteins. In order to construct a 3D model, high-resolution experimental methods such as crystallography [8, 9], cryo-EM [10], and nuclear magnet resonance spectroscopy may be taken advantage of [11]. But experimental methods are tedious, time-consuming, pricey, requires special equipment, and, thus, cannot often be implemented. Methods for statistical simulation are an option and parallel to experimental approaches. Additionally, the findings from the RNA-Puzzles [12], joint research on the estimation of RNA structures, suggest that computational methods can be employed for effective RNA modeling. However, there is still space for improvement.

Like proteins, RNAs can be divided into families [13], which originated from a common ancestor. RNA sequences from the same family will have higher similarity, and the study of sequence conservation may be used for identifying important conserved areas, such as areas binding ligands, active sites, or other important functions. The Watson crick basis pairing pattern for the RNA is often used to forecast SSs. According to the CompaRNA [14], RNA alignments methods such as PETfold [15] outweigh the predictive single sequence methods of the secondary RNA structure. RNA alignments may also be used to enhance the prediction of the tertiary structure [16]. For instance, recently, a group of researchers employed a novel approach for exploring tertiary structure predictions [13]. The methodology examines the usage of multiple alignment knowledge and simultaneous RNA homolog simulation to strengthen ab initio RNA structure modeling techniques. A new technique, called EvoClustRNA, is focused on a conventional strategy to predict RNA structures, utilizing evolutionary knowledge from distant sequence homologs [17]. On the basis of the empirical finding that RNA sequences of the same RNA family normally fold into identical 3D structures, they have checked whether computational modeling can be driven by searching for a global helical sequence for the target sequence, which is shared through de novo models of various sequence homologs. EvoClustRNA is the first effort to use this method for RNA 3D prediction. Thus, in this chapter, the authors attempt to understand the sources, form, and role of the RNA structure and how different computational approaches that researchers are adopting for determining various RNA structures.

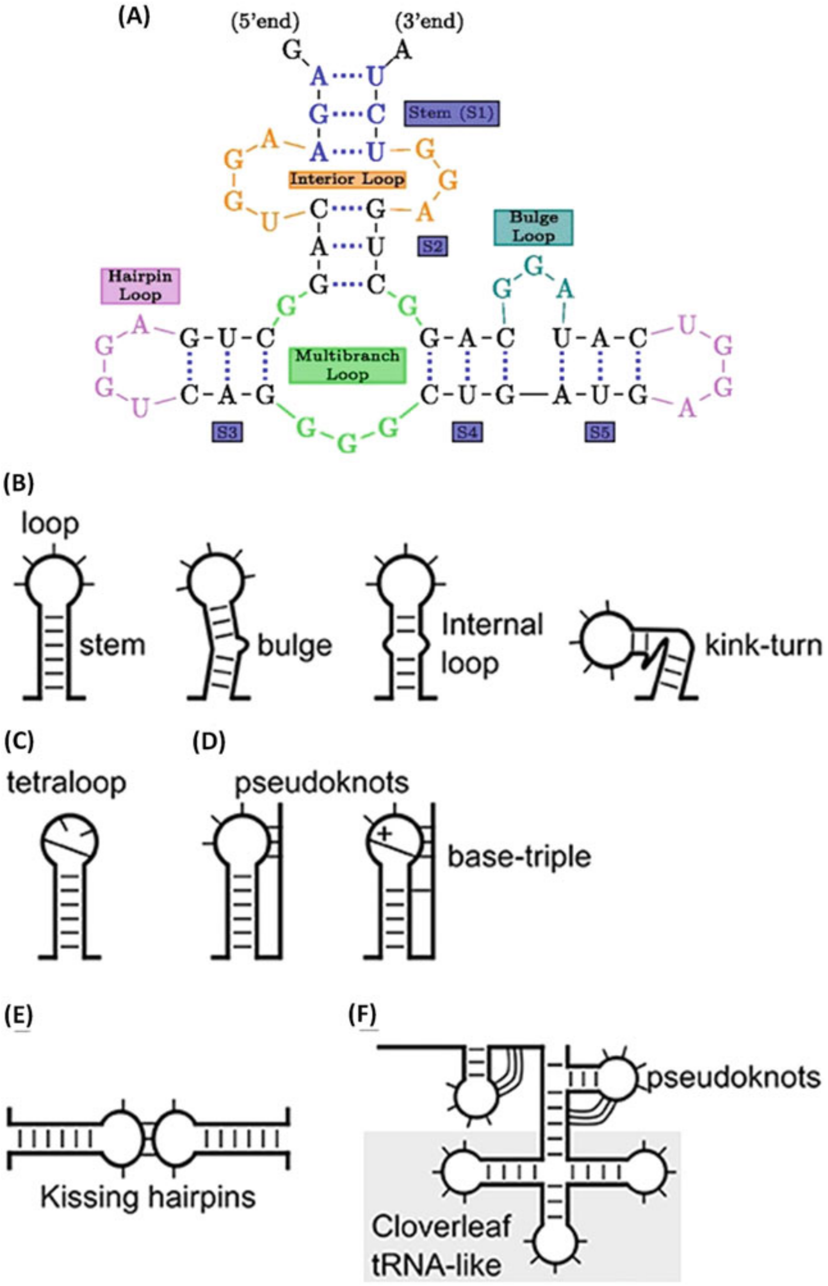


Fig. 10.1 (**a**) Basic structural motifs depicted within RNA secondary structures. (**b**) The simplest form of RNA structure is a stem-loop. A stem-loop is shown with a bulge, internal loop, or (**c**) tetraloop. (**d**) The loop can also base-pair with upstream or downstream sequences to form a pseudoknot. (**e**) Interaction between the loops of two stem-loops forms kissing hairpins. (**f**) A relatively complex structure is a cloverleaf or tRNA-like structure that often consists of multiple stem-loops and pseudoknots. (Adapted from [5, 6])

10.2 RNA Structure

The RNA is classified mainly according to its existence, role, and structure into three groups: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). The rRNA produces complex three-dimensional structures that interact with polypeptides to shape ribosomes responsible for protein synthesis in organelles. The ribosomes act as an mRNA encoding tool. The mRNA includes instructions that dictate protein amino acid sequences. The tRNA serves as an adapter to convert mRNA codons into those amino acids [18]. In addition, as discussed below, there are also other forms of RNAs, like long noncoding RNA (lncRNAs), small nucleolar RNAs (snoRNAs), microRNA(miRNAs), and short interfering RNA (siRNA) [19].

10.2.1 Messenger RNA

The "messenger" RNA is mRNA. The mRNA in the nucleus is synthesized using the DNA nuclear sequence as a reference. This process needs nucleotide triphosphates as substrate and is catalyzed by the RNA polymerase II enzyme. The DNA to mRNA processing is called transcription and takes place in the nucleus. The mRNA produced in the nucleus is transferred to ribosomes out of the nucleus and into a cytoplasm. Subsequently, the mRNA guides the protein synthesis that takes place in the cytoplasm. On the ribosomes, proteins are packaged using the mRNA sequence as a reference. Thus, mRNA bears a "message" to the cytoplasm from the nucleus for encoding protein. The processing of mRNA to proteins is called translation [20]. Earlier studies have reported that while the configuration and mode of action of the prokaryotic and eukaryotic mRNAs vary, similarities still exist. In mRNA, genetic information is encoded into a four-base nucleotide alphabet, which forms codons of three bases. Each codon codes for a certain amino acid except for stop codons that specify when the synthesis of protein stops. The mRNA is translated by the codon-reading ribosome. For all prokaryotes and eukaryotes, the beginning or initiator codon is an AUG sequence, and the sequences are read in $5'$ to $3'$ direction. Eukaryotic mRNA normally codes for one specific (monocistronic) protein, whereas the prokaryotic mRNA typically codes for a set of similar (polycistronic) proteins on the same mRNA. Polycistronic mRNA guides the synthesis of each coded polypeptide, which is more or less simultaneous. For example, the *trp* operon is a DNA, which is transcribed in mRNA and codes for six polypeptides, catalyzes the synthesis of tryptophan.

The mRNA has a shorter life in comparison to the DNA. An mRNA molecule may be stored, edited, and transported before translation following transcription [21]. For many factors, mRNA stabilization is an important control point in the regulation of gene expression. At first, an equilibrium between its synthesis and degradation is highly required for consistent normal function. Secondly, the consistency of individual mRNAs may be altered because of multiple environmental stimuli, such as carbon source, viral diseases, and developmental transformations, allowing rapid shifts in the expression of the gene. Thirdly, a method of competent mRNA degradation to remove deleterious errors during mRNA synthesis. Finally, successful mRNA degradation is essential for the growth of both the prokaryotes and the eukaryotes [22].

10.2.2 Ribosomal RNA

The biosynthesis of rRNA and its integration into the ribosomes is a surprisingly complex process, which for over three decades, has been the focus of intensive study. Ribosome biogenesis starts in the nucleolus, in accordance with the "RNA-base machine," through the synthesis of the large primary mRNA transcripts via the RNA polymerase I (Pol I) [23]. In eukaryotes, the mature 80S cytoplasmic ribosome is composed of the 60S larger subunit and the 30S smaller subunit. The small subunit is composed of 18S rRNAs and more than 30 ribosomal proteins. The large subunit comprises 5.8S, 25S/28S, and 5S rRNAs and over 40 ribosomal proteins. Biogenesis of ribosome includes replication of ribosomal DNA (rDNA), production of precursors-rRNAs (pre-rRNA), modifications to the RNA, and assembly of ribosomal protein and assembly factors in rRNA. Ribosome biogenesis is an important, complicated, and energy-intensive mechanism strictly controlled by endogenous signals and environmental factors, such as ambient temperature. Within eukaryotic cells, irregular biogenesis of rRNA stimulates "RNA Nucleus Quality Regulation," inducing higher polyadenylation of some intermediate rRNA products as well as by-products, known as TRAMPs (Trf/Air/Mtr4 polyadenylation complex). The nuclear exosome complex sequentially degrades these intermediates. Ribosomal biogenesis failure results in significant developmental of deficiencies in higher plants and extreme hereditary disorders in mammals [24].

The catalytic function of rRNA was first shown by Harry Noller and his colleagues' 1992 experiments. These researchers found that even after about 95% of the ribosomal proteins have been discarded via traditional protein extraction methods, the large ribosomal unit would catalyze peptide bond formation ("Peptidyl Transferase Reaction"). In comparison, RNase treatment fully abolishes the development of peptide bindings, which clearly supports the theory that peptide binding development is an "RNA-catalyzed reaction." Further experiments have also validated as well as expanded these findings by showing that the "peptidyl transferase reaction" can be catalyzed by synthetic fragments of 23S rRNA in the complete absence of any ribosomal protein. These findings support that rRNA catalyzes the basic reaction in protein synthesis.

Apart from being the ribosomes' basic catalytic constituents, ribosomal proteins can also be used for promoting proper rRNA folding and for boosting the ribosomes' activity through proper tRNAs' positioning [25]. The direct presence of rRNA during the peptidyl transmission response has significant evolutionary consequences. RNAs are considered to be the first macromolecules that have self-replicating properties. Earlier studies have also supported this theory by stating that ribozymes like RNase P as well as self-splicing introns can catalyze RNA substrate reactions. The rRNA's function in the peptide attachment formation expands the catalytic action of RNA to direct participation in the synthesis of protein. Few studies have also revealed that the "Tetrahymena rRNA ribozyme" can catalyze the amino acid binding of RNA, thus adding credence to the likelihood of RNAs, rather than protein, being the initial aminoacyl tRNA synthesis. Thus, the RNA molecules may also serve as a significant biomarker toward understanding the early evolution of cells in catalyzing the reactions needed for self-replication as well as for protein synthesis [25].

10.2.3 transfer RNA

tRNA is a small nucleotide chain. The tRNA acts as an "adapter" molecule with an L-shape configuration, which converts the three-nucleotide codon sequence of the mRNA into the required amino acid of that codon. The tRNAs define the genetic code as the bond between amino acids and nucleic acids. However, their functions extended beyond protein translation, providing a remarkable set of tasks in the synthesis of bacterial cell walls, viral replication, cell tension, and even regulation of animal behavior [26]. Bacteria have multiple antibiotic mechanisms, which in the clinic is a growing obstacle. Within bacterial outer membrane lipids, tRNA-dependent aminoacylation offers improved virulence and tolerance to the cationic antimicrobial peptide [27]. Earlier, Fields and his team have studied the well-documented pathways of lipid aminoacylation to illustrate the usage of aminoacyl-tRNA substrates as an amino acid donor in lipid changes for improved antibiotic tolerance by the aminoacyl-phosphatidylglycerol synthases [28]. Emerging data also suggest that the tRNA genes perform a new function in bacterial conjugation. For instance, Alamos and his team found that 36 out of its 95 tRNAs are encoded in an integrative-conjugative genetic variable within *acidithiobacillus ferrooxidans* [29]. Castillo and his team have also shown that the integrases encoded inside the conjugative factor recognize the area of the tRNA stem-loop for active and location-specific recombination [30].

Mature tRNAs are abundant in nucleotide-based post-transcriptional modifications. These improvements perform important roles in the management of translation and reading frames [31], tRNA reliability, and transport [32]. Modification may occur within the anticodon as well. Phylogenetic analysis has been performed by Rafels-Ybern and his team to show the production of adenine base modification to Inosin (I) at location 34. The A to I shift affects the tRNA's ability to decipher the codon's third nucleotide location. Whereas A34 forms an optimal relationship with U, I34 is equally well informed of U, C, or A. The I34 function is to broaden wobble decoding, as investigated by another group of researchers [33]. The switch to I34 requires one tRNA to read three codons of the same amino acid. While the alteration is widely used in eukaryotes, there were limited earlier examples in bacteria. Earlier, researchers have also found many possible I34 locations for tRNAs' modifications in Firmicutes as well as Cyanobacteria genomes [34].

Changes can often involve the shipment of tRNA within cellular compartments. In an interesting study, Kessler et al. explain how the transportation of tRNAs between various areas of the cell influences the maturation and alteration of tRNAs [35]. Seminary tests of the sleeping causative agent, *Trypanosoma brucei*, indicate that tRNA-Tyr is transferred to the cytoplasm where the intron in an immature tRNA is broken. The tRNA is then reimported to the nuclease in a process called "retrograde transport," in which the spliced tRNA is necessary for modification with queuosin [32]. Some tRNA variations are retained in all life types. In a few tRNAs, the modification threonyl-carbamoyl adenosine (t6A) is found to decipher the ANN codons and is necessary for the stabilization of the duplex codon–anticodon. The t6A modification at loci 37, next to the anticodon loop, has been found to be necessary for the operation of *Streptococcus mutans*' anticodon nuclease PrrC [31]. The nuclease facilitates the bacterial cell death under stress conditions or during phage infection when the $\text{tRNA}^{\text{Lys}}{}_{\text{UUU}}$ anticodon loop is precisely cleaved.

10.2.4 Small Nucleolar RNAs

snoRNAs are generally composed of 60–170 nuclear nucleotides (with few exceptions) [36, 37] and are mainly involved in directing post-transcriptional alteration of nonprotein-coding RNAs (rRNAs, snRNAs) [38]. snoRNAs are broadly categorized as either a "C/D box" or "H/ACA box" based on the given sequence as well as SSs component [37]. "C/D box" directs 2′-O-methylation and "H/ACA" nuclear pseudouridylation upon target molecules. Since the 5′ as well as 3′ ends of the molecule fold into a stem configuration, which in turn creates a "kink switch," the "C box" ("RUGAUGA", R = A or G) and the "D box" ("CUGA") sequence motifs of the "C/D box" are brought closely into contact. The majority of the C/D boxes have another less conserved C as well as D box motifs, namely the C′ and D′ boxes, within the "Central SnoRNA region." C/D box is mainly involved in the ribonucleoprotein (RNP) complexes that also include 15.5 K, NOP56, NOP58, and fibrillarine proteins [39, 40]. The latter catalyzes the 2′-O-methylation of ribose molecules within the target RNA [40]. "H/ACA box" is a well-designed SSs comprising two hairpins connected together through a single-stranded area designated as the "H box" ("ANANNA", N = A, C, G or U) as well as the "ACA box" ("AYA", Y = C or U) at the 3′ end [41]. "H/ACA" produces "H/ACA" snoRNA and a group of four proteins, namely, Nop10, Gar1, Dyskerin, and Nhp2, where Dyskerin functions as pseudouridine synthase [42]. Primary identification of "H/ACA box" often includes RNA–RNA interactions between single-stranded area within the inner loops of the two snoRNA hairpin systems, mostly with target RNA [43, 44].

10.2.5 microRNA

miRNAs are small noncoding RNAs that have a mean length of ~22 nucleotides. The majority of the miRNAs are transcribed into prime miRNAs from DNA sequences and converted into precursor miRNAs as well as mature miRNAs. In certain instances, miRNA interacts with the 3′ UTR of the objective mRNA for suppressing expression. However, there have also been records of the association of miRNAs with other regions, including the coding sequence, 5′ UTR, and gene promoters. Moreover, under some circumstances, miRNAs have been shown to cause gene expression. In recent research, miRNAs have been shuttled between various subcellular cells to regulate the rate of translation and also transcription [45]. miRNAs are critical for the natural growth of animals and active in various biological processes [46]. Aberrant expression of miRNAs is related to a variety of human diseases [47, 48]. miRNAs are often secreted within extracellular fluids. Extracellular miRNAs can serve as plausible biomarkers for a number of diseases and signaling molecules for cell–cell interaction [49].

10.2.6 Short Interfering RNA

siRNAs are derived from double-strand RNAs (dsRNAs), consisting of two antisenses as well as a sense RNA strand that forms 19–25 bp duplex with 3′ dinucleotide overhangs. The antisense strand is a perfect reverse complement to the expected mRNA target. Few important functions of siRNAs include mainly post-transcriptional gene silencing or translation inhibition, exogenous DNA defense, intervention in epigenetic processes, and preserving genome integrity by transcriptional silencing. It has been used for industrial purposes to easily research in vivo gene expression owing to its capacity to knock out genes. Many of the siRNA measurement applications are therefore planned to aim siRNA sequences optimally to knock out genes. Subsequently, siRNAs' prediction can also be used to establish protocols for screening and can be used to classify new pathways to confirm cellular targets correlated with diseases such as hepatitis, cancer, and HIV infection [50].

10.2.7 Long Noncoding RNA

lncRNAs are classified as >200 nucleotide RNA molecules. While this differentiation is rather subjective and dependent on functional aspects of RNA separation techniques, lncRNAs vary from miRNAs as well as other sRNAs. In significant amounts, lncRNAs are present within the genome. They may not usually have working open read frames (ORFs). However, the discovery of bifunctional RNAs with coding-independent and protein-coding functions is flexible by this distinction, which increases the probability that certain protein-coding genes might have non-coding functions, as well [51]. Many lncRNAs are poorly expressed and, thus, researchers experience difficulties during exploring lncRNAs and understanding why lncRNAs were always considered to be "transcriptional noise." RNA-sequences in various tetrapods indicate that mostly (~81%), primate-specific lncRNAs are poorly retained in the DNA chain. However, it is worth remembering that many lncRNAs are highly conserved within the DNA sequence, and ~3% of lncRNAs might have originated earlier than 300 million years ago [52].

lncRNAs may be fast-evolving species of RNA that can play key roles in lineage specifying. A comparison of the matching tissues in *Rattus norvegicus*, *Mus musculus castaneus*, and *Mus musculus domesticus* indicates that shifts in the transcription levels of the adjacent protein-coding genes are linked with the appearance or disappearance of the lncRNAs [53]. There are several instances of lncRNAs with retained biological roles but low-level sequence survival, such as TUNA/megamind correlated with the growth of the brain in zebrafish, mouse, and humans [54, 55], and X-inelective unique transcript (*Xist*) involved in X-inactivation [56]. RNA molecules can require fewer sequence retention in order to maintain their function than proteins. Conversely, lncRNA promoters have a strong sequence conservation, which is even higher in comparison to protein-coding-gene promoters [57], indicating that lncRNA expression control is significant.

10.3 RNA Structure Prediction

RNA plays various cellular functions, and, thus, recognizing RNA structure is essential to understand its action mechanism [58]. Because the prediction of the three-dimensional RNA structure is difficult and expensive, scientists mainly depend on RNA's SSs. Hence, to date, numerous algorithms have been designed for predicting the SSs of RNA [19]. However, since three-dimensional structures are highly required for getting insight into the function of the RNA, few approaches have also been developed for predicting tertiary structures of RNA atoms [59] (Fig. 10.2 and Tables 10.1 and 10.2).

10.3.1 RNA SSs Prediction Methods

Present methods of prediction of SSs of RNA may be broadly categorized into comparative sequence analysis (CSA) and folding algorithms with thermodynamic, predictive, or probabilistic scoring schemes [81]. CSA distinguishes base pairs between homologous sequences. These approaches are incredibly accurate [82] if there are a sufficient number of compatible sequences and are aligned with professional expertise manually. However, to date, only a few thousand RNA families have been identified. Therefore, the most popular method used for the RNA SSs prediction is to fold an individual RNA sequence according to a suitable scoring feature. In this method, the RNA structure is separated into substructures, such as loops and trunks in the closest model [83]. Dynamic programming (DP) algorithms are then used to find minimal or probabilistic global structures from such substructures. Subsequently, experimental technique [84] (e.g., RNAshapes [85],

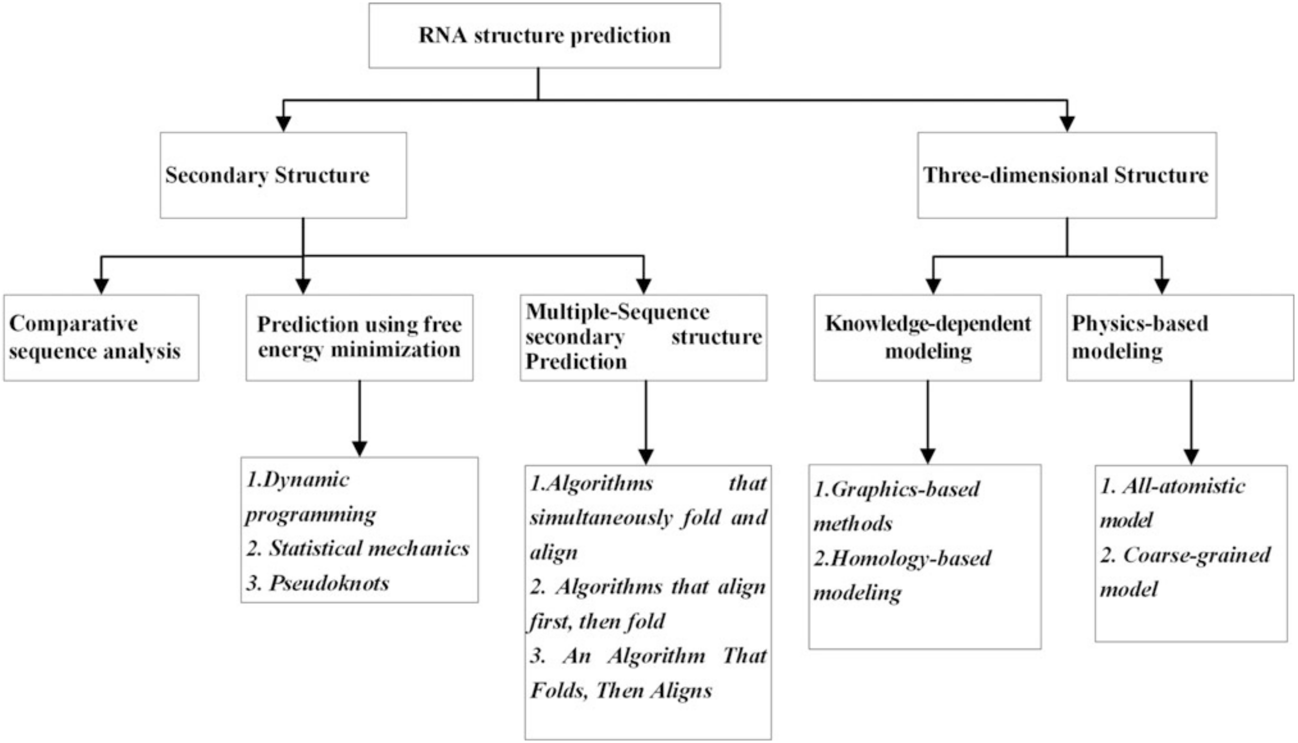

RNA structure prediction
Secondary Structure
Three-dimensional Structure
Comparative sequence analysis
Prediction using free energy minimization
Multiple-Sequence secondary structure Prediction
Knowledge-dependent modeling
Physics-based modeling
1.Dynamic programming
2. Statistical mechanics
3. Pseudoknots
1.Algorithms that simultaneously fold and align
2. Algorithms that align first, then fold
3. An Algorithm That Folds, Then Aligns
1.Graphics-based methods
2.Homology-based modeling
1. All-atomistic model
2. Coarse-grained model

Fig. 10.2 Different approaches to predict secondary and three-dimensional structure of RNA

Table 10.1 Softwares and tools for predicting secondary structure of RNA

Name	Description	References
CentroidAlifold	Employs generalized centroid estimator	[60]
DAFS	Align and fold RNA sequences through dual decomposition.	[61]
MASTR	Uses Markov chain Monte Carlo in a simulated annealing framework	[62]
Multilign	Utilizes multiple Dynalign calculations for finding a low free energy structure that is common to numerous sequences. It does not need sequence identity.	[63]
Murlet	Uses iterative alignment dependent on Sankoff's algorithm having sharply decreased computational time as well as memory.	[64]
MXSCARNA	Employs progressive alignment	[65]
PARTS	Probabilistic model and requires pseudo free energies	[66]
Pfold	Utlizes a SCFG trained on rRNA alignments.	[67]
PETfold	Combines both the energy-based and evolution-based approaches	[15]
PhyloQFold	Consider the evolutionary history of a group of aligned RNA sequences	[68]
TurboFold	Utilizes probabilistic alignment as well as partition functions for mapping conserved pairs among sequences, and subsequently iterates the partition functions for improving the accuracy of structure prediction	[69]
Context Fold	Dependent on feature-rich trained scoring models.	https://www.cs.bgu.ac.il/~negevcb/contextfold/
E2Efold	A deep learning approach that employs a constrained optimization solver, without using dynamic programming.	[70]
SwiSpot	Detects alternative (secondary) configurations of riboswitches	[71]
Mfold	Prediction based on minimum free energy	[72]

RNAstructure [86], and RNAfold [87]) or machine learnings approaches (e.g., ContextFold [88] and CentroidFold [89]) are required to calculate the score parameters of every substructural device. But total accuracy (the percentage of correctly predicted basic pairs in all predicted base pairs) seems to have hit a "efficiency ceiling" [81] at around 80% [90, 91]. This is because all current approaches do not recognize some of all the base pairs arising from tertiary interactions [92]. These base-pairs are mostly pseudo-knotted (non-nested), lone (unstacked), and noncanonical base pairs (not G-U, A-U, and G-C) and triple interactions [92, 93]. While some methods can predict secondary RNA structures with pseudoknots (e.g., Knotty [94] and Probknot [95], pknotsRG [96]) and others

Table 10.2 Softwares and tools for predicting three-dimensional structure of RNA

Name	Description	References
BARNACLE	Employs probabilistic approach	[73]
FARNA	de novo prediction.	[74]
iFoldRNA	3D structure prediction as well as folding	[75]
MC-Fold MC-Sym Pipeline	Thermodynamics as well as nucleotide cyclic motifs for RNA structure prediction algorithm	[76]
ModeRNA	Based on a template RNA structure as well as a user-defined target-template sequence alignment	[77]
NAST	Coarse-grained modeling having knowledge-based potentials as well as structural filters	[78]
MMB	Turning limited experimental information into 3D models of RNA	[79]
RNA123	de novo and homology modeling of RNA 3D structures.	[80]
RNAComposer	Automated generation of large RNA 3D structures.	http://rnacomposer.cs.put.poznan.pl/

may predict noncanonical base pairs (e.g., CycleFold [97], MC-Fold-DP [98], and MC-Fold [76]).

10.3.1.1 Comparative Sequence Analysis

The most reliable way of the prediction of SSs for RNA is CSA as well as it is the method of first preference when deciding the SSs of a new RNA. It is built on the hypothesis that structure is more commonly conserved than sequences via evolution [58]. CSA was first used to address tRNA SSs [99, 100]. This research is used to be done with the assistance of modern structure prediction algorithms. Subsequently, tRNA crystal structures predicted was found to be right. Later, a CSA of the SSs of rRNAs have revealed that over 88% of the expected pairs find crystal structures subsequently fixed, and nearly all of the expected tertiary as well as noncanonical interactions were considered to be right [82]. Almost no means of SSs prediction can give something similar to this degree of precision, especially for longer RNAs, or have a similar insight into higher-order contacts that might also have functional or structural worth. CSA is typically the criterion under which structure prediction algorithms are tested since only a small number of sequences of such RNA families have been crystallized [58].

Identifying regions with orchestrated mutations that do not represent nucleotide identities but retain base pairs is a good indicator of an underlying structure that is conserved and of practical significance. The two-nucleotide sequence modifications that maintain base pairing are considered compensating base-pair changes. For example, the G-C base pair is more likely to mutate in one sequence into another canonical pair (AU, UA, CG, UG, GU) as the modification may include modifying two nucleotides rather than one, than mutating into one noncanonical pair or deleting one of the pairing partners with a single nucleotide. The bases of homologous RNAs from distant species can have low identification, but SSs are fully conserved, as each transition under which the sequences diverge conserves the structure [58]. Further accessible descriptions of structures of variance sequences resolved through CSA are accessible in the Rfam database seed alignments [101]. Irrespective of all these signs, CSA is not necessarily feasible, especially when a sequence is not defined. Free energy minimization (FEM) is one of the most common approaches in such scenarios [58].

10.3.1.2 Secondary RNA Structure Prediction Using Free Energy Minimization

The most common approach for predicting SSs is the FEM, where only a single sequence is defined for a certain function [102, 103]. This approach mainly employs DP, statistical mechanics, and pseudoknots algorithms to achieve its aim.

Dynamic Programming

The most common methods of FEM for RNA prediction are focused on dynamic algorithms of programming [102–104]. In principle, these algorithms can take into account indirectly all potential SSs for a particular sequence with the explicit construction of these structures. To do this, the lowest folding free energies are calculated for all sequence fragments of the entire sequence and the outcomes retained. As the least folding-FE (FE) for longer fragments is measured, the mechanism speeds up to the free energies for shorter fragments. DP algorithms have been preferred because they are computationally powerful and usually produce the same results to ensure that the lower FE structure is provided with the stability laws.

Statistical Mechanics

The lowest FE configuration is the most possible configuration for RNA in equilibrium. When the expected lowest FE structure is compared to the well-known secondary sequence, the precision may be defined by sensitivity or positive predictive value (PPV). Sensitivity is the proportion of recognized base pairs in the SSs predicted. The PPV is the proportion of expected pairs in the established structure. Therefore, sensitivity states that the proportion of identified pairs can be estimated independently of erroneously estimated pairs. Good predictive value is the percentage of expected, accurate pairs influenced by inaccurate pair predictions. It is usually less than sensitivity since FE reduction expects more base pairs than the so-called base changes. It is usually less than sensitivity since FE reduction expects more base pairs than the so-called base compensating base changes.

In 2003, Ding and Lawrence [105] established a statistical sampling procedure for RNA SSs prediction. The SSs are sampled according to Boltzmann's likelihood by means of a partition-function approximation using a stochastic DP algorithm. The likelihood of any given base pair is the frequency of its existence in the ensemble of structures within sampled structures. Moreover, several new structural properties can be calculated, including the likelihood of the single-stranded of two neighboring nucleotides. This information is not given by the partition function estimation alone in a single estimation since the base pairs' pairing chances are not independent. In other words, the possibility that two base pairs will appear in the same structure is not the consequence of its partition function. The predictive sampling approach improves the estimation of the SSs by identifying the SSs in the ensemble that better describes all of the structures [106]. This "centroid" configuration is selected as the least aggregate variance of all systems. The centroid of the ensemble is also not the lowest FE system. On average, centroids have slightly better sensitivity to base-pair predictions for different sequence databases of the established SSs but have a substantially higher predictive value. Thus, statistical sampling should also be used to boost the precision of SSs prediction.

Pseudoknots

Pseudoknots are troublesome since most DP algorithms cannot anticipate them, although 1.4% of simple pairs in various defined SSs are pseudo-knotted. The fundamental issue is that most DP algorithms presume, in order to speed up the estimation, that the overall folding FE change of a secondary system with two branches is the amount of the FE change of each branch calculated separately. If pseudoknots develop between the divisions, it does not work anymore. With increasing sequence length, DP algorithms that forecast pseudoknots scale poorly in time and are quite sluggish. For instance, the standard set of DP algorithms for FEM as well as partition function calculation scales $O(N^3)$ in time, whereby N is the number of nucleotides in the sequence. PKNOTS is a DP algorithm that can forecast the most known topology, yet $O(N^6)$ scales [107]. This means that doubling the sequence length takes 8 times more computing time by conventional methods but 64 times more when pseudoknots are taken into account. This restricts the use of these algorithms to sequences of up to 100–200 nucleotides. Numerous different DP algorithms scale better $O(N^5)$ or $O(N^4)$, but cannot forecast as many established pseudoknot topologies [108, 109]. Pknots RG, by Reeder and Giegerich [110], which scale $O(N^4)$ are one of the best to this group of algorithms.

Pseudoknots can be predicted using heuristics in acceptable computational time. However, the trade-off is that no certainty exists for estimating the lowest FE structure. In a software called ILM [111], one heuristic algorithm is introduced. It is based on a repeated (iterated) prediction of the structure with the so-called "loop matching algorithm." Each repetition forecasts a non-pseudoknotted structure, from which the highest score helix will be selected for the final structure. The paired nucleotides from the previously selected helixes are discarded in the next structure prediction iteration for each repeat. Since nucleotides are eliminated from successive measurements in pairs, the selected helices' collection may be pseudo-knotted in the final assembly. The algorithm scales $O(mN^3)$ in the worst case for m loop matching calculations. Another heuristic algorithm is introduced in the software HotKnots [112]. HotKnots commonly use many calls to a DP algorithm to assemble pseudoknots constructs, but at each point, several alternative helixes are expected simultaneously. This tests a number of SSs, which are ordered by increased FE changes at the end of the measurement.

10.3.1.3 Multiple-Sequence SSs Prediction

The multi-sequence SSs estimate tries to mimic a CSA by forecasting structures retained in two or more sequences. These techniques are not as reliable as manual CSA, but they can greatly increase precision over single sequence methods [63]. However, many of these techniques are more computationally costly than single sequence approaches.

Algorithms That Simultaneously Fold and Align

David Sankoff suggested the first method for folding homologous RNAs [113]. It concurrently considers the alignment as well as the folding of any amount of RNA sequences in a single measurement. The algorithm formally scales $O(N^{3s})$ in time, and $O(N^{2s})$ in storage, i.e., the memory usage, for s sequences of having length N. As other prediction methods of the single-sequence structure, this algorithm cannot predict pseudoknots. This algorithm is costly, particularly for more than two sequences. However, the limitation of alignment to eliminate impossible biological alignments and pairs provides major time-cost improvements. FOLDALIGN [114] was the first minimal version of the Sankoff algorithm. This algorithm used base-pair optimization instead of the FEM and nearest neighbor approach. It also removed branched structures from consideration, which reduced the algorithm to $O(N^4)$ in time, but removed a common and significant motif within RNA structure. Latest FOLDALIGN updates also provided support for the provision of branched systems, a FE model, and a heuristic trimming to speed up computation, which greatly increases algorithm accuracy [115].

Another approach that employs the Sankoff algorithm is LocARNA [116]. LocARNA maximizes the sum of pair probabilities for both sequences that are computed by different single-sequence partition function estimates and a similarity score for alignment, rather than minimizing energy. LocARNA runs easily since only significant base pairs are regarded, reducing the order of the algorithm to $O\ (N^2(N^2 + M^2))$, for two sequences of length N, and where M is the number of significant base pairs, also on order N. It does, however, lose any precision in contrast to FOLDALIGN [66]. While it is typically challenging to expand Sankoff's algorithm directly to more than two sequences in terms of computational costs, other means have been used to adapt it for the infinite number of sequences with greatly reduced complexity. FOLDALIGNM employed pair frequencies for all pair-wise FOLDALIGN sequence measurements [117].

Initially, LocARNA was also able to work on several sequences [116]. In order to generalize to several sequences, mLocARNA employs the output of the LocARNA multi-sequence alignment calculations in pairs. The chance of a pair of aligned columns is the square root of the pairing odds of the two alignments. RAF (99) is a distinct method that aligns and plies an infinite number of sequences concurrently. RAF functions in two sets at a time and aligns successive implementations of alignments rather than sets. mLocARNA and FOLDALIGNM seemed to work better at shorter sequences in benchmarks, and RAF & Multilign appeared to operate better for longer sequences, and both seemed to outperform single and double sequence approaches for most forms of RNAs [118]. Many of these algorithms cost the current hardware fairly.

Algorithms That Align First, Then Fold

The second paradigm to approximate the arrangement of more than one sequence is to first align and fold the sequences. This model is seen in RNAalifold [119]. RNAalifold defines a minimum FE consensus framework that a community of compatible input sequences may build. Input alignment mostly emerges from a series alignment algorithm; however, individually curated alignments are endorsed and may increase precision. RNAalifold is quick and productive for matching sequences. Its accuracy is hindered if the input is incorrectly matched, which may happen when the sequence identity of the pairs is <60%. In those instances, automatic sequence alignment algorithms battle [120]. The alignment of sequences of low identification, rendered by a professional investigator, or at least modified, could have provided good results. The CentroidFold algorithm [121] also exists in the model "align, then fold." It investigates the central frame in a way that is similar to the one of Sfold [122] rather than considering a minimum free-energy consensus structure. The central structure of the largest structural cluster represents the central structure, created by the stochastic sampling of homologous structures series. The chances of identifying the core consensus structure can be calculated by using an experimental discovery [123] or by using the nearest model from a database of unique structure sequences.

Recently, the TurboFold model requires an unlimited amount of sequences in the "align, then fold" model and then tests their pair-wise probabilistic alignment and their base-pair probabilities [124]. These alignments are employed for configurations among sequences. The single-sequence base pair probabilities for a provided sequence within the collection are referred to as the "intrinsic information." For any other sequence, the combined probabilistic aligning and the base couple probabilities are referred to as the "extrinsic information." The updated probabilities of the pair would then be used to recalculate extrinsic information. Many iterations strengthen and improve the predicted chances of the pair for series. The architectures are designed with the highest expected accuracy algorithm after the required number of iterations. In random assortments, TurboFold typically outperforms other predictive algorithms in multiple sequences that normally include identities that are less than 60% in pairs and that are commonly comparable in PPV sequences [124]. Although it may be more expensive to compute than any of the above alignment and then fold algorithms, up to 10 RNA sequences of normal lengths per minute are required. One of TurboFold's important advantages over most of the above-noted algorithms is that it does not enforce a common structure. Variable elements can also be properly predicted in homologous sequences, including the variable stem in tRNAs, allowing TurboFold a convincing alternative for structural prediction, where different sequences can be used in divergent identities and ambiguous alignment.

An Algorithm That Folds, Then Aligns

In multi-sequence structure prediction, the third paradigm is to “fold, then align.” RNAshapes [125] follows this method. This algorithm lists separately the abstract “shape” space accessible for each sequence and determines the probability of each form, and then defines the thermodynamically optimal configuration with the typical form. Instead of full pairing details, abstract shapes encode RNA structure features. There are far less low FE sources than systems, so the solution is feasible. RNAshapes is fast; after single sequence structure measurements, it is roughly linear in time. It gives precision comparable to the above multi-sequence approaches. It does not provide series alignments but can be created with RNAforester [126] from the retained structure.

10.3.2 Three-Dimensional Structure Prediction Methods

Although the SSs offers the blueprint for an RNA molecule, information about the RNA 3D structure remains key to an overall understanding of its role. Initial 3D structure modeling was carried out successfully by RNA structure experts with the 3D structures of several typical RNA molecules, like tRNAs [100], the group I introns [127], and RNase P [128]. In recent years, a range of computational models for the prediction of RNA 3D structures has been developed [1]. These models may be broadly categorized into two groups, i.e., depending on the knowledge or physical property.

10.3.2.1 Knowledge-Dependent Modeling

RNA 3D structures may be predicted by assembling established motifs or the aligning sequence with already available experimentally defined structures in the database. Knowledge-based modeling primarily involves modeling on the basis of graphics and homology-based modeling (HBM) [129].

Graphics-Based Methods

The graphics modeling typically offers a visual interface, which enables users to create 3D RNA constructs by controlling or assembling segments of RNA [130–134]. Few of the major graphics-based algorithms are MANIP, ERNA-3D, and RNA2D3D. The MANIP helps users to design known 3D models on the computer screen using the corresponding SSs predicted through CSA [130]. Although the MANIP is not an automated process, it provides a quick and simple way to construct 3D RNA structures, particularly large RNAs, such as the RNase P RNA [130]. Moreover, multiple relationship tables, as well as base-pair tables that specifically contain RNAs’ topological information, can be used to precisely model RNAs’ interactions [130].

In order to create RNA 3D structures from sequences as well as SSs, the ERNA-3D offers users a graphic interface in order to freely position the A-form helixes and to explicitly draw the single inter-helical strands [135]. The 3D structures of mRNA, rRNAs, and tRNAs, including 16S rRNAs, 5S rRNA, and 23S rRNA, were successfully produced using ERNA-3D [135]. The RNA2D3D will forecast rough 3D structures for large RNAs easily based on their SSs, e.g., viral kissing loops, ribozymes, and various RNA nanostructures [136]. Manual handling, though, must be performed to create a graphical interface to achieve a better structure, like compacting, energy refinement, stalking, and segment-positioning [137]. While the graphics-based approaches introduced above can be used for creating 3D structures for large RNAs with hundreds of nucleotides quickly and intuitively, since they are manual techniques, they require users to set up and optimize the RNA structure models according to particular concepts utilizing the tools provided in the software packages. Thus, in order to construct plausible systems, it is important for users to have extensive knowledge of RNA systems.

Homology-Based Modeling

Although a macromolecule's 3D structure experiences change much slower in comparison to its sequence, evolutionarily associated macromolecules normally preserve similar 3D structure though divergences at the sequence level. On this basis, 3D macromolecule structures are able to be built by aligning the target molecule sequence to molecular structure templates [134]. HBM, also referred to as comparative modeling or template-based modeling, was very effective in the prediction of 3D protein structure [138, 139]. Additionally, HBM has been expanded to include fragment assembly methods like 3dRNA [140] and RNAComposer [141]. 3dRNA is a quick and automatic 3D algorithm designed to construct RNA structure by assembling A-form helixes and various loops, whose structures are extracted in a database from known structures [140]. 3dRNA predicts reliable 3D structures based on its SSs for 300 RNA tested, including pseudoknots, duplexes, and hairpins. In addition, 3dRNA can also be used freely online as a database server, and the projected 3D structure can be accessed rapidly with the sequence and SSs as data [140].

ModeRNA enables both the simplified structure forecast from a series of templates/alignments as well as user-controlled structure manipulations, i.e., the fragment assembly [142]. ModeRNA understands as well as models post-transcriptional alteration of nucleosides compared with other modeling algorithms. It is pertinent to note that even though ModeRNA is not a method focused on graphics, it also demands that users should have alignment among the RNAs template and the target RNA and define the base pairs between the embedded fragment and the rest of the RNA [142]. RNAComposer is another web server that can use the RNA FRABASE database for predicting 3D structures for large RNAs [141]. The RNA FRABASE database can be considered as a dictionary linked to the RNA SSs with established fragments of a tertiary structure. The SSs that a user provides in the RNAComposer is, first of all, broken into elements such as stem, loops, and individual strands and subsequently scanned the related tertiary structural elements automatically from an RNA FRABASE database as well as assembled into full 3D structures.

The key benefit of HBM is that the size of the RNAs to be modeled is not necessarily limited. The consistency of the projected structures depends, however, on the sequence alignment consistency, template structures, and secondary frameworks identified by the user. Although the amount of identified RNA structures stored, the PDB/NDB database is growing quickly, and it might yet be challenging to locate accurate template RNAs for a given target RNA. In addition, owing to their strong stability, the configurations of their RNA are normally modified with solutions such as ion conditions and temperature [143], and other ligands or macromolecules. Moreover, the creation of a good alignment of RNAs with complicated systems typically involves laborious manual planning dependent on proven expertise in the most significant RNA families. HBM is, therefore, not always accurate.

10.3.2.2 Physics-Based Modeling

Physics-related methods are based on biophysical concepts that concurrently scan for the conformation to fold with minimum free energy. Since complete atomic structure modeling for an RNA typically requires several degrees of freedom and thus tremendous computational sophistication, many CG predictive models with physical simplifications have also been developed at various resolution levels.

All-Atomistic Model

Until today, the "all-atomistic molecular dynamics" are highly required for understanding macromolecule simulation, which in turn provides an insight into the real movement of atoms, such as AMBER [140, 144] and CHARMM [145] with physics-based force fields. However, considering the several degrees of freedom, it remains challenging for folding RNA 3D structures even with advanced computing strategies. The models were then evolved considering the recognized or secondary fragments [146], such as the MC-fold/Mc-Sym Pipeline (MMP) [76] and FARNA/FARFAR [147]. Because SSs can provide enough structural constraints for automated construction of 3D structures, the MMP infers RNA SSs from sequence data and subsequently assembles a set of 3D structures based on their SSs [76]. Unlike the thermodynamics approaches such as Mfold [148]. The MC-Fold can forecast RNA SSs, including noncanonical and canonical base pairs, for the usage of a knowledge-related scoring function associated with the NCM (nucleotide cycle motif) databases. The NCMs that are circularly bound through covalent bonds, pairing or stacking interactions, were actually developed from a study of the X-ray crystallographic structures. The MC-Sym along with the 3D NCMs as well as the Las Vegas algorithm was employed for the fragment insertion simulation. MMP has been authenticated by constructing 3D structures of precursor microRNA as well as human immunodeficiency virus (HIV1) *cis*-acting-1 frame-shifting segment [76].

Das and Baker discussed FARNA's completely automatic, energy-based solution to predict RNA 3D structure [74]. FARNA integrates trinucleotide fragments obtained from the ribosome crystal structure into a completely atomistic structure that is compatible with the particular sequence by utilizing the Monte Carlo algorithm as well as the simpler knowledge-based energy feature that favors stacking, base pairing, and geometry. The CG core pairing capacity employed in FARNA is focused on the mathematical study of the ribosomal basis, and not only Watson–

Crick base pairs. However, the interactions along with Hoogsteen as well as sugar edges may be taken into account. FARFAR implements a high-resolution process of refining into FARNA in order to forecast and design the atomic precision of noncanonical small RNA structures [147]. In another study, RNAnbds was built for predicting RNA's 3D structures through fragmentary assembly, on the basis of statistics of bases as per their sequence/space neighbors in databases [149]. RNAnbds offers a good predictor for short fragments (<15 nucleotides), in specific RMSD loops <4, together with statistical potentials like base stacking and base pairing.

Coarse-Grained Model

Another important approach for minimizing the computational expense is to decrease the number of objects by handling a set of usable atoms with a single bead [150, 151]. The bead can either denote a few or a large number of atoms on the basis of the model's resolution. Following the initially "one-bead RNA model" designed by Malhotra and Harvey [152], several CG models were implemented for the purposes of predicting RNA 3D structures [153] or modeling interactions among RNAs as well as other molecules [154, 155], for example, NAST [156], YUP [157], and iFold [75]. The YUP is a very versatile molecular mechanic algorithm for CG and multi-scaling modeling [157]. The YUP is employed for modeling RNA, protein, and DNA structures on the basis of the related energy potentials and approaches such as Monte Carlo, energy minimization, and molecular dynamics. In YUP, one nucleotide is substituted by a pseudo-atom at the middle of phosphorous atoms in order to model high RNAs, which decreases device costs efficiently. While YUP needs users to supply the information about the SSs of RNAs and the force field, YUP is an adaptive RNA modeling kit for automated CG modeling [157].

Like YUP [157], NAST is another "one-bead RNA model" in which a nucleotide's C3' atom is picked to reflect the whole nucleotide [158]. The NAST will sample conformations that fulfill a certain range of secondary structure as well as tertiary interaction limit, with an RNA basic knowledge-based ability and a simple molecular dynamic algorithm. One benefit of NAST is its capacity to integrate experimental data as a filter for structurally equivalent conformation clusters, for example, with perfect small-angle X-ray dispersing data as well as experimental solvent accessibility data. Earlier, NAST was employed for predicting the yeast phenylalanine tRNA's 3D structures (76 nucleotides) and the Tetrahymena thermophile group I intron's the P4-P6 domain of 158 nucleotides within 8 Å as well as 16 Å RMSDs retrieved from experimental structures, respectively [158]. The iFold is another web-based algorithm that is being built by the Doknolyan community and can be used for predicting RNA's 3D structures [75]. The model employs three-nucleotide beads of CG representation and efficient molecular dynamic simulations with step-by-step potentials like the base pairing. iFold's strength has been seen in forecasting the 3D structures of 150 RNA with different sequences, with <4 Å deviations in experimental structures in the majority of predicted structures [159].

10.4 Conclusion and Future Perspective

In conclusion, to date, numerous computational approaches have been developed for predicting the secondary as well as the three-dimensional structure of RNA. However, the authors believe that there is still scope for the development of novel tools and techniques which can predict a more accurate structure [160]. Combining experimental and computational approaches for predicting the structure of RNA will enable us to understand its structure and function more precisely. For instance, smFRET as well as NMR spectroscopy are useful tools to evaluate the several states that the RNA can embrace. The PARIS approach for in vivo crosslinking also has a significant potential to include several instances of multiple-folded RNA. Improvements in cryoelectron microscopy, as well as tomography along with direct electron detectors, advanced contrast methods as well as single particle detection, can allow direct observation of single RNA molecules feasible in several configurations. The cryoelectron microscopy group has already designed novel sophisticated systems for categorizing related configurations of complex macromolecular assemblies. The least number as well as the length of RNA helices obtained from crystallography or cryoelectron microscopy often offer an important restriction for RNA folding [160]. Good measurements are also needed to assess differences among RNA structures. For instance, the analysisDist tool available within the program kit of the Vienna RNA provides several alternatives for measuring matrix distances with Ward's method, Shapiro's cost matrix for coarse structures, or Saitou's neighbor-joining method [87]. Thus, RNA structure prediction continues to progress with novel metrics as well as more experimentally specified examples of multi-conformation RNA assemblies. A single minimum free energy configuration for the RNA sequence would become a greater appreciation of the several potential different configurations encrypted in the RNA sequence and precise forecasts will direct the evaluation of detailed transcript and transcriptome-wide research studies.

Conflict of Interest None

Additional Information Figure 10.1 (CC BY 4.0) [5, 6] has been used under the terms of the Creative Commons Attribution License.

References

1. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. PNAS. 2013;110(14):5498–503.
2. Vandivier LE, Anderson SJ, Foley SW, Gregory BD. The conservation and function of RNA secondary structure in plants. Annu Rev Plant Biol. 2016;67:463–88.
3. Kashi K, Henderson L, Bonetti A, Carninci P. Discovery and functional analysis of lncRNAs: methodologies to investigate an uncharacterized transcriptome. Biochim Biophys Acta. 2016;1859(1):3–15.
4. Cech TR, Steitz JA. The noncoding RNA revolution-trashing old rules to forge new ones. Cell. 2014;157(1):77–94.

5. Achar A, Sætrom P. RNA motif discovery: a computational overview. Biol Direct [Internet]. 2015 [cited 2020 Dec 15];10. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4600295/.
6. Lim CS, Brown CM. Know your enemy: successful bioinformatic approaches to predict functional RNA structures in viral RNAs. Front Microbiol [Internet]. 2018 [cited 2020 Dec 15];8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02582/full#h10.
7. Li H, Zhu D, Zhang C, Han H, Crandall KA. Characteristics and prediction of RNA structure, vol. 2014 [Internet]. BioMed Research International. Hindawi; 2014 [cited 2020 Oct 25]. p. e690340. Available from: https://www.hindawi.com/journals/bmri/2014/690340/.
8. Reyes FE, Garst AD, Batey RT. Chapter 6—Strategies in RNA crystallography. In: Methods in enzymology [Internet]. Biophysical, chemical, and functional probes of RNA structure, interactions and folding: part B; vol. 469. Academic Press; 2009 [cited 2020 Oct 25]. p. 119–39. Available from: http://www.sciencedirect.com/science/article/pii/S0076687909690066.
9. Westhof E. Twenty years of RNA crystallography. RNA. 2015;21(4):486–7.
10. Fernandez-Leiro R, Scheres SHW. Unravelling the structures of biological macromolecules by cryo-EM. Nature. 2016;537(7620):339–46.
11. Fürtig B, Richter C, Wöhnert J, Schwalbe H. NMR spectroscopy of RNA. Chembiochem. 2003;4(10):936–62.
12. Miao Z, Adamiak RW, Blanchet M-F, Boniecki M, Bujnicki JM, Chen S-J, et al. RNA-puzzles round II: assessment of RNA structure prediction programs applied to three large RNA structures. RNA. 2015;21(6):1066–84.
13. Magnus M, Kappel K, Das R, Bujnicki JM. RNA 3D structure prediction guided by independent folding of homologous sequences. BMC Bioinform [Internet]. 2019 [cited 2020 Oct 25];20. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6806525/.
14. Puton T, Kozlowski LP, Rother KM, Bujnicki JM. CompaRNA: a server for continuous benchmarking of automated methods for RNA secondary structure prediction. Nucleic Acids Res. 2013;41(7):4307–23.
15. Seemann SE, Gorodkin J, Backofen R. Unifying evolutionary and thermodynamic information for RNA folding of multiple alignments. Nucleic Acids Res. 2008;36(20):6355–62.
16. Weinreb C, Riesselman AJ, Ingraham JB, Gross T, Sander C, Marks DS. 3D RNA and functional interactions from evolutionary couplings. Cell. 2016;165(4):963–75.
17. Bonneau R, Strauss CEM, Baker D. Improving the performance of rosetta using multiple sequence alignment information and global measures of hydrophobic core formation. Proteins. 2001;43(1):1–11.
18. Sun XS. 2—Plant materials formation and growth. In: Wool RP, Sun XS, editors. Bio-based polymers and composites [Internet]. Burlington: Academic Press; 2005 [cited 2020 Oct 25]. p. 15–32. Available from: http://www.sciencedirect.com/science/article/pii/B9780127639529500034.
19. Sharma D, Singh S, Chand T, Kumar P. RNA: structure, prediction, and visualization tools. In: Intelligent communication, control and devices. New York: Springer; 2018. p. 335–45.
20. Feher J. 2.2—DNA and protein synthesis. In: Feher J, editor. Quantitative human physiology. 2nd ed [Internet]. Boston: Academic Press; 2017 [cited 2020 Oct 25]. p. 120–9. Available from: http://www.sciencedirect.com/science/article/pii/B9780128008836000112.
21. Goss DJ, Domashevskiy AV. Messenger RNA (mRNA): the link between DNA and protein. In: Bradshaw RA, Stahl PD, editors. Encyclopedia of cell biology [Internet]. Waltham: Academic Press; 2016 [cited 2020 Oct 25]. p. 341–5. Available from: http://www.sciencedirect.com/science/article/pii/B9780123944474100409.
22. Dunckley T, Parker R. RNA turnover. In: Brenner S, Miller JH, editors. Encyclopedia of genetics [Internet]. New York: Academic Press; 2001 [cited 2020 Oct 25]. p. 1748–51. Available from: http://www.sciencedirect.com/science/article/pii/B0122270800011381.

23. Nazar RN. Ribosomal RNA processing and ribosome biogenesis in eukaryotes. IUBMB Life. 2004;56(8):457–65.
24. Hang R, Wang Z, Deng X, Liu C, Yan B, Yang C, et al. Ribosomal RNA biogenesis and its response to chilling stress in Oryza sativa. Plant Physiol. 2018;177(1):381–97.
25. Cooper GM. The cell: a molecular approach. Washington, DC/Sunderland, MA: ASM Press/ Sinauer Associates; 2000.
26. Doherty J, Guo M. Transfer RNA. In: Bradshaw RA, Stahl PD, editors. Encyclopedia of cell biology [Internet]. Waltham: Academic Press; 2016 [cited 2020 Oct 25]. p. 309–40. Available from: http://www.sciencedirect.com/science/article/pii/B9780123944474100392.
27. O'Donoghue P, Ling J, Söll D. Transfer RNA function and evolution. RNA Biol. 2018;15 (4–5):423–6.
28. Fields RN, Roy H. Deciphering the tRNA-dependent lipid aminoacylation systems in bacteria: novel components and structural advances. RNA Biol. 2018;15(4–5):480–91.
29. Alamos P, Tello M, Bustamante P, Gutiérrez F, Shmaryahu A, Maldonado J, et al. Functionality of tRNAs encoded in a mobile genetic element from an acidophilic bacterium. RNA Biol. 2018;15(4–5):518–27.
30. Castillo A, Tello M, Ringwald K, Acuña LG, Quatrini R, Orellana O. A DNA segment encoding the anticodon stem/loop of tRNA determines the specific recombination of integrative-conjugative elements in Acidithiobacillus species. RNA Biol. 2018;15 (4–5):492–9.
31. Bacusmo JM, Orsini SS, Hu J, DeMott M, Thiaville PC, Elfarash A, et al. The t6A modification acts as a positive determinant for the anticodon nuclease PrrC, and is distinctively nonessential in Streptococcus mutans. RNA Biol. 2018;15(4–5):508–17.
32. Kessler AC, Kulkarni SS, Paulines MJ, Rubio MAT, Limbach PA, Paris Z, et al. Retrograde nuclear transport from the cytoplasm is required for tRNATyr maturation in T. brucei. RNA Biol. 2018;15(4–5):528–36.
33. Agris PF, Eruysal ER, Narendran A, Väre VYP, Vangaveti S, Ranganathan SV. Celebrating wobble decoding: half a century and still much is new. RNA Biol. 2018;15(4–5):537–53.
34. Rafels-Ybern À, Torres AG, Grau-Bove X, Ruiz-Trillo I, de Pouplana LR. Codon adaptation to tRNAs with inosine modification at position 34 is widespread among eukaryotes and present in two bacterial phyla. RNA Biol. 2018;15(4–5):500–7.
35. Kessler AC, d'Almeida GS, Alfonzo JD. The role of intracellular compartmentalization on tRNA processing and modification. RNA Biol. 2018;15(4–5):554–66.
36. Marz M, Gruber AR, Zu Siederdissen CH, Amman F, Badelt S, Bartschat S, et al. Animal snoRNAs and scaRNAs with exceptional structures. RNA Biol. 2011;8(6):938–46.
37. Jorjani H, Kehr S, Jedlinski DJ, Gumienny R, Hertel J, Stadler PF, et al. An updated human snoRNAome. Nucleic Acids Res. 2016;44(11):5068–82.
38. Decatur WA, Fournier MJ. rRNA modifications and ribosome function. Trends Biochem Sci. 2002;27(7):344–51.
39. Kiss T. New embo member's review. EMBO J. 2001;20(14):3617–22.
40. McKeegan KS, Debieux CM, Boulon S, Bertrand E, Watkins NJ. A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly. Mol Cell Biol. 2007;27 (19):6782–93.
41. Ganot P, Caizergues-Ferrer M, Kiss T. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev. 1997;11(7):941–56.
42. Lafontaine DL, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D. The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. Genes Dev. 1998;12(4):527–37.
43. Ganot P, Bortolin ML, Kiss T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. Cell. 1997;89(5):799–809.

44. Bortolin ML, Ganot P, Kiss T. Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. EMBO J. 1999;18(2):457–69.
45. O'Brien J, Hayder H, Zayed Y, Peng C. Overview of microRNA biogenesis, mechanisms of actions, and circulation. Front Endocrinol [Internet]. 2018 [cited 2020 Oct 25];9. Available from: https://www.frontiersin.org/articles/10.3389/fendo.2018.00402/full.
46. Fu G, Brkić J, Hayder H, Peng C. MicroRNAs in human placental development and pregnancy complications. Int J Mol Sci. 2013;14(3):5519–44.
47. Tüfekci KU, Öner MG, Meuwissen RLJ, Genç Ş. The role of microRNAs in human diseases. In: Yousef M, Allmer J, editors. miRNomics: microRNA biology and computational analysis [Internet]. Methods in molecular biology. Totowa, NJ: Humana Press; 2014 [cited 2020 Oct 25]. p. 33–50. Available from: https://doi.org/10.1007/978-1-62703-748-8_3.
48. Paul P, Chakraborty A, Sarkar D, Langthasa M, Rahman M, Bari M, et al. Interplay between miRNAs and human diseases. J Cell Physiol. 2018;233(3):2007–18.
49. Huang W. MicroRNAs: biomarkers, diagnostics, and therapeutics. In: Huang J, Borchert GM, Dou D, Huan J (Luke), Lan W, Tan M, et al., editors. Bioinformatics in microRNA research [Internet]. Methods in molecular biology. New York, NY: Springer; 2017 [cited 2020 Oct 25]. p. 57–67. Available from: https://doi.org/10.1007/978-1-4939-7046-9_4.
50. Hari R, Parthasarathy S. Prediction of coding and non-coding RNA. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology [Internet]. Oxford: Academic Press; 2019 [cited 2020 Oct 25]. p. 230–40. Available from: http://www.sciencedirect.com/science/article/pii/B978012809633820099X.
51. Fang Y, Fullwood MJ. Roles, functions, and mechanisms of long non-coding RNAs in cancer. Genomics Proteomics Bioinform. 2016;14(1):42–54.
52. Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, et al. The evolution of lncRNA repertoires and expression patterns in tetrapods. Nature. 2014;505(7485):635–40.
53. Kutter C, Watt S, Stefflova K, Wilson MD, Goncalves A, Ponting CP, et al. Rapid turnover of long noncoding RNAs and the evolution of gene expression. PLoS Genet. 2012;8(7): e1002841.
54. Ulitsky I, Shkumatava A, Jan CH, Sive H, Bartel DP. Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. Cell. 2011;147 (7):1537–50.
55. Lin N, Chang K-Y, Li Z, Gates K, Rana ZA, Dang J, et al. An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. Mol Cell. 2014;53(6):1005–19.
56. Brockdorff N, Ashworth A, Kay GF, McCabe VM, Norris DP, Cooper PJ, et al. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell. 1992;71(3):515–26.
57. Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, et al. The transcriptional landscape of the mammalian genome. Science. 2005;309(5740):1559–63.
58. Seetin MG, Mathews DH. RNA structure prediction: an overview of methods. In: Keiler KC, editor. Bacterial regulatory RNA: methods and protocols [Internet]. Methods in molecular biology. Totowa, NJ: Humana Press; 2012 [cited 2020 Oct 25]. p. 99–122. Available from: https://doi.org/10.1007/978-1-61779-949-5_8.
59. Waldispühl J, Reinharz V. Modeling and predicting RNA three-dimensional structures. In: Picardi E, editor. RNA bioinformatics [Internet]. Methods in molecular biology. New York, NY: Springer; 2015 [cited 2020 Oct 25]. p. 101–21. Available from: https://doi.org/10.1007/978-1-4939-2291-8_6.
60. Hamada M, Sato K, Asai K. Improving the accuracy of predicting secondary structure for aligned RNA sequences. Nucleic Acids Res. 2011;39(2):393–402.
61. Sato K, Kato Y, Akutsu T, Asai K, Sakakibara Y. DAFS: simultaneous aligning and folding of RNA sequences via dual decomposition. Bioinformatics. 2012;28(24):3218–24.

62. Lindgreen S, Gardner PP, Krogh A. Measuring covariation in RNA alignments: physical realism improves information measures. Bioinformatics. 2006;22(24):2988–95.
63. Xu Z, Mathews DH. Multilign: an algorithm to predict secondary structures conserved in multiple RNA sequences. Bioinformatics. 2011;27(5):626–32.
64. Kiryu H, Tabei Y, Kin T, Asai K. Murlet: a practical multiple alignment tool for structural RNA sequences. Bioinformatics. 2007;23(13):1588–98.
65. Tabei Y, Kiryu H, Kin T, Asai K. A fast structural multiple alignment method for long RNA sequences. BMC Bioinform. 2008;9:33.
66. Harmanci AO, Sharma G, Mathews DH. PARTS: probabilistic alignment for RNA joinT secondary structure prediction. Nucleic Acids Res. 2008;36(7):2406–17.
67. Knudsen B, Hein J. Pfold: RNA secondary structure prediction using stochastic context-free grammars. Nucleic Acids Res. 2003;31(13):3423–8.
68. Doose G, Metzler D. Bayesian sampling of evolutionarily conserved RNA secondary structures with pseudoknots. Bioinformatics. 2012;28(17):2242–8.
69. Seetin MG, Mathews DH. TurboKnot: rapid prediction of conserved RNA secondary structures including pseudoknots. Bioinformatics. 2012;28(6):792–8.
70. Chen X, Li Y, Umarov R, Gao X, Song L. RNA secondary structure prediction by learning unrolled algorithms. In 2019 [cited 2020 Dec 16]. Available from: https://openreview.net/forum?id=S1eALyrYDH.
71. Barsacchi M, Novoa EM, Kellis M, Bechini A. SwiSpot: modeling riboswitches by spotting out switching sequences. Bioinformatics. 2016;32(21):3252–9.
72. Zuker M, Stiegler P. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. Nucleic Acids Res. 1981;9(1):133–48.
73. Frellsen J, Moltke I, Thiim M, Mardia KV, Ferkinghoff-Borg J, Hamelryck T. A probabilistic model of RNA conformational space. PLoS Comput Biol. 2009;5(6):e1000406.
74. Das R, Baker D. Automated de novo prediction of native-like RNA tertiary structures. PNAS. 2007;104(37):14664–9.
75. Sharma S, Ding F, Dokholyan NV. iFoldRNA: three-dimensional RNA structure prediction and folding. Bioinformatics. 2008;24(17):1951–2.
76. Parisien M, Major F. The MC-Fold and MC-Sym pipeline infers RNA structure from sequence data. Nature. 2008;452(7183):51–5.
77. Rother M, Milanowska K, Puton T, Jeleniewicz J, Rother K, Bujnicki JM. ModeRNA server: an online tool for modeling RNA 3D structures. Bioinformatics. 2011;27(17):2441–2.
78. Jonikas MA, Radmer RJ, Laederach A, Das R, Pearlman S, Herschlag D, et al. Coarse-grained modeling of large RNA molecules with knowledge-based potentials and structural filters. RNA. 2009;15(2):189–99.
79. Flores SC, Altman RB. Turning limited experimental information into 3D models of RNA. RNA. 2010;16(9):1769–78.
80. Eriksson ESE, Joshi L, Billeter M, Eriksson LA. De novo tertiary structure prediction using RNA123—benchmarking and application to Macugen. J Mol Model. 2014;20(8):2389.
81. Rivas E. The four ingredients of single-sequence RNA secondary structure prediction. A unifying perspective. RNA Biol. 2013;10(7):1185–96.
82. Gutell RR, Lee JC, Cannone JJ. The accuracy of ribosomal RNA comparative structure models. Curr Opin Struct Biol. 2002;12(3):301–10.
83. Domer JE, Ichinose H. Cellular immune responses in guinea pigs immunized with cell walls of Histoplasma capsulatum prepared by several different procedures. Infect Immun. 1977;16(1):293–301.
84. Schroeder SJ, Turner DH. Optical melting measurements of nucleic acid thermodynamics. Methods Enzymol. 2009;468:371–87.
85. Janssen S, Giegerich R. The RNA shapes studio. Bioinformatics. 2015;31(3):423–5.
86. Reuter JS, Mathews DH. RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinform. 2010;11(1):129.

87. Lorenz R, Bernhart SH, Höner zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA Package 2.0. Algorithms Mol Biol. 2011;6(1):26.
88. Zakov S, Goldberg Y, Elhadad M, Ziv-ukelson M. Rich parameterization improves RNA structure prediction. J Comput Biol. 2011;18(11):1525–42.
89. Sato K, Hamada M, Asai K, Mituyama T. CentroidFold: a web server for RNA secondary structure prediction. Nucleic Acids Res. 2009;37(Web Server issue):W277–80.
90. Do CB, Woods DA, Batzoglou S. CONTRAfold: RNA secondary structure prediction without physics-based models. Bioinformatics. 2006;22(14):e90–8.
91. Xu X, Chen S-J. Physics-based RNA structure prediction. Biophys Rep. 2015;1(1):2–13.
92. Nowakowski J, Tinoco I. RNA structure and stability. Semin Virol. 1997;8(3):153–65.
93. Westhof E, Fritsch V. RNA folding: beyond Watson–Crick pairs. Structure. 2000;8(3): R55–65.
94. Jabbari H, Wark I, Montemagno C, Will S. Knotty: efficient and accurate prediction of complex RNA pseudoknot structures. Bioinformatics. 2018;34(22):3849–56.
95. Bellaousov S, Mathews DH. ProbKnot: fast prediction of RNA secondary structure including pseudoknots. RNA. 2010;16(10):1870–80.
96. Reeder J, Giegerich R. Design, implementation and evaluation of a practical pseudoknot folding algorithm based on thermodynamics. BMC Bioinform. 2004;5(1):104.
97. Sloma MF, Mathews DH. Base pair probability estimates improve the prediction accuracy of RNA non-canonical base pairs. PLoS Comput Biol. 2017;13(11):e1005827.
98. zu Siederdissen CH, Bernhart SH, Stadler PF, Hofacker IL. A folding algorithm for extended RNA secondary structures. Bioinformatics. 2011;27(13):i129–36.
99. Madison JT, Everett GA, Kung H. Nucleotide sequence of a yeast tyrosine transfer RNA. Science. 1966;153(3735):531–4.
100. Levitt M. Detailed molecular model for transfer ribonucleic acid. Nature. 1969;224 (5221):759–63.
101. Gardner PP, Daub J, Tate J, Moore BL, Osuch IH, Griffiths-Jones S, et al. Rfam: Wikipedia, clans and the “decimal” release. Nucleic Acids Res. 2011;39(Database issue):D141–5.
102. Baxevanis AD, Ouellette BFF, editors. Bioinformatics: a practical guide to the analysis of genes and proteins. Hoboken, NJ: Wiley-Interscience; 2004. 560 p
103. Mathews DH, Turner DH, Watson RM. RNA secondary structure prediction. Curr Protoc Nucleic Acid Chem. 2007;CHAPTER 11:Unit-11.2.
104. Eddy SR. How do RNA folding algorithms work? Nat Biotechnol. 2004;22(11):1457–8.
105. Ding Y, Lawrence CE. A statistical sampling algorithm for RNA secondary structure prediction. Nucleic Acids Res. 2003;31(24):7280–301.
106. Ding Y, Chan CY, Lawrence CE. RNA secondary structure prediction by centroids in a Boltzmann weighted ensemble. RNA. 2005;11(8):1157–66.
107. Mathews DH, Disney MD, Childs JL, Schroeder SJ, Zuker M, Turner DH. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. PNAS. 2004;101(19):7287–92.
108. Condon A, Davy B, Rastegari B, Zhao S, Tarrant F. Classifying RNA pseudoknotted structures. Theor Comput Sci. 2004;320(1):35–50.
109. Dirks RM, Pierce NA. An algorithm for computing nucleic acid base-pairing probabilities including pseudoknots. J Comput Chem. 2004;25(10):1295–304.
110. Stoddard CD, Montange RK, Hennelly SP, Rambo RP, Sanbonmatsu KY, Batey RT. Free state conformational sampling of the SAM-I riboswitch aptamer domain. Structure. 2010;18 (7):787–97.
111. Ruan J, Stormo GD, Zhang W. An iterated loop matching approach to the prediction of RNA secondary structures with pseudoknots. Bioinformatics. 2004;20(1):58–66.
112. Ren J, Rastegari B, Condon A, Hoos HH. HotKnots: heuristic prediction of RNA secondary structures including pseudoknots. RNA. 2005;11(10):1494–504.
113. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math. 1985;45(5):810–25.

114. Gorodkin J, Heyer LJ, Stormo GD. Finding the most significant common sequence and structure motifs in a set of RNA sequences. Nucleic Acids Res. 1997;25(18):3724–32.
115. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol. 2007;3(10):e193.
116. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 2007;3 (4):e65.
117. Torarinsson E, Havgaard JH, Gorodkin J. Multiple structural alignment and clustering of RNA sequences. Bioinformatics. 2007;23(8):926–32.
118. Do CB, Foo C-S, Batzoglou S. A max-margin model for efficient simultaneous alignment and folding of RNA sequences. Bioinformatics. 2008;24(13):i68–76.
119. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC Bioinform. 2008;9(1):474.
120. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res. 2005;33(8):2433–9.
121. Hamada M, Kiryu H, Sato K, Mituyama T, Asai K. Prediction of RNA secondary structure using generalized centroid estimators. Bioinformatics. 2009;25(4):465–73.
122. Ding Y, Lawrence CE. Statistical prediction of single-stranded regions in RNA secondary structure and application to predicting effective antisense target sites and beyond. Nucleic Acids Res. 2001;29(5):1034–46.
123. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29(6–7):1105–19.
124. Harmanci AO, Sharma G, Mathews DH. TurboFold: iterative probabilistic estimation of secondary structures for multiple RNA sequences. BMC Bioinform. 2011;12(1):108.
125. Steffen P, Voss B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics. 2006;22(4):500–3.
126. Höchsmann M, Voss B, Giegerich R. Pure multiple RNA secondary structure alignments: a progressive profile approach. IEEE/ACM Trans Comput Biol Bioinform. 2004;1(1):53–62.
127. Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol. 1990;216(3):585–610.
128. Harris ME, Nolan JM, Malhotra A, Brown JW, Harvey SC, Pace NR. Use of photoaffinity crosslinking and molecular modeling to analyze the global architecture of ribonuclease P RNA. EMBO J. 1994;13(17):3953–63.
129. Shi Y-Z, Wu Y-Y, Wang F-H, Tan Z-J. RNA structure prediction: progress and perspective. Chinese Phys B. 2014;23(7):078701.
130. Massire C, Westhof E. MANIP: an interactive tool for modelling RNA. J Mol Graph Model. 1998;16(4–6):197–205, 255–7
131. Zwieb C, Wower I, Wower J. Comparative sequence analysis of tmRNA. Nucleic Acids Res. 1999;27(10):2063–71.
132. Hammann C, Westhof E. Searching genomes for ribozymes and riboswitches. Genome Biol. 2007;8(4):210.
133. Bindewald E, Grunewald C, Boyle B, O'Connor M, Shapiro BA. Computational strategies for the automated design of RNA nanoscale structures from building blocks using NanoTiler. J Mol Graph Model. 2008;27(3):299–308.
134. Jossinet F, Ludwig TE, Westhof E. Assemble: an interactive graphical tool to analyze and build RNA architectures at the 2D and 3D levels. Bioinformatics. 2010;26(16):2057–9.
135. Lu F, Ammiraju JSS, Sanyal A, Zhang S, Song R, Chen J, et al. Comparative sequence analysis of MONOCULM1-orthologous regions in 14 Oryza genomes. Proc Natl Acad Sci U S A. 2009;106(6):2071–6.
136. Martinez HM, Maizel JV, Shapiro BA. RNA2D3D: a program for generating, viewing, and comparing 3-dimensional models of RNA. J Biomol Struct Dyn. 2008;25(6):669–83.
137. Jossinet F, Westhof E. Sequence to structure (S2S): display, manipulate and interconnect RNA data from sequence to structure. Bioinformatics. 2005;21(15):3320–1.

138. Wu S, Zhang Y. A comprehensive assessment of sequence-based and template-based methods for protein contact prediction. Bioinformatics. 2008;24(7):924–31.
139. Zhang Y, Skolnick J. Segment assembly, structure alignment and iterative simulation in protein structure prediction. BMC Biol. 2013;11(1):44.
140. Zhao Y, Huang Y, Gong Z, Wang Y, Man J, Xiao Y. Automated and fast building of three-dimensional RNA structures. Sci Rep. 2012;2(1):734.
141. Popenda M, Szachniuk M, Antczak M, Purzycka KJ, Lukasiak P, Bartol N, et al. Automated 3D structure composition for large RNAs. Nucleic Acids Res. 2012;40(14):e112.
142. Rother M, Rother K, Puton T, Bujnicki JM. RNA tertiary structure prediction with ModeRNA. Brief Bioinform. 2011;12(6):601–13.
143. Tan Z-J, Chen S-J. Chapter 22—Predicting electrostatic forces in RNA folding. In: Methods in enzymology [Internet]. Biophysical, chemical, and functional probes of RNA structure, interactions and folding: part B; vol. 469. Academic Press; 2009 [cited 2020 Oct 29]. p. 465–87. Available from: http://www.sciencedirect.com/science/article/pii/S0076687909690224.
144. Case DA, Cheatham TE, Darden T, Gohlke H, Luo R, Merz KM, et al. The Amber biomolecular simulation programs. J Comput Chem. 2005;26(16):1668–88.
145. Brooks BR, Brooks CL, MacKerell AD, Nilsson L, Petrella RJ, Roux B, et al. CHARMM: the biomolecular simulation program. J Comput Chem. 2009;30(10):1545–614.
146. Bida JP, Maher LJ. Improved prediction of RNA tertiary structure with insights into native state dynamics. RNA (New York, NY). 2012;18(3):385–93.
147. Das R, Karanicolas J, Baker D. Atomic accuracy in predicting and designing noncanonical RNA structure. Nat Methods. 2010;7(4):291–4.
148. Zuker M. Mfold web server for nucleic acid folding and hybridization prediction. Nucleic Acids Res. 2003;31(13):3406–15.
149. Zhang J, Bian Y, Lin H, Wang W. RNA fragment modeling with a nucleobase discrete-state model. Phys Rev E. 2012;85(2):021909.
150. Paliy M, Melnik R, Shapiro BA. Coarse-graining RNA nanostructures for molecular dynamics simulations. Phys Biol. 2010;7(3):036001.
151. de Pablo JJ. Coarse-grained simulations of macromolecules: from DNA to nanocomposites. Annu Rev Phys Chem. 2011;62:555–74.
152. Harvey SC, Malhotra A, Tan RK-Z. Molecular modeling studies on the ribosome. Mol Eng. 1995;5(1):213–8.
153. Xia Z, Gardner DP, Gutell RR, Ren P. Coarse-grained model for simulation of RNA three-dimensional structures. J Phys Chem B. 2010;114(42):13497–506.
154. Hori N, Takada S. Coarse-grained structure-based model for RNA-protein complexes developed by fluctuation matching. J Chem Theory Comput. 2012;8(9):3384–94.
155. Denesyuk NA, Thirumalai D. Coarse-grained model for predicting RNA folding thermodynamics. J Phys Chem B. 2013;117(17):4901–11.
156. Jonikas MA, Radmer RJ, Altman RB. Knowledge-based instantiation of full atomic detail into coarse-grain RNA 3D structural models. Bioinformatics. 2009;25(24):3259–66.
157. Tan RKZ, Petrov AS, Harvey SC. YUP: a molecular simulation program for coarse-grained and multi-scaled models. J Chem Theory Comput. 2006;2(3):529–40.
158. Kerpedjiev P, Höner zu Siederdissen C, Hofacker IL. Predicting RNA 3D structure using a coarse-grain helix-centered model. RNA. 2015;21(6):1110–21.
159. Woodson SA. Metal ions and RNA folding: a highly charged topic with a dynamic future. Curr Opin Chem Biol. 2005;9(2):104–9.
160. Schroeder SJ. Challenges and approaches to predicting RNA with multiple functional structures. RNA. 2018;24(12):1615–24.

Structural Proteomics

11

Manoj Kumar Gupta, Gayatri Gouda, S. Sabarinathan, Ravindra Donde, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

Structural proteomics identifies three-dimensional (3D) protein structures at an atomic resolution on a genome-wide scale to better understand the interaction among protein sequence, structure, and function. The 3D structure of proteins is mostly estimated via x-ray crystallography or nuclear magnetic resonance spectroscopy. However, for the overwhelming majority of protein sequences, no experimental structure is available to date. This gap in structural proteomics can be overcome by computational approaches. The prediction of protein structure through computational approaches may be addressed in three major ways: (1) computer simulation focused on empirical energy calculations; (2) knowledge-based approaches that employ information obtained from structural-sequence relationships retrieved from already available experimentally defined 3D protein structures; and (3) ab initio methods. Irrespective of all these, the creation of an exact model is not often feasible and may sometimes generate

M. K. Gupta (✉) · G. Gouda · R. Donde · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_11

incorrect models. Analysis with incorrect models may provide wrong information about the structure and function of biomolecules. Hence, validation of the model generated is highly required prior to downstream analysis. The precision of a generated model is calculated by different factors, including the existence of solved protein structures that can be employed as reference models. In the near future, the chapter's information will be highly useful for constructing a more accurate three-dimensional structure of proteins, which, in turn, will help us understand the biological function in a more comprehensive way.

Keywords

Protein structures · Homology modeling · Model validation · Threading · ab initio

Abbreviations

3D	Three-dimensional
CASP	Critical Assessment of Techniques for Protein Structure Prediction
HMM	Hidden Markov model
I-TASSER	Iterative threading assembly refinement
MQA	Model Quality Assessment
MSA	Multiple sequence alignment
NMR	Nuclear magnetic resonance
PPA	Profile-profile alignment
RMSD	Root-mean-square deviation
SVM	Support vector machine
XRC	X-ray crystallography

11.1 Introduction

The accumulation of genomic sequence has contributed to the development of structural proteomics, which, in turn, had led to the recognition of a significant number of protein architectures. The vast number of protein structures anticipated from numerous approaches can provide useful insights into the guidelines for the prediction of protein folding and biochemical activity. One of the key objectives of the structural proteomics initiative is to find the best plausible technologies and efficient processes for converting gene sequence to three-dimensional (3D) structure [1]. Interestingly, most of the technologies evaluate the best result obtained from either x-ray crystallography (XRC) or nuclear magnetic resonance (NMR) spectroscopy. XRC is currently regarded as a potential workhorse for structural proteomics since it is possible to evaluate a 3D structure in hours when supplied with a well-diffracting crystal. However, the performance of structural determination employing XRC remains an open question, as the rate determination stage still generates only crystals that are well-diffracted. An erratic method may take hours and months [1]. On the other side, NMR research involves no crystals, and structure-specific samples can be detected in minutes of purification. Additionally, the determination of the NMR structure is currently limited to size restrictions and lengthy data collection and evaluation periods (often months). This process is better applicable for proteins having <250 amino acids. In brief, XRC and NMR spectroscopy tend to have complementary defects, and the relative effectiveness of such approaches is still to be identified in structural proteomics. Irrespective of all these advancements, for the overwhelming majority of protein sequences, no experimental structure is available. Earlier studies have reported that this gap in structural proteomics can be overcome easily by computational approaches [2, 3].

Structural bioinformatics, initially referred to as structural computational biology, supersedes bioinformatics' other forms [4]. It may be claimed, in reality, that Watson and Crick's seminal paper of 1953 is a model work and perhaps the first structural document on bioinformatics [5]. For Martin Karplus, Arie Warshel, and Michael Levitt, the 2014 Nobel Award for "multiscale modeling" is a major hallmark that acknowledges the effect of structural bioinformatics on science. In his description of the emergence of the sector, Levitt explains how computation is needed to correctly refine Crick's tRNA model for the creation of an actual model [6]. Simulation has also been an important part of structural biology since its very beginning and has played an important role in biochemistry and molecular biology through the years. Indeed, we have now entered the stage of millisecond simulations following the first simulations of small systems and a couple of picoseconds recognized by the Nobel Committee [4].

The protein structures obtained from Structural Proteomics Centers may be used to evaluate the current protein structure using molecular substitution. Structural proteome structures may also be exploited to build homology models for whole protein families utilizing computational bioinformatics techniques, and these structural models may also be used for understanding these proteins' roles without the requirement for experimental 3D structures. For the systematic investigation of proteins that are annotated as a hypothetical protein or with vague functions, structural proteomic structures also play an important role. Information on precise 3D protein structures is also a prerequirement for reasonable drug design. Structural proteomics has a significant effect on drug development pipelines and plays role in understanding the molecular processes at the atomic level controlling human diseases [7]. Thus, in this chapter, the authors attempt to understand recent advances in structural proteomics for protein structure determination. At first, we provide a detailed description of instrumental approaches like NMR spectroscopy and XRC; subsequently, we provide brief descriptions of computational procedures, including comparative and de novo structure prevision. In the end, we explain different approaches to test the accuracy of the generated models.

11.2 Approaches for Structure Prediction

As stated above, the 3D structure of any protein can be estimated by both experimental and computational approaches [8] (Fig. 11.1).

11.2.1 Experimental Approaches

At present, XRC and NMR are the key experimental methods that are widely used for evaluating the 3D structure of proteins [8].

X-Ray Crystallography

The exploration, creation, and precision of XRC are the result of several groundbreaking activities [9, 10]. In 1895, X-rays were invented by W. C. Röentgen [10]. In 1912, M von Laue illustrated the potential of X-rays that diffracted from single crystals. This discovery awarded him the Nobel Physics Award in 1914 [11]. In the following year, W.L. Bragg coined the rule on diffraction, identified as Bragg's law, which showed the usage of different patterns of diffraction in the assessment of crystal structure of NaCl [12]. The invention of instrumentation for observing diffraction patterns was pioneered by W. H. Bragg (father of W.L. Bragg). Bragg's rule (only named after the son W. L. Bragg) was used to examine the pattern of diffraction, and in 1915, their combined and separate experiments earn them Nobel Prize. To date, 29 Nobel prizes have been awarded for the significance as well as the applications of x-ray diffraction in many research fields [10].

XRC is a technique used to establish a crystal's atomic and molecular composition. The fundamental theory is that a beam of X-rays diffracts through several unique directions of crystalline atoms [13]. A crystallographer may create a 3D image of the electron density in the crystal by calculating the angles and intensities of these diffracted beams. The average locations of the atoms within the crystal, their chemical bonds, and their disorder and numerous other details maybe calculated from this picture of electron densities. The system shows how certain biological

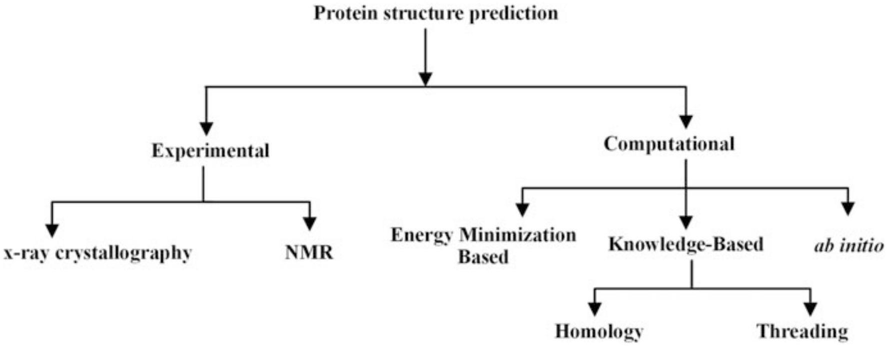


Fig. 11.1 Different approach for predicting protein structure

molecules, including vitamins, medications, proteins, and nucleic acids, such as DNA, are organized as well as function. It is pertinent to note that James Watson & Francis Crick observed the double helix structure of DNA via XRC. The recent improvement in image restoration technology rendered XRC appropriate for the structural study of even greater complexes. The biggest failure in XRC is that it is impossible to obtain a virus particle crystal, a precondition required for successful XRC. Another drawback is that XRC normally demands that samples are positioned in nonphysiologic conditions, which may often induce technically insignificant conformational shifts [13].

NMR

NMR imaging is arguably today the most versatile biochemical testing methods. The NMR concept was first noticed in the 1940s and was mainly the work of physicists. During the next 50 years or so, NMR applications were quickly established and first utilized by chemists. In the late 1960s, the usage of NMR to research protein and other biological molecule structure was greatly enhanced with the development of superconducting magnets and Fourier Transform NMR. It was not until the mid-1970s, nevertheless, that the first NMR applications were recorded for the metabolism analysis of living biological systems. At around the same time, it has been shown the usage of magnetic field gradients could be used to spatially encode NMR signals, and hence, the idea of MRI has been born. Soon after, MRI images were collected from the human body, and, as early as 1980, the evaluation of MRI as a clinically effective imaging tool had begun [14]. In recent years, NMR has also been rendered as a viable biomarker exploration technique by integrating advanced hardware and chemometric techniques. NMR is ideally adapted for mixture analyses and has many attractive biomarker exploration characteristics such as limited sample preparation, nondestructive screening, and simultaneous metabolite identification with complex physiochemical properties. NMR's excellent intra- and interlaboratory analytical replicability has been well-documented [15, 16]. Under suitable circumstances, NMR spectra are quantitative since the peak region is directly proportional to the corresponding number of nuclei.

The effectiveness of biomarker discovery via NMR spectroscopy has been well-documented in cancer [17], genetic disorders [18], and toxicology research [19]. Biomarker detection can be performed with urine and blood and other biological fluids and tissues obtainable through minimum intrusive techniques. In one study, researchers reviewed the usage of NMR spectroscopy to identify possible cancer molecular markers in human biofluids [17]. The authors presented a detailed list of 15 forms of cancer testing, tabled details on gathering and storing samples, data processing, and statistical analysis, described metabolic modifications within disease versus control, and recommended pathway agitations [17]. Other articles have also analyzed numerous NMR implementations and examined the significance of sample selection, storage, and experimental techniques [20, 21]. "Metabolome-wide association studies" (MWASs) have grown rapidly as a tool to diagnose disease risk factors depending on metabolite concentrations in the broad biological sample, rather than relying on detecting particular metabolic biomarkers in pre-existing conditions

[22, 23]. Bernini et al. [24] demonstrated the stability of human metabolic phenotypes over many years. Another 1H NMR analysis showed that consistent hereditary and environmental effects accounted for 47% in urine and 60% in plasma for biological differences within the metabolite levels [25]. As stated above, irrespective of all these findings, for the overwhelming majority of protein sequences, no experimental structure is available. This gap in structural proteomics can be overcome by computational approaches [2, 3].

11.2.2 Computational Approaches

The prediction of protein structure through computational approaches may be addressed in three major ways: (1) computer simulation focused on empirical energy calculations; (2) knowledge-based approaches that employ information obtained from structural-sequence relationships retrieved from experimentally defined 3D protein structures; and (3) ab initio methods [8].

11.2.2.1 Energy Minimization-Based Methods

Protein structure prediction based on energy minimization techniques is based on measurements of the equivalence of native protein structures to a thermodynamic equilibrium regime with limited free energy. Energy-based approaches should not make pre-existing conclusions regarding amino acid coding properties. Instead, efforts to establish the minimum global amount of the protein molecule’s surface-free energy are believed to conform to its native configuration. Methods focused on energy minimization may be roughly divided into two categories: (1) static minimization methods and (2) dynamic minimization methods. ECEPP, AMBER, CHARMS, and GROMOS are the most appropriate energy minimization program packages [26, 27]. The benefit of energy measurements is that they are focused on physicochemical theories, but they are complicated by a huge number of degrees of freedom and the restricted performance of energy functions. There are generally two big issues with energy minimization-based methods. First, the equations needed for the allocation of protein structures dependent on the reduction of energy are beyond the scope of today’s computers. Second, for such measurements, the interaction potential used is not adequate to model the native structure of the protein at the atomic detail [8].

11.2.2.2 Knowledge-Based Approaches

Knowledge-based approaches are subdivided into homology modeling and threading [8, 28]. Table 11.1 lists software for predicting protein structure.

Homology Modeling

Homology modeling (or comparative modeling) works on the biological reality that two sequences that have a close similarity/identity also have similar structures. This approach generates the 3D protein structure employing the following steps: (1) the best plausible template is recognized for the specified target sequence by using

Table 11.1 Software for predicting protein structure

Name	Name	Method	Link
Homology modeling	Biskit	Employs external tools like T-coffee, BLAST, and MODELLER for generating three-dimensional structure	[29]
	ESyPred3D	Template identification, alignment, and three-dimensional modeling	[30]
	FoldX	Energy estimation and three-dimensional modeling	[31]
	MODELLER	Spatial restraints satisfaction	[32]
	CONFOLD	Contact and distance restraints satisfaction	[33]
	ROBETTA	Rosetta homology modeling and ab initio fragment assembly through Ginzu domain identification	[34]
	BHAGEERATH-H	Augmentation of ab initio folding and homology approaches	[35]
	SWISS-MODEL	Local similarity or fragment assembly	[36]
	Yasara	Templates identification, alignment, three-dimensional modeling including ligands as well as oligomers, and model fragments hybridization	[37]
	AWSEM-Suite	Molecular dynamics simulation based on template-guided, coevolutionary-enhanced optimized folding landscapes	[38]
Homology and threading modeling	RaptorX	Identification of remote homology, three-dimensional modeling, and prediction of binding site	[39]
	HHpred	Template identification, alignment, and three-dimensional modeling	[40]
	Phyre2	Identification of remote homology, alignment, three-dimensional modeling, multitemplates, and ab initio	[41]
Homology and ab initio modeling	ROBETTA	Rosetta homology modeling and ab initio fragment assembly through Ginzu domain identification	[34]
Ab initio modeling	trRosetta	It predicts the structure of the protein on the direct energy minimizations with a restrained Rosetta.	[42]
	I-TASSER	Threading fragment structure reassembly	[43]

BLAST search, (2) sequence alignment correction, (3) to ensure the alignment of conserved or functionally essential residues, (4) backbone prediction, (5) modeling of the loop, and (6) sidechain modeling using rotamer libraries [44]. MODELLER (https://salilab.org/modeller/), "iterative threading assembly refinement" (I-TASSER) [45], and SWISS-MODEL (https://swissmodel.expasy.org/) are a few main homology modeling methods and servers. MODELLER constructs a protein model by comparative modeling between the template and target sequences provided [32]. In the easiest form, the input is to match the query sequence that is about to be modeled with the template structure(s), the template(s), atomic coordinates, and a basic script file. MODELLER then generates a model that includes all the nonhydrogen atoms, without even a user intervention and on a desktop machine within minutes [46]. In addition to model construction, MODELLER may conduct auxiliary operations, like the fold allocation [47], the alignment of two protein profiles or their sequences [48], multiple protein sequence and/or structure alignments [49], sequence and/or structure clustering, and ab initio protein structure loop modeling [46]. The modeler is often used to model loops and optimize protein [32].

SWISS-MODEL is another workplace for the homology modeling of protein structures [50–52]. The first completely automatic protein homology simulation server, SWISS-MODEL (https:/swissmodel.expasy.org), has constantly improved over the last 25 years [43, 50, 53, 54]. Its modeling functionality has recently been expanded into consideration of the amino acid sequences of interaction partners and involves the modeling of homomeric and heteromeric complexes. The latest modeling engine ProMod3 has been developed with improved precision of the generated models and along with a refined method of assessment of the local model output (QMEANDisCo) focuses on a recent version of QMEAN [50, 55], which has been released recently [52].

The I-TASSER server is an optimized protein structure prediction tool based on the paradigm for the sequence to structure to function. I-TASSER initially produces three-dimensional atomic modeling from several threading alignments and structural iterative assembly simulations focused on an amino acid chain. The protein function is then deduced by comparing the 3D models with other recognized proteins structurally [45]. C-score is a confidence score for the quality evaluation of estimated I-TASSER models. It is determined depending on the context and the convergence parameters of structural assembly simulations of threading template alignments. C-score is usually mostly in [−5,2] range, where a higher value C-score reflects a high trust model and vice versa (https://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html). TM score is another newly proposed scale for the structural resemblance calculation between the two systems. The aim of proposing a TM score is to solve the RMSD issue that is susceptible to local error. Because RMSD is an average distance of all residual pairs in two structures, local error (e.g., tail misorientation) will exist at a great RMSD value even though the global topology is right. However, in TM score, the distance is weighted more heavily than the distance, which renders the score indifferent to the error in local modeling. A TM score > 0.5 implies the right topology model, and a TM score > 0.17 is a spontaneous similarity. This cutoff does not rely on the duration of the protein (https://zhanglab.ccmb.med.umich.edu/I-TASSER/about.html).

Threading Approach

In the threading, a new sequence is positioned on a collection of established folds to find the best score (lowest energy) for one fold. The standard quasienergetic scoring scheme allots energy to alignment, in the expectation that a similar structure to the native fold of the query sequence will have the least energy. Threading is basically similar to sequence alignment. The statistical value of the findings was calculated by an integral feature of sequence alignment procedures [56]. After its first use in the early 1990s, threading has been one of the protein structure prediction's most popular areas. Threading methods contain structural profile alignments, sequence profile alignments, deep learning, and hidden Markov model (HMM) [57]. The sequence "profile-profile alignment" (PPA) is perhaps the most widely employed, stable threading approach. Rather than aligning the single target and template sequences, PPA aligns multiple sequence alignment (MSA) targets to an MSA template. The alignment score throughout the PPA is generally determined as a product of the amino acid frequency at each target MSA location with the log-odds of the matching amino acid in the MSA template, although there are alternate methods for measuring the profile-alignment score. Profile-profile alignment approaches displayed benefits in many recent blind experiments. For example, in LiveBench-8, the top four servers (SFST/STMP, BASD/MASP/MBAS, ORF2 / ORFS, and FFAS03) are all sequence-profile alignment-dependent. Several sequence profile-based approaches were rated top of single threading servers in the CASP Server Segment [58] and the CAFASP [59]. Earlier, Wu and Zhang [60] briefly demonstrated that by adding a variety of additional structural details, the precision of sequence profile alignments could be further increased by nearly 5–6%.

HHsearch, an HMM-HMM alignment tool, has been distinguished as the best individual threading server in CASP7 [61]. The concepts of the HMM-HMM alignments, as well as the profile alignments, are similar, as each tries to pair the target MSA to the MSA template. Instead of portraying the MSAs through sequence profiles, HHsearch employs HMM profiles that can produce sequences with some probabilities calculated by the amino acid product emission and the likelihood of insertion/deletion. HHsearch aligns the target with the HMM blueprint by optimizing the chance that two models coroduce the same amino acid sequence. This optimizes the amino acid frequencies and insertions/deletions in both HMMs [61].

3D-PSSM (http://www.sbg.bio.ic.ac.uk/~3dpssm/index2.html) is a web-based software package that uses a method for structural profile recognition of protein folds. The profiles of each superfamily protein are created by integrating many smaller profiles. The primary aim is to superimpose protein structures within a superfamily focused on the SCOP definition by adding secondary structures and solvent accessibility details for the corresponding residue. Furthermore, each member of a structural protein superfamily has its own PSI-BLAST sequence-based profile determined. In conjunction with the profile of the structure, these sequence profiles are used to shape a broad superfamily profile in which each location includes sequence and structural details. For the query sequence, a sequence-based profile is created by PSI-BLAST. To forecast the secondary structure, PSI-PRED is used. The sequence profile and the expected secondary construction are contrasted by dynamic programming with the precomputed protein superfamily profiles. The matching scores are measured in secondary structure, solvation, and sequence profiles and are graded as the top scoring structure folded [28].

GenThreader (http://www.ebisu.co.uk/chemogenomix.com/chemogenomix/GenThreader.html) is a web-based application that uses a profile hybrid and a pair-wise energy system. The initial stage is similar to the 3D-PSSM; 3 rounds of PSI-BLAST are needed for the sequence of the query protein. The resulting several sequence hits are used for profile creation. PSIPRED is employed for predicting its secondary structure. Both are employed as a threaded measurement input dependent on a pair-wise energy potential process. The outcome is assessed by neural networks that incorporate energy capacity, sequence alignment, and longitude data to establish a single score reflecting the relation among the template and query proteins [28].

11.2.2.3 *ab initio* Protein Structure Prediction

Both homology and fold recognition methods, are focused on the usability of template structures mostly in the database to forecast. If the database contains no right structure, these methods malfunction. Proteins in nature, though, fold individually without testing what their homolog configurations are in databases. There is, of course, detail in the sequences that instruct proteins to "form" their native structures. Early biophysical experiments indicate that most proteins spontaneously fold into an almost minimal energy stable form. This structural condition is called the native state. This folding method seems to be nonrandom, but the structure is uncertain. The minimal understanding of protein folding is the foundation for the ab initio forecast. The ab initio procedure, as the name implies, aims to generate all-atomic protein models focused solely on sequence data without the aid of existing protein structures. The presumed benefit of this approach is that projections are not constrained by existing folds, and new protein folds are discovered. However, as the physicochemical rules regulating protein folding are currently not well-known, the energy functions used in the ab initio forecast are currently very imprecise. The topic of folding remains one of the major problems in bioinformatics today [28].

Present ab initio algorithms cannot yet approximate the protein folding mechanism correctly. They function by utilizing heuristics of some kind. Since the native state of a protein structure is nearly the minimal energy cost, the prediction programs are thus built utilizing the energy minimization theory. These algorithms aim to find the one with the least global energy in any conformation. However, in practice, it might not be correct to find a fold with absolute minimal energy. This points to one of the main drawbacks of this strategy. Moreover, it is not computationally viable to check for all potential structural conformations. It was predicted that it would take 10–20 years to sample all potential conformations of a 40-residue protein using one of the largest supercomputers in the world (1 trillion operations per second). Any kind of heuristics must then be used to minimize the room to be checked for conformational purposes. Any new ab initio strategies combine the quest for fragments with threading to create an unknown protein model [28].

Rosetta is a three-dimensional protein conformation prediction server that utilizes the ab initio process [28]. A customized database of three and nine residue segments is used in ROSETTA. For each section, a group of structural motifs has been identified, and the simulation relies on the assembly of these motives into a consistent 3D framework. For a standard goal, ~10,000 attempts are required. The minimization of intramolecular associations is determined by an individual scoring feature that takes into account hydrophobicity, disulfide bands, α-helices, β-twists, and β-sheet assemblies. ROSETTA proposes that local connexions may be modeled to various collections of structural motifs through mapping chain fragments and that nonlocal interactions choose low-free tertiary structures from local mapping-compliant conformations [62]. Nonlocal contact is fairly well-modeled at the level necessary for folding, without comprehensive side chains. ROSETTA-ab initio is an ab initio protein structure prediction kit, focused on the premise that local interactions direct the conformation of short segments, whereas global interactions decide a three-dimensional structure, consistent with local biases [62]. Once these configurations are identified, a Monte Carlo protocol may involve each query sequence in a variety of higher-order combinations. The resulting structures in a semiempirical force field are subject to energy minimization, taking into account hydrophobic and electrostatic interactions, central associations of hydrogen, and omitted amounts. Finally, structures consistent with local and nonlocal connexions are categorized according to their total energy, measured via the minimization method. The EUChinaGrid project employed ROSETTA to model the protein set known as "never born proteins." About 10,000 (60 random amino acids per protein) were produced, with an equal likelihood of each amino acid occurring. The collection should include a variety of pharmacologically active proteins. The protein structures were folded using both the ROSETTA model and "fuzzy oil decrease" and produced RMSD of about 6.7–7.7 Å in the most common structures and 25–32 Å for the most varying structures. The comparative study of the two groups indicates that the structures created by the "fuzzy oil drop" model (which was in line with expectations) have a well-defined hydrophobic center. ROSETTA repeated this phenomenon not consistently: although the folded proteins often constituted strongly hydrophobic regions, these zones were not necessarily situated close to the middle of the molecule [62].

Models obtained by de novo prediction techniques are useful to gain biological insight through either functional site recognition or functional annotation by fold identification. The Rosetta method is quick enough to allow genome-scale research to be made feasible. Earlier, ~500 PfamA family structures with no relation to the established structure were expected [63]. Irrespective of all these advancements, earlier studies have reported that the algorithms of ab initio prediction are far from mature. Their estimation accuracies are too poor for realistic usage. Initially, protein structure estimation remains an imaginary target for the future. However, considering the latest high-throughput structural evaluation by the structural proteomics initiative, which aims at solving all protein pliers in a decade, it may soon come when the ab initio modeling method is unnecessary since homology modeling and threading will deliver far higher quality predictions for all potential protein pliers. Regardless of the advances achieved in structural proteomics, the study of protein structures using the ab initio prediction method will still offer insight into the mechanism of protein folding [28].

11.3 Model Evaluation

As stated above, protein structure prediction techniques have advanced rapidly over the past couple of decades, with expanded software availability and regular usage of biological studies in computational models [64]. However, also with state-of-the-art methods [65, 66], the creation of an exact model is not often feasible. The precision of a structure model is calculated by different factors, including the existence of solved protein structures that can be used as simulation models. Since the reliability of the models varies, it is important to know the accuracy of a given computational model for the realistic application of the biological research model. The prediction of the accuracy of protein-tertiary structure models is named the Model Quality Assessment (MQA) and became an active focus in structural bioinformatics research [67]. MQA is not just for the detection of high-precision models. For instance, models with modest precision are useful for several purposes [68]. Models with atomic-detailed precision with an RMSD of 1.5–2 Å to the native structure are useful for almost every application in which the structural knowledge is helpful, such as enzyme and protein engineering studies and drug design [69]. For applications that need residue level precision, e.g., the design and analysis of site-directed mutagenesis studies, models of proper backbone orientation (e.g., RSMD 4 to 6 Å) may be used. Models with a slightly higher (worse) RMSD but nearly right overall fold can be used to predict the feature in their global fold [70], to explain structural data with a low resolution [71] or to classify local functional sites [72–74]. Appropriate usage of a model for the above-described applications is only feasible if consumers know about the model's accuracy. It is, therefore, necessary to develop methods to determine the consistency of the models of the protein structure such that they are appropriate for their predicted accuracy in applications. MQA is also an important step in refining models of structures [75].

There are two groups of MQA protein methods that estimate the global and local consistency of a model structure, respectively. In the former class, RMSD or other associated metrics showing the global structural similarities of a model with the protein's native structure was anticipated. By contrast, the local consistency prediction methods are structured to show the exactness or mistake of each model residue, such as a distance between the Cα atoms' location in the expected and the native structures. More MQA approaches for global rather than local quality are produced. The former is also much more effective in the "Critical Assessment of Techniques for Protein Structure Prediction" (CASP) [67], a communal protein structure prediction experiment. The global quality of models, mostly formulated in a machine learning system, may be expected from structural and sequence (target-template alignment) features of models or their combinations. But useful structural attributes include the residue/atomic interaction potentials [76, 77], main-chain torsion angles, [78] and residue exposure or burial propensity [79], while the alignment characteristics include target alignment scores measured for a target sequence/profile sequence identity and statistical importance [80, 81]. Their structural characteristics are useful. Methods like regression [82] or linear combination [83], support vector machine (SVM) [84, 85], and neural network [86–88] were used to merge functions.

In the CASP, consensus strategies investigating the coherence of models developed through various methods of structural prediction were successfully carried out [89, 90].

In comparison to global MQA, local MQA approaches need major changes in practice. The correlation coefficients of expected and real, local errors in structural model models were stated to be significantly lower than those of the global accuracy prediction [67]. Local consistency knowledge informs users of regions or residues in a precise projected structure model. It also offers knowledge of considerable significance for the functional execution of structural models. Current local MQA approaches adopt similar approaches to global MQA methods: in a machine learning setting, structural and sequence features of a model are known as residue, e.g., neural networks [91] and SVM [92]. SMOQ [93], MULTICOM [94, 95], and Wang deep [96] are also recently established local QA methods.

11.4 Conclusions and Future Perspective

In conclusion, in bioinformatics and computational biology, the prediction of the protein tertiary structure from amino acid sequence is very important. A lot of protein tertiary structure prediction techniques have been established in recent decades. One method category adopts a template-based methodology, which utilizes experimentally defined frameworks as models for the creation of structural templates for an undefined target protein. Another group uses a template-free method, which aims to fold the protein without using established template structures. Both approaches were frequently merged in order to cope with a wide variety of protein structure prediction issues from comparatively simple homology simulation to hard-de-novo estimation. One essential role in the prediction of protein structure is to evaluate the consistency of structural models provided by methodologies for the prediction of protein structures. For assessment, refinement, and collection of models, a sample quality assessment process used in a pipeline for prediction of protein structure is crucial. Generally, a sample quality evaluation approach may estimate a global quality score that measures the total quality of a model protein structure and a set of local quality measurements that measure the local quality of each residue in the product. A global quality score can be a global distance test score, which estimates the structural similarity between the model and the unknown native protein structure. The Euclidean distance from the location of the residue in a model and that in the unknown native structure after superimposition may be a local quality score of a residue. In the near future, information present in the chapter will be highly useful for constructing the three-dimensional structure of proteins, which, in turn, will help us to understand the biological function in a more comprehensive way.

Conflicts of Interest None.

References

1. Yee A, Chang X, Pineda-Lucena A, Wu B, Semesi A, Le B, et al. An NMR approach to structural proteomics. PNAS. 2002 Feb 19;99(4):1825–30.
2. Gupta MK, Vadde R. Insights into the structure–function relationship of both wild and mutant zinc transporter ZnT8 in human: a computational structural biology approach. J Biomol Struct Dyn. 2020 Jan 2;38(1):137–51.
3. Gupta MK, Vadde R. A computational structural biology study to understand the impact of mutation on structure–function relationship of inward-rectifier potassium ion channel Kir6.2 in human. J Biomol Struct Dyn. 2020 Feb 23;39(4):1–14.
4. Samish I, Bourne PE, Najmanovich RJ. Achievements and challenges in structural bioinformatics and computational biophysics. Bioinformatics. 2015 Jan 1;31(1):146–50.
5. Watson JD, Crick FH. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature. 1953 Apr 25;171(4356):737–8.
6. Levitt M. The birth of computational structural biology. Nat Struct Biol. 2001 May;8(5):392–3.
7. Manjasetty BA, Turnbull AP, Panjikar S. The impact of structural proteomics on biotechnology. Biotechnol Genet Eng Rev. 2009 Jan 1;26(1):353–70.
8. Bhasin M, Raghava GPS. 8 - computational methods in genome research. In: Arora DK, Berka RM, Singh GB, editors. Applied mycology and biotechnology [internet], Applied mycology and biotechnology, vol. 6. Amsterdam: Elsevier; 2006. p. 179–207. [cited 2020 Nov 2]. Available from: http://www.sciencedirect.com/science/article/pii/S1874533406800110.
9. Waseda Y, Matsubara E, Shinoda K. X-Ray diffraction crystallography: introduction, examples and solved problems [internet]. Berlin Heidelberg: Springer-Verlag; 2011 [cited 2020 Nov 2]. Available from: https://www.springer.com/gp/book/9783642166341
10. Aitipamula S, Vangala VR. X-Ray crystallography and its role in understanding the physicochemical properties of pharmaceutical Cocrystals. J Indian Inst Sci. 2017 Jun 1;97(2):227–43.
11. Friedrich W, Knipping P, Laue M. Interferenzerscheinungen bei Röntgenstrahlen. Ann Phys. 1913;346(10):971–88.
12. Bragg WH. IX. Bakerian lecture.— X-rays and crystal structure. In: Philosophical transactions of the Royal Society of London series a, containing papers of a mathematical or physical character, vol. 215(523–537); 1915 Jan 1. p. 253–74.
13. Ryu W-S. Chapter 2 - virus structure. In: Ryu W-S, editor. Molecular virology of human pathogenic viruses [internet]. Boston: Academic Press; 2017. p. 21–9. [cited 2020 Nov 2]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128008386000023.
14. Chatham JC, Blackband SJ. Nuclear magnetic resonance spectroscopy and imaging in animal research. ILAR J. 2001 Jan 1;42(3):189–208.
15. Keun HC, Ebbels TMD, Antti H, Bollard ME, Beckonert O, Schlotterbeck G, et al. Analytical reproducibility in 1H NMR-based Metabonomic urinalysis. Chem Res Toxicol. 2002 Nov 1;15 (11):1380–6.
16. Dumas M-E, Maibaum EC, Teague C, Ueshima H, Zhou B, Lindon JC, et al. Assessment of analytical reproducibility of 1H NMR spectroscopy based metabonomics for large-scale epidemiological research: the INTERMAP study. Anal Chem. 2006 Apr 1;78(7):2199–208.
17. Rocha CM, Carrola J, Barros AS, Gil AM, Goodfellow BJ, Carreira IM, et al. Metabolic signatures of lung Cancer in biofluids: NMR-based Metabonomics of blood plasma. J Proteome Res. 2011 Sep 2;10(9):4314–24.
18. Wu B, Barile E, De SK, Wei J, Purves A, Pellecchia M. High-throughput screening by nuclear magnetic resonance (HTS by NMR) for the identification of PPIs antagonists. Curr Top Med Chem. 2015;15(20):2032–42.
19. Coen M, Holmes E, Lindon JC, Nicholson JK. NMR-based metabolic profiling and metabonomic approaches to problems in molecular toxicology. Chem Res Toxicol. 2008 Jan;21(1):9–27.
20. Gowda GAN, Zhang S, Gu H, Asiago V, Shanaiah N, Raftery D. Metabolomics-based methods for early disease diagnostics. Expert Rev Mol Diagn. 2008 Sep;8(5):617–33.

21. Issaq HJ, Veenstra TD. Proteomic and Metabolomic approaches to biomarker discovery. Boston: Academic Press; 2019. 506 p.
22. Van QN. Chapter 7 - current NMR strategies for biomarker discovery. In: Issaq HJ, Veenstra TD, editors. Proteomic and Metabolomic approaches to biomarker discovery. 2nd ed. Boston: Academic Press; 2013. p. 103–31. [cited 2020 Nov 2]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128186077000074.
23. Rattray NJW, Deziel NC, Wallach JD, Khan SA, Vasiliou V, Ioannidis JPA, et al. Beyond genomics: understanding exposotypes through metabolomics. Hum Genomics [Internet]. 2018 Jan 26;12. [cited 2020 Nov 2]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5787293/
24. Bernini P, Bertini I, Luchinat C, Nepi S, Saccenti E, Schäfer H, et al. Individual human phenotypes in metabolic space and time. J Proteome Res. 2009 Sep;8(9):4264–71.
25. Nicholson G, Rantalainen M, Maher AD, Li JV, Malmodin D, Ahmadi KR, et al. Human metabolic profiles are stably controlled by genetic and environmental variation. Mol Syst Biol. 2011 Aug 30;7:525.
26. Brooks JM, Wiesenburg DA, Roberts H, Carney RS, MacDonald IR, Fisher CR, et al. Salt, seeps and Symbiosis in the Gulf of Mexico. EOS Trans Am Geophys Union. 1990;71 (45):1772–3.
27. van Gunsteren WF, Berendsen HJC. Computer simulation of molecular dynamics: methodology, applications, and perspectives in chemistry. Angew Chem Int Ed Engl. 1990;29 (9):992–1023.
28. Xiong J. Essential bioinformatics. Cambridge: Cambridge University Press; 2006. 360 p.
29. Grünberg R, Nilges M, Leckner J. Biskit—a software platform for structural bioinformatics. Bioinformatics. 2007 Mar 15;23(6):769–70.
30. Lambert C, Léonard N, De Bolle X, Depiereux E. ESyPred3D: prediction of proteins 3D structures. Bioinformatics. 2002 Sep 1;18(9):1250–6.
31. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W382–8.
32. Fiser A, Sali A. Modeller: generation and refinement of homology-based protein structure models. Meth Enzymol. 2003;374:461–91.
33. Adhikari B, Cheng J. CONFOLD2: improved contact-driven ab initio protein structure modeling. BMC Bioinformatics. 2018 Jan 25;19(1):22.
34. Kim DE, Chivian D, Baker D. Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res. 2004 Jul 1;32(Web Server issue):W526–31.
35. Jayaram B, Dhingra P, Mishra A, Kaushik R, Mukherjee G, Singh A, et al. Bhageerath-H: a homology/ab initio hybrid server for predicting tertiary structures of monomeric soluble proteins. BMC Bioinformatics. 2014 Dec 8;15(Suppl 16):S7.
36. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018 Jul 2;46(W1):W296–303.
37. Land H, Humble MS. YASARA: a tool to obtain structural guidance in biocatalytic investigations. Methods Mol Biol. 1685;2018:43–67.
38. Jin S, Contessoto VG, Chen M, Schafer NP, Lu W, Chen X, et al. AWSEM-suite: a protein structure prediction server based on template-guided, coevolutionary-enhanced optimized folding landscapes. Nucleic Acids Res. 2020 Jul 2;48(W1):W25–30.
39. Wang S, Li W, Liu S, Xu J. RaptorX-property: a web server for protein structure property prediction. Nucleic Acids Res. 2016 Jul 8;44(Web Server issue):W430–5.
40. Söding J, Biegert A, Lupas AN. The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W244–8.
41. Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJE. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845–58.
42. Yang J, Anishchenko I, Park H, Peng Z, Ovchinnikov S, Baker D. Improved protein structure prediction using predicted interresidue orientations. PNAS. 2020 Jan 21;117(3):1496–503.

43. Yang J, Zhang Y. I-TASSER server: new development for protein structure and function predictions. Nucleic Acids Res. 2015 Jul 1;43(Web Server issue):W174–81.
44. Gromiha MM, Nagarajan R, Selvaraj S. Protein structural bioinformatics: an overview. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology [internet]. Oxford: Academic Press; 2019. p. 445–59. [cited 2020 Nov 2]. Available from: http://www.sciencedirect.com/science/article/pii/B9780128096338202781.
45. Zhang Y. I-TASSER server for protein 3D structure prediction. BMC Bioinformatics. 2008 Jan 23;9(1):40.
46. Fiser A, Do RK, Sali A. Modeling of loops in protein structures. Protein Sci. 2000 Sep;9 (9):1753–73.
47. Eswar N, Eramian D, Webb B, Shen M-Y, Sali A. Protein structure modeling with MODELLER. Methods Mol Biol. 2008;426:145–59.
48. Marti-Renom MA, Madhusudhan MS, Sali A. Alignment of protein sequences by their profiles. Protein Sci. 2004 Apr;13(4):1071–87.
49. Madhusudhan MS, Marti-Renom MA, Sanchez R, Sali A. Variable gap penalty for protein sequence-structure alignment. Protein Eng Des Sel. 2006 Mar;19(3):129–33.
50. Schwede T, Kopp J, Guex N, Peitsch MC. SWISS-MODEL: an automated protein homology-modeling server. Nucleic Acids Res. 2003 Jul 1;31(13):3381–5.
51. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, Schwede T. Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc. 2009 Jan;4(1):1–13.
52. Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, et al. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res. 2014 Jul 1;42(W1):W252–8.
53. Peitsch MC. ProMod and Swiss-model: internet-based tools for automated comparative protein modelling. Biochem Soc Trans. 1996 Feb 1;24(1):274–9.
54. Arnold K, Bordoli L, Kopp J, Schwede T. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics. 2006 Jan 15;22 (2):195–201.
55. Benkert P, Künzli M, Schwede T. QMEAN server for protein model quality estimation. Nucleic Acids Res. 2009 Jul;37(Web Server issue):W510–4.
56. Mirny LA, Finkelstein AV, Shakhnovich EI. Statistical significance of protein structure prediction by threading. PNAS. 2000 Aug 29;97(18):9978–83.
57. Wu S, Zhang Y. Protein structure prediction. In: Edwards D, Stajich J, Hansen D, editors. Bioinformatics: tools and applications [internet]. New York: Springer; 2009. [cited 2020 Nov 2]. p. 225–42. https://doi.org/10.1007/978-0-387-92738-1_11.
58. Battey JND, Kopp J, Bordoli L, Read RJ, Clarke ND, Schwede T. Automated server predictions in CASP7. Proteins: Structure, Function, and Bioinformatics. 2007;69(S8):68–82.
59. Fischer D, Rychlewski L, Dunbrack RL, Ortiz AR, Elofsson A. CAFASP3: the third critical assessment of fully automated structure prediction methods. Proteins: Structure, Function, and Bioinformatics. 2003;53(S6):503–16.
60. Wu S, Zhang Y. MUSTER: improving protein sequence profile–profile alignments by using multiple sources of structure information. Proteins: Structure, Function, and Bioinformatics. 2008;72(2):547–56.
61. Söding J. Protein homology detection by HMM–HMM comparison. Bioinformatics. 2005 Apr 1;21(7):951–60.
62. Roterman-Konieczna I. 9 - a short description of other selected ab initio methods for protein structure prediction. In: Roterman-Konieczna I, editor. Protein Folding in Silico [Internet], Woodhead publishing series in biomedicine. Woodhead Publishing; 2012. p. 165–89. [cited 2020 Nov 3]. Available from: http://www.sciencedirect.com/science/article/pii/B9781907568176500098.
63. Rohl CA, Strauss CEM, Misura KMS, Baker D. Protein structure prediction using Rosetta. In: Methods in enzymology [internet], Numerical computer methods, part D, vol. 383. Academic

Press; 2004. p. 66–93. [cited 2020 Nov 3]. Available from: http://www.sciencedirect.com/science/article/pii/S0076687904830040.

64. Padilla-Sanchez V, Gao S, Kim HR, Kihara D, Sun L, Rossmann MG, et al. Structure-function analysis of the DNA translocating portal of the bacteriophage T4 packaging machine. J Mol Biol. 2014 Mar 6;426(5):1019–38.
65. Kim H, Kihara D. Protein structure prediction using residue- and fragment-environment potentials in CASP11. Proteins. 2016 Sep;84(Suppl 1):105–17.
66. Kinch LN, Li W, Monastyrskyy B, Kryshtafovych A, Grishin NV. Evaluation of free modeling targets in CASP11 and ROLL. Proteins. 2016 Sep;84(Suppl 1):51–66.
67. Kryshtafovych A, Fidelis K, Tramontano A. Evaluation of model quality predictions in CASP9. Proteins. 2011;79(Suppl 10):91–106.
68. Baker D, Sali A. Protein structure prediction and structural genomics. Science. 2001 Oct 5;294 (5540):93–6.
69. Shin W-H, Christoffer CW, Wang J, Kihara D. PL-PatchSurfer2: improved local surface matching-based virtual screening method that is tolerant to target and ligand structure variation. J Chem Inf Model. 2016 Sep 26;56(9):1676–91.
70. Kihara D, Skolnick J. Microbial genomes have over 72% structure assignment by the threading algorithm PROSPECTOR_Q. Proteins: Structure, Function, and Bioinformatics. 2004;55 (2):464–73.
71. Trip JA, van Dam J, Eibergen R, Que GS. Investigations on correlations between serum enzymes and histological findings in liver disease. With special reference to transaminases and urocanase. Acta Med Scand. 1973 Feb;193(1–2):113–8.
72. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res. 2004 Nov 1;32(21):6226–39.
73. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W89–93.
74. Li B, Turuvekere S, Agrawal M, La D, Ramani K, Kihara D. Characterization of local geometry of protein surfaces with the visibility criterion. Proteins: Structure, Function, and Bioinformatics. 2008;71(2):670–83.
75. Kosinski J, Gajda MJ, Cymerman IA, Kurowski MA, Pawlowski M, Boniecki M, et al. FRankenstein becomes a cyborg: the automatic recombination and realignment of fold recognition models in CASP6. Proteins: Structure, Function, and Bioinformatics. 2005;61 (S7):106–13.
76. Shen M, Sali A. Statistical potential for assessment and prediction of protein structures. Protein Sci. 2006 Nov;15(11):2507–24.
77. Lu M, Dousis AD, Ma J. OPUS-PSP: an orientation-dependent statistical all-atom potential derived from side-chain packing. J Mol Biol. 2008 Feb 8;376(1):288–301.
78. Tosatto SC, Battistutta R. TAP score: torsion angle propensity normalization applied to local protein structure evaluation. BMC Bioinformatics. 2007 May 15;8:155.
79. Lüthy R, Bowie JU, Eisenberg D. Assessment of protein models with three-dimensional profiles. Nature. 1992 Mar;356(6364):83–5.
80. Lee M, Jeong C, Kim D. Predicting and improving the protein sequence alignment quality by support vector regression. BMC Bioinformatics. 2007 Dec 3;8:471.
81. Chen H, Kihara D. Estimating quality of template-based protein models by alignment stability. Proteins: Structure, Function, and Bioinformatics. 2008;71(3):1255–74.
82. Yang YD, Spratt P, Chen H, Park C, Kihara D. Sub-AQUA: real-value quality assessment of protein structure models. Protein Eng Des Sel. 2010 Aug;23(8):617–32.
83. Benkert P, Tosatto SCE, Schomburg D. QMEAN: a comprehensive scoring function for model quality assessment. Proteins: Structure, Function, and Bioinformatics. 2008;71(1):261–77.
84. Eramian D, Shen M, Devos D, Melo F, Sali A, Marti-Renom MA. A composite score for predicting errors in protein structure models. Protein Sci. 2006 Jul;15(7):1653–66.
85. Li J, Deng X, Eickholt J, Cheng J. Designing and benchmarking the MULTICOM protein structure prediction system. BMC Struct Biol. 2013 Feb 27;13:2.

86. Lundström J, Rychlewski L, Bujnicki J, Elofsson A. Pcons: a neural-network–based consensus predictor that improves fold recognition. Protein Sci. 2001 Nov;10(11):2354–62.
87. Wallner B, Elofsson A. Can correct protein models be identified? Protein Sci. 2003 May;12 (5):1073–86.
88. McGuffin LJ. Benchmarking consensus model quality assessment for protein fold recognition. BMC Bioinformatics. 2007 Sep 18;8:345.
89. Benkert P, Schwede T, Tosatto SC. QMEANclust: estimation of protein model quality by combining a composite scoring function with structural density information. BMC Struct Biol. 2009 May 20;9:35.
90. Wang Q, Vantasin K, Xu D, Shang Y. MUFOLD-WQA: a new selective consensus method for quality assessment in protein structure prediction. Proteins. 2011;79(Suppl 10):185–95.
91. Wallner B, Elofsson A. Identification of correct regions in protein models using structural, alignment, and consensus information. Protein Sci. 2006 Apr;15(4):900–13.
92. Ray A, Lindahl E, Wallner B. Improved model quality assessment using ProQ2. BMC Bioinformatics. 2012 Sep 10;13:224.
93. Cao R, Wang Z, Wang Y, Cheng J. SMOQ: a tool for predicting the absolute residue-specific quality of a single protein model with support vector machines. BMC Bioinformatics. 2014 Apr 28;15:120.
94. Cao R, Wang Z, Cheng J. Designing and evaluating the MULTICOM protein local and global model quality prediction methods in the CASP10 experiment. BMC Struct Biol. 2014 Apr 15;14:13.
95. Cao R, Bhattacharya D, Adhikari B, Li J, Cheng J. Large-scale model quality assessment for improving protein tertiary structure prediction. Bioinformatics. 2015 Jun 15;31(12):i116–23.
96. Liu T, Wang Y, Eickholt J, Wang Z. Benchmarking deep networks for predicting residue-specific quality of individual protein models in CASP11. Sci Rep [Internet]. 2016 Jan 14;6. [cited 2020 Nov 3]. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4725912/

Gene Ontology and Pathway Enrichment Analysis

12

Manoj Kumar Gupta, Gayatri Gouda, S. Sabarinathan, Ravindra Donde, Goutam Kumar Dash, Ramakrishna Vadde, and Lambodar Behera

Abstract

Over the past 10 years, gene set analysis has been the first option for studying gene expression and gene interaction for gaining insights into the fundamental dynamic biology of disease/traits. It, therefore, reduces the complexity of traditional statistical research and increases the illustrating strength of the outcomes achieved. Although approaches to gene set analysis are commonly utilized in gene expression analytics, the statistical framework and steps generally employed in these methods have not yet been thoroughly explored, restricting their usefulness. Thus, in this chapter, the authors include an outlined statistical framework and steps for the analysis of gene set used for various genome studies, ranging from microarrays, RNA sequencing, and the analysis of genomic widespread association results. The drawbacks of these approaches and strengths have also been addressed depending on their separate components such as their gene score, null hypotheses, and essential evaluation methods. The authors believe that a standardized approach for testing the methods of gene set analysis can also be used for correcting the lack of agreement on the method of preference for a specific experiment. The benchmark expression data sets will reflect actual expression data characteristics and prevent oversimplifying conclusions, such

M. K. Gupta · G. Gouda · R. Donde · G. K. Dash · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_12

as naturally distributed data with zero or constant gene-gene correlations. In the near future, information present in this chapter will be highly useful for underlying process with any disease or trait in more comprehensive way.

Keywords

Enrichment methods · Functional annotation · Gene ontology · Pathway enrichment analysis

Abbreviations

BP	Biological process
CC	Cellular component
DAG	Directed acyclic graph
FCS	Functional class scoring
GO	Gene ontology
GSA	Gene set analysis
GSEA	Gene Set Enrichment Analysis
HTBD	High-throughput biological data
MF	Molecular function
ORA	Over-representation analysis
PA	Pathway-based analysis
PTB	Pathway topology-based analysis
PTB	Pathway-based topology
SNP	Single-nucleotide polymorphism
SPIA	Signaling Pathway Impact Analysis
SRT	SNP-ratio research

12.1 Introduction

Traditional biological experiments typically research one or few genes at a time. In comparison, high-throughput sequencing approaches, like microarrays and RNA sequencing, produce huge lists of “interesting” genes as their final outputs. The analysis of these huge numbers of “interesting” gene lists (>100 or >1000 genes) involves complex process [1]. To address this problem, since the last two decades, various Gene Ontology [2] and pathway enrichment analysis [3, 4] approaches, such as Onto-Express [5], GoMiner [6], MAPPFinder [7], DAVID [8], GeneMerge [9], FunSpec [10], and FuncAssociate [11], had been developed. Since then, the sector of enrichment research has become extremely efficient, culminating in the availability of more related tools.

These tools are very close in that they often measure the *P-values* of pathway enrichment for a user-specific list of interesting genes, employing different statistical approaches, like the χ^2-test, binomial distribution tests, Fisher’s exact test, and the hypergeometric test. Fisher's exact test is optimal for the studying pathways that include a limited number of genes, and the χ^2-test is sufficient if gene count is larger than 5. Like Fisher's exact test, the hypergeometric distribution is used for sampling a few genes but is close to the binomial distribution when the gene numbers are high [12]. The binomial distribution tests are more suitable for evaluating a large number of genes, whereas the other three are valid for research using a limited number of genes. When most of the significant pathways are known, the variations between these statistical approaches are not dramatic [13, 14]. In 2014, employing both microarrays and RNA-seq results, Hong and the team demonstrated that a separate examination of up- and downregulated genes can identify more significant pathways, which are truly important for differentiating phenotypic in various cancers [12]. However, as no gold standard norm is there for retrieving significant biological information from any genes sets, recently, several researchers demanded designing of benchmark datasets, which, in turn, will represent the true existence of actual databases. If publicly accessible, these benchmarks will promote the assessment of gene analysis strategies more precisely and will facilitate the creation of new approaches. Thus, in the present chapter, the author attempted to understand the underlying principle of gene ontology and pathway enrichment analysis and how we can use this information to retrieve more significant biological information from the any gene sets, which, in turn, will help us to understand biological phenomenon more precisely.

12.2 Gene Ontology

The Gene Ontology (GO) project offers a standardized, regulated vocabulary of "term" or "class" that describes the roles of gene and equates them with their genetic products. It was designed primarily to illustrate biological processes via computational approaches. The development of strong conceptual information and appropriate tools has rendered the GO a very famous resource in biological science and a valuable resource for the study of computational data [15]. GO can be broadly categorized into three nonoverlapping ontologies, namely, biological process (BP), cellular component (CC), and molecular function (MF). The CC explains genetic products' positions in cells, e.g., the "nucleus," in subcellular systems. The MF defines actions at the molecular level, like "catalytic behavior." The terms MF correspond to activities, instead of molecules, that carry out the operations and do not define when or where the activity occurs. The BP contains terminology that characterizes numerous events in order through assemblies of one or more MF like "signal transduction". It might not be simple to discern between molecular and biological processes, but in general, there must be more than one distinct phase in a biological method. Note the GO words theoretically describe genes rather than pathways since they are not gene dependencies. GO is also not a gene sequence database, and instead of gene product, it contains the properties of gene products [16].

12.2.1 Molecular Function Annotation

Under most basic biological scenario, almost every biomolecule m is an instance of a category M that has certain ability to act as an instance of the molecular function type F (depicted through a corresponding GO term). For instance, alcohol dehydrogenase 1, Adh1's gene product, functions as an instance for the alcohol dehydrogenase molecules. This implies that an alcohol dehydrogenase 1 has the capability to play its role in a specific way. In this way, the word "action" is intended to be used in a biochemical framework and is best interpreted as representing "potential activity." Notice that while the "alcohol dehydrogenase" term is used both by the name of the gene and in the molecular function, the term itself refers to various entities: in the former, it refers to the type of molecules, in the latter, it refers to the type of function which the molecule may perform. This complexity is embedded in the propensity to label molecules depending on their performing roles, and it is highly required to distinguish this differences, since the molecules name and the function upon which molecule is named may not generally match, for example, because the molecule may perform many functions [17]. However, if certain gene product class has the ability to perform a certain role, this does not imply that every incident of this molecule type necessarily executes that same function. For instance, the molecules of the *Zp2* mouse gene are located in the oocyte and are likely to bind the molecules of the *Acr* gene during fertilization [18]. However, when an oocyte is not fertilized, the molecules are still present and they are vulnerable to performing the binding role, but never performed.

12.2.2 Biological Process Annotation

A MF instance is a gene product instance's ongoing ability to function in a certain manner [17]. A BP is the execution of one or more such molecular functional instances that act together to fulfill a biological goal. At the cellular or organismal level, a BP instance is how a function operates at the level of the molecule. There is a connection between MF and BP. However, at present, this relationship is not clearly expressed in GO. From a viewpoint of genetic annotation, we want to move beyond the connections between cellular and organism instances and to establish the capacity to infer type-type relationships that incorporate genetic component forms at the molecular level of granularity with process types at the cell or organism level. It is worth noting that molecules of one gene product class may be correlated with instances of a molecular function type (known or unknown) whose success leads to the incidence of a specific biological phase. Information may be produced on certain type ratio when studies are structured to investigate what occurs when defined biological parameters are encountered in normal circumstances—circumstances in which disturbing incidents do not interfere because of the experimenter's efforts. Experiments are structured to be reproducible and predictive, representing instances that are supposed to follow the specified conditions in biological systems. If future studies reveal that previous experiments did not describe the expected typical scenario, the results of the previous experiments are disputed, re-examined, and reinterpreted, or even completely denied, and the annotations relating to that need to be changed accordingly [17]. Such annotations often point to errors in the standard relationships defined in an ontology. Annotations thus point to errors in the type relationships defined in ontology. An example is the recent elimination of the type serotonin secretion as a child of the neurotransmitter secretion from the GO BP. This change was rendered as a consequence of a paper annotation, which indicates that serotonin may be secreted by immune system cells where it does not function as a nervous transmitter [17].

12.2.3 Cellular Component Annotation

In most instances, a correlation between the gene product and the cell component location is rendered on the basis of a direct examination of a cellular component observed under the microscope [17]. For instance, in one study [19], which reports an experiment that uses a genetic antibody recognized in the *Atp1a1* gene to mark the location of its products within preimplantation mouse embryos. The fluorescent staining indicates that the gene product is found on the plasma membrane of the embryo cells. In this case, the gene products are the molecules connected by fluorescent antibodies, and the location of the CC is the plasma membrane as seen under the microscope. The findings of this experiment were used by a curator to annotate the GO CC of the *ATP1A1* gene product as a plasma membrane. As in MF and BP, the MF and CC also have a connection among themselves. Thus, it is simple to conclude that if a molecule of a gene product is present in a cell component instance, then the gene product will perform its role in that cellular component as well [17]. These molecules often perform their work in such a way that these operations become biologically important. As MF and BP are often separable, experimental evidence for each annotation is often not the same. Thus, from a functional perspective, both ontologies must be built separately [17].

12.3 GO Structure and Data Representation

A significant characteristic of GO is that its terms are structured, i.e., genes with similarly associated features are categorized in the same group [16]. Each subontology, i.e., MF, BP, and CC, organizes their terminology as a directed acyclic graph (DAG), basically a hierarchical system that allows several parent terms to be used for each term in the child. The bond between child and parent can be ‘is_a” or “part_of.” The lower a term in the DAG is, the more descriptive it is. This helps the biomedical researcher to find the most appropriate uniform descriptors for the functions of a single gene element, depending on available knowledge. When a gene or gene product is annotated with GO, the GO DAG will demonstrate how the annotations inside a subontology refer to each other and to the annotations of other genes compositionally [20].

In fact, there are several published methods that exploit the structure, which can be narrowly divided into two groups: ones that are specialized for capturing the local graph properties into account and ones that are designed for making the global graph structures. In the former, we have several current GO resources that use a limited subset of words to illustrate some nuances of the hierarchical relationships between terms. One of the most commonly used tools under this category is QuickGO [21], which enables us to query and search a single GO term with the similar term (and annotations) as DAG. It is intended to provide convenient access to the electronic and manual GO annotations given by the GO Consortium annotation communities. Populating a collection of GO terms can be done using a broad variety of GO features, which can be sorted through hierarchical taxonomic values and genomic annotations. The user then chooses one or more GO terms of interest, and the results of all these terms in the selected taxonomic group appear into a single table format. The user may also download the data as one browses the GO results and modify the data presentation or loading. REVIGO [22] is another tool that summarizes long, generic, complicated lists of GO words by finding a representative subset of terms using a simple algorithm that uses a relatively semantic similarity measure. REVIGO was developed by integrating manual and automated processes that create new information about the relationship of certain biological features with neuronal and behavioral functions. One of the important functions of REVIGO, in addition to providing a valid output to an underlying problem of how knowledge and semantic function in the cell or organisms become manifests in data, is to give a visual representation of the underlying ongoing trends in a large dataset of modern cell biology. Additional tools provide displays, where nodes and connections can be obtained with more immersive functionality [23, 24] or where node enrichment ratings can be illustrated with enhanced customization [25, 26].

In the latter, a few visualization tools do not tend to vividly illustrate all the Go functions [27, 28], considering detailed literature on displaying large graphs or network features [28–30]. These types of graph displays, as opposed to the tiny graph displays, are normally not flexible enough to highlight the node or the link-specific information owing to many visual elements. These kinds of tools may offer a global view of the overall structure of a graph, such as clusters or entire hierarchies, which are particularly useful for understanding patterns, outliers, or the overall structure of a graph. Recently developed new open-source software, namely, AEGIS [31], aims at bridging the merits of all visual approaches and facilitates robust connections with knowledge encrypted in an ontology. Using the rules of a classic visual information system, we first look at the whole graph, then zoom-in on individual nodes (or pieces of information), and process them through a set of filters, and then when we are satisfied, we look at the particulars of the information. AEGIS is based upon infrastructure that uses a pipeline of instantaneous pictures for the reconstitution of dynamic maps into overlapping, interactive displays. Users can even manipulate the interface to visualize the results of existing pipelines, and they can also select any vantage point to explore the experiment with their visualization before collecting data or running pipelines. The user can also manipulate the graph in several other ways to gather many other data. During the exploratory phase, AEGIS helps them to become informed of biological knowledge that is important for simulations and hypothesis generation and measure power needed for study design [31].

12.4 Pathway Enrichment Analysis

The pathway-based analysis (PA) overcomes the drawbacks found with other single-locus research approaches. The end product of PA provides a thorough understanding of the mechanism underlying complex diseases [32]. Principally, a PA is similar to the GO analysis [33]. However, the PA is more descriptive and detailed; it also measures the interaction of a pathway with a disease phenotype. Its ability to address biological interactions among genes and provide power and robustness has been well-recognized [13, 34]. The early use of PA was demonstrated during microarrays analysis [32, 35], which was specifically expanded from the Gene Set Enrichment Analysis (GSEA) [36, 37]. Now, PA analysis is used in a wide range of research studies, including gene set analysis (GSA) [38] and SNP-ratio research (SRT) [39]. Methods employed for pathway analysis can be broadly classified as over-representation analysis (ORA), functional class scoring (FCS), and pathway topology-based analysis (PTB) (Table 12.1).

12.4.1 Over-representation Analysis

In an ORA, the basic principle is that it is possible to classify the appropriate pathways if the proportion of differently expressed genes, inside the defined pathway, exceeds the proportion of genes that would be randomly expected [82]. In this context, ORA approaches operate along the key workflow in which the fraction of the pathways contained in the collection of biological components chosen by the user is evaluated (Fig. 12.1). This list normally follows some requirements, usually log fold transition, statistical value, or both, rating and cutting off most sections of an original list, for instance, all genes that have been examined in a microarray. Then, the confidence value is determined using statistical techniques, such as the hypergeometric distribution, the chi-square, the binomial, or Fisher's exact test, etc. Additional correction is normally carried out with several tests, because data evaluation simultaneously (in this case, pathways) for many hypotheses may lead to false-positive findings. The end outcome of an ORA approach typically consists of a *p*-value and/or a multiple-hypothesis-test-corrected *p*-value of the most important pathways [82]. The key benefits of utilizing ORA methodologies relative to nonknowledge (i.e., solely data-driven) analyses are the biological background of omic data, which makes the formulation of hypothesis and eventual experimental research. Hence, it quickly becomes an information creation loop that is compatible with the Systems Biology method. GoMiner [6] is one of the most cited ORA techniques and was developed to interpret the gene expression microarray results. It requires a list of over- and underexpressed genes plus the entire setlist of the

Table 12.1 Approaches for pathway enrichment analysis (Adapted from [40])

Approaches	Methods	Advantages	Limitation	Software packages/tools
Over representation analysis	Hypergeometric distribution/ Fisher's test, binomial distribution, chi-square distribution, etc.	• Easy to perform. • Allots simple interpretable values like *p*-values for the whole gene set.	• Highly reliant on a threshold/ cutoff value, which is impossible to assess at the discretion of the user. • Statistical evaluation independent of the differential expression score of chromosomes. • Using only the most important hard threshold-based genes and discarding others contribute to information loss. • Assumes that each gene leads to the phenotype/trait similarly. • Assumes that each gene is separate and avoids the gene association or duplication in the gene collection. • Assumes that each predefined gene collection, which is incorrect, is separate from others.	AgriGO [41], GenMAPP [42], Onto-Express [5], GoMiner [6], DAVID [8], GOstat [43], FuncAssociate [11], GOToolBox [44], FatiGO [45], GOEAST [46], ClueGO [25], FunSpec [10], GeneMerge [9], GARBAN [47], GO: TermFinder [48], WebGestalt [26], GOFFA [49], WEGO [50], GOTM [51], GSAQ [52], Pathview [53], Wholepathwayscope [54], and ShinnyGO [55]
Functional class scoring	Wilcoxon signed-rank test, median, sum, or mean, of the gene-level statistic(s), and max-mean statistic	• No requirement of threshold/ cutoff value to divide gene space into different nonselected and selected part. • Study dependence among genes within the gene set. • The test statistic is dependent on the differential gene	• Each gene set is studied separately. • Consider only the number of genes in a gene set (pathway to executing GSA) but disregard the additional details accessible at the bioknowledge basis. • If the predefined gene sets	SAFE [56], GSA [57], GSEA [35], Random set [58], sigPathway [59], GlobalTest [60], SAM-GS [61], LIMMA [58], Catmap [62], T-profiler [63], FunCluster [64], GeneTrail [65], Gazer [66], GSAQ [52], CAMERA [67],

		enrichment score of genes within the gene set.	have become mutually exclusive, however, in genetics, these gene sets have overlapping consequences.	GAGE [68], SGSE [69], PAGE [70], GSNCA [71], GSA-SDR [72], GenePattern [73], plantGSEA [74], and GSAR [75]
Pathway-based topology	Graph/network theory	• Consider the relation between genes and experimental condition modification • Consider gene or pathway topology or • Consider the topology of the pathways/genes during modelling.	• Depending on the form of cell owing to the cell-specific GE profiles and the seldom accessible condition being investigated. • It is not as common as it needs rarer knowledge and more intense computing. • Interactions between gene sets cannot be considered (pathways). Much based on annotations.	PathwayExpress [76], ScorePAGE [77], SPIA [78], NetGSA [79], TopoGSA [80], and CliPPER [81]

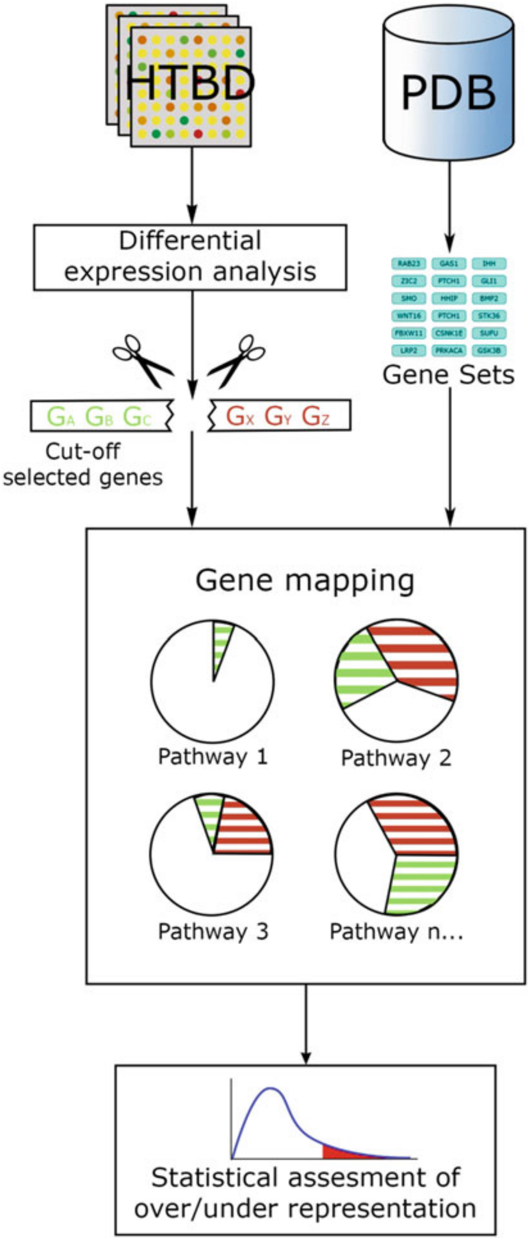


Fig. 12.1 Generalized workflow for over-representation analysis. (Adapted from [82])

microarrays and then measures over-representation and under-representation for the gene ontology groups using Fisher's exact test. WebGestalt [26], a Web-based tool first released in 2005, but regularly revised, is another example of ORA (last update in 2017). It operates by transforming the ORA into a user-friendly integrated interface with a range of central public PDBs. The approach thus allows an interpretation of data on multiple biological contexts, such as metabolic, gene-phenotype, gene-disease, gene-drug mixture.

Although ORA rapidly classifies large datasets, these methods have many limitations. Because of the cutoff approach chosen by the user, these methods set a large number of basal level records. The study also omits theoretically significant components near the cutoff threshold. This would also have an effect on stability, as there is no thumb law to define the cutoff threshold, since multiple cutoff strategies show contradictory outcomes [83], and the choosing of cutoff thresholds is subjective. They assess every aspect of the pathway, give them equivalent weight or significance, and discard all information (e.g., position in the pathway, gene expression level, and interaction among genes) inherent in the interactions. This often results in a study of two pathways with the same genes but separate topologies [84]. They also believe that pathways are separate, contrary to the awareness of association and overlaps between pathways [85]. These drawbacks lead to the development of second generation tool for PA.

12.4.1.1 Functional Class Scoring

The key principle for these approaches is that not only major variations in genetic expression have important consequences on a pathway, but also less coordinated changes in genes assembling the pathway have an influence on the general pathway condition. Thus, FCS methods use all the metrics available in high-throughput biological data (HTBD) for calculating their enrichment scores, eliminate the ORA cutout limit, but then use pathways as gene sets to conduct their calculations. Basically, each FCS system operates in a three-stage workflow (Fig. 12.2). A baseline statistic using all HTBD, estimating differential expressions of individual components, is determined. The most important baseline statistics used in PA are fold change, *t*-statistic, log-likelihood ratio, and signal-to-noise ratio. If the sample is limited, basic versions of these test statistics are employed. The baseline statistics of each route are subsequently aggregated into a single pathway-level statistic, like the Kolmogorov-Smirnov statistic, Wilcoxon sum rank statistic, max-mean statistic, and χ^2 (chi-squared) test. The predictive validity of pathway-level data is essentially measured on the basis of the null hypothesis chosen. The fundamental benefits of FCS approaches are that they do not require an arbitrary cutoff level of differential genes, and they utilize all available information. They can distinguish variations between pathways that barely pass the differentially articulated thresholds and those that pass across them to multiple magnitude stages. They can identify slight yet coordinated interactions between molecular gene expression and their pathways. Some approaches may often classify the most important genes in a specified way; GSEA, for example, considers these genes as the center of the pathway [82].

One of the first and best-known methods of FCS is the Gene Set Enrichment Analysis (GSEA) [35] developed for microarray data gene expression analysis. In short, the genes are classified according to their distinctive gene expression among two phenotypical groups (by signal-to-noise ratio basal-level statistic). Then, their distribution is assessed by a given gene set (e.g., MSigDB gene sets), and an

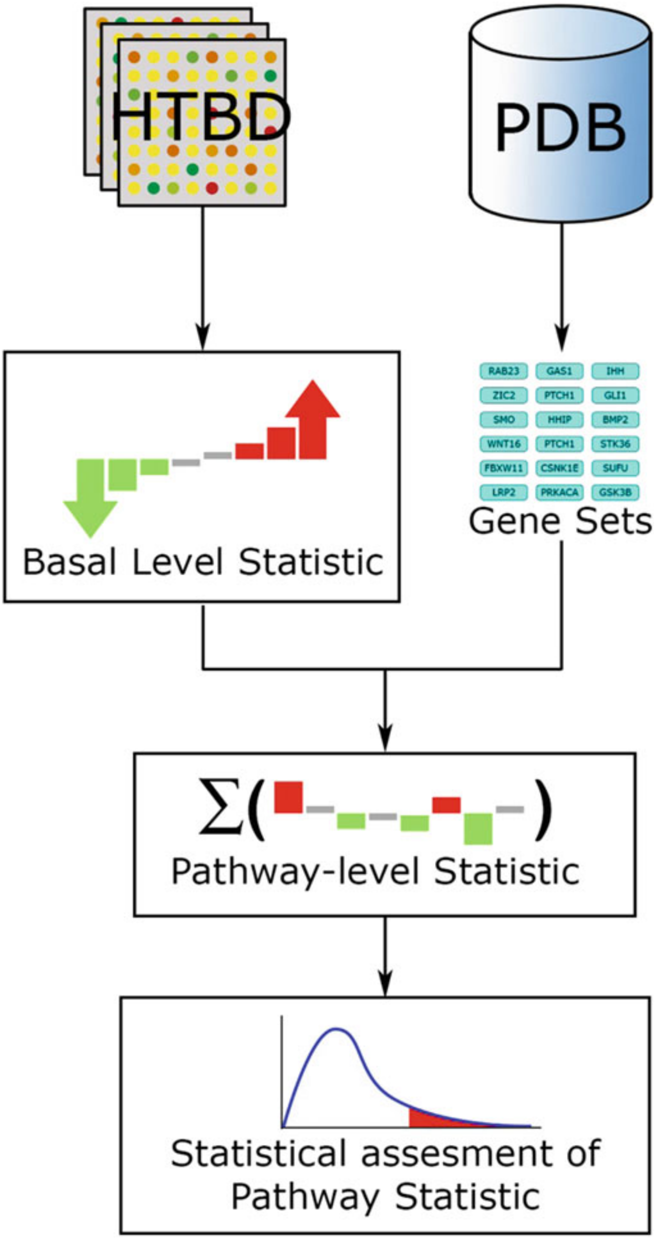


Fig. 12.2 Generalized workflow for functional class scoring. (Adapted from [82])

enrichment score (ES) is defined for each set of genes (through a Kolmogorov-Smirnov pathway-level statistic). The importance of the ES and correction for multiple testing hypotheses would then be evaluated. Earlier, Folger and the team used GSEA methodology as an illustration of the methodology to verify the proliferating function of growth-enhancing genes, which have been forecast to be of importance in cancer treatment [86]. This research stresses the usage of an alternative gene collection of some PDB derived from the Luo and collaborators' shRNA screening in 12 cancer cell lines [87]. In particular, the usage of external pathway data is mainly utilized in ORA and FCS methodologies because topology information is difficult to determine and derive primarily from PDB's multitude of expertise. However, knowledge regarding the cell cycle and replication can already be sought via most PDBs such as KEGG and Gene Ontology.

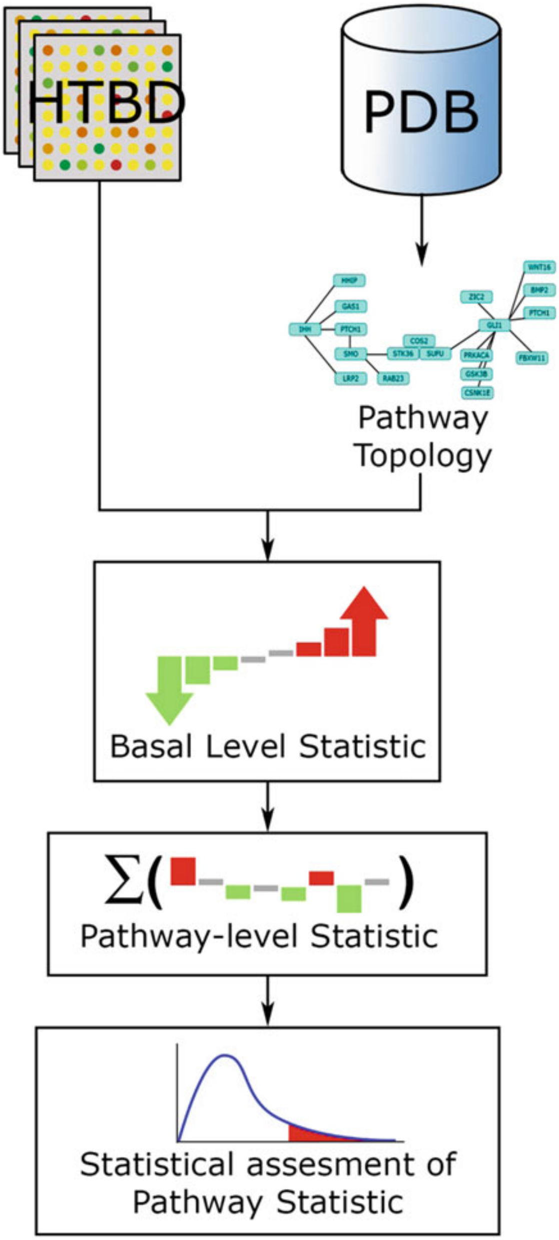
Given the fact that FCS analyzes addressed ORA shortcomings, they do have issues. The usage of pathways as gene sets and not as networks is primarily accountable. Examples of those limits are as follows: Most of the components in the pathway also provide the same weight to decide the path statistics, as GSEA did in its first publication, regardless of previous awareness of the pathway [84]. The knowledge from several PDBs remains unused since these approaches do not take into consideration the interactions between pathway components and other information about the network configuration of pathways. This may contribute to reduced identification of the related paths [88].

12.4.2 Pathway-Based Topology

Following developments in pathway annotation from PDBs, the topology of the underlying pathway networks was released through various databases following their immediate incorporation with PA methodologies. This was strongly promoted by the expanded focus network theory [89, 90]. The key theory for pathway-based topology (PTB) research is that pathway-topology relationships annotated in PDBs provide details for understanding of associated shifts in pathway components. PTB methods may be called extensions of the ORA and FCS methods since they are typically implemented in the same general measures, but they include topology of the pathways to determine their statistical significance. In the case of expanded ORA systems, the pathway topology maps user-chosen genes and the resulting network and statistical study were carried out. In the case of expanded FCS approaches, HTBD and topology are used to measure the base level statistics and to continue in a similar manner to the FCS methodology (Fig. 12.3).

Through evaluating network routes, the PTB research tackles ORA and FCS method shortcomings, as such, on the basis of topology data, biologically important component variations may be taken into account by increasing the effect of shifts in genes with a greater impact on the pathway. By taking different topological details into account, they make a more exact study of the same group of pathway elements, as certain interactions under some biological conditions are known to be different. Account for causal interactions in the networks, when shifts in the upstream elements are supposed to affect the behavior. Pathways Express as part of the Onto-Tool Suite was one of the first tools for the study of pathway topology [88, 91]. Driven by the sense that further detail becomes accessible, the query "Is

Fig. 12.3 Generalized workflow for pathway-based topology. (Adapted from [82])



there a known pathway containing my gene(s) of interest?" would convert into "How do I find the most interesting pathway(s) involving my gene(s)?" The system computes two probabilistic parameters, the gene perturbation factor that is identical to that of Google's Page-Rank index (http://ilpubs.stanford.edu:8090/422/), taking into account the uniform fold shift of each input gene and the sum of gene perturbation downstream, representing the relevance of each gene controlled by differentiation, finally, a collection of pathways according to their effect factor and their adjustment of several assumptions, as seen by their incorrect rate of detection. The same evolving Pathway-express team recently created the "Signaling Pathway Impact Analysis" (SPIA) [78], which is an updated variant of the first. SPIA attempts to assess and integrate two kinds of facts to ensure the independence of each probability: (1) the over-representation of differentially expressed genes in a given pathway and (2) the irregular disruption of the pathway, determined by shifts in propagating expression across the pathway topology. PARADIGM [92] is another PTB technique that has been evolved to incorporate different omic datasets to deduce paths modified in a patient-specific or sample-specific maker. It utilizes a probabilistic graphical model system to learn the underlying causal networks in conjunction with the observations made by HTBD.

PTB approach constraints are challenging to overcome since they portray a transition in one of the existing life science paradigms. Life elements do not work individually. However, they act specifically to carry out life tasks and their action as an entire mechanism is complex, adaptable, and resilient [93]. Existence is, in short, a dynamic framework. However, some drawbacks of PTB methods can be recognized that would definitely be dealt with in future methods, as obstacles to studies and annotations are overcome. Any PTB strategies do not understand the way the pathways are related. The deregulation chain consequences may be skipped [78]. For the better identification of relevant routes, the PTB methods do not take into consideration the interconnections among pathways. Take for example, "pathway A" downstream components may be upstream in "pathway B," and thus, path A is supposed to have an effect on path B [94]. Time and spatial distribution of pathway components in their models is not taken into account. Pathway behavior, for example, transcriptional control within the nucleus, protein transport within the endoplasmic reticulum, and mitochondrial-mediated signaling, can rely on biomolecule compartmentalization. Regular pathways such as BioPAX, SBGN, and SBML now endorse compartments in their ontology pathways [95]. In addition, time-scale molecular control is also critical for understanding the mechanisms through which pathways function in cells. As equipment becomes cheaper, experimental costs will decrease and timescale will increase the need for better PA instruments that can interpret these data [96]. The majority of approaches cannot consider a pathway component's various states and variants. Most PA techniques, for example, collapse splicing variants from gene expression data into a single HGNC gene symbol. Thus, the inclusion of this information will help us to understand the single nucleotide polymorphism (SNP), splicing variations, epigenetic changes, and post-translation changes, as well as their probable effect on the working of phenotype and route in a more comprehensive way [84].

12.4.2.1 Pitfalls in Gene Ontology Analysis

There are several approaches for analyzing genes with no agreement about best practices available. In the absence of gold standard datasets, the key problems in gene set research are lack of reproductiveness. Gold standard datasets may help us to identify and solve these problems more effectively. Although, to date, several approaches have been developed, nearly all approaches proposed are incomplete [97]. For certain gene set analysis, whether competitive or autonomous, the analysis findings have been found to be unreproducible for limited sample sizes [98]. However, regardless of this problem, limited sample size studies ($n < 5$ per group) are still analyzed using these techniques [99]. The size of a dataset is a significant factor when deciding on a reasonable approach for the gene set analysis or whether gene set analysis is sufficient. Also, their vulnerability to sample size should be examined in the validation phase while designing new gene analysis methods. The evaluation of gene set analytical methods has become a significant research area [100, 101]. Methods for gene set analysis are analyzed on the basis of actual and virtual expression datasets. True datasets with supposed enrichment status are widely used for the assessment of methods for the gene set study [102]. Unfortunately, enrichment status predictions of the gene sets cannot be justified with certainty. This ambiguity in gene enrichment status also contributes to uncertainty in the outcomes of the evaluation.

Because of the absence of gold standards for the assessment of gene set analysis techniques, simulated expression datasets have been used [103]. These datasets were produced with regularly distributed values with constant means and standard deviations. Furthermore, these simulated datasets either presume no gene-gene association [104] or continuous associations between genes [103] in genetic classes. However, expression data never adopt a normal distribution in operation. In addition, gene-gene interaction is considered to have a significant influence on the effects of enrichment analytical approaches in actual expression evidence [105]. These super-simplifications may lead to judgments that support some methods of study of genes. Ackermann and Strimmer [103] simulated expression datasets utilizing a standard multivariate distribution with variances of 1. They simulated noninformative gene expression significance using a standard multivariate normal distribution. They modeled enhanced genetic sets with constant shifts in mean expression and constant gene-gene associations. As noninformative gene expression values that constitute a majority of a dataset followed typical multivariate standard distribution, competitive methods and parametric methods were easily able to detect gene sets' enrichment status. This results in a skewed judgment in favor of these approaches. Often, distributed values of continuous mean and standard deviation do not realize the variance variability in high-performance results [106]. Gene set collections were also simulated as a limited number of nonoverlapping, equal-size sets, which vary considerably from actual gene-set databases. The assessment of the gene set study utilizing these datasets led to contradictory and contrary findings because of oversimplifying assumptions [107].

The explanation is that the inclusion of unrelated genes modifies the distribution of context genes. Compared to randomly assembled gene sets of the same scale as Gi, the importance of the gene set score S(Gi) is determined using competitive methods. The inclusion of unrelated genes enhances the disparity between S(Gi) and the randomly assembled gene sets as nonrelated genes frequently have a poor, nonconcordant pattern of expression. They also report that GAGE, a nonparametric process, achieves higher power if the expression dataset is augmented with unrelated genes. This can also be demonstrated by the way GAGE measures their gene set ratings, based on the discrepancy between average gene expression values in the gene set and typical gene expression values. By incorporating unrelated genes, sometimes displaying smaller average expression values, there are more drastic gene set scores that contribute to a deceptive increase in intensity. The usage of strategic strategies such as GSEA (with gene sampling) and GAGE is highly discouraged by Tripathi and the team [108].

12.5 Conclusion and Future Perspective

In conclusion, since no gold standard datasets are available for performing GO and PA analysis of gene set [104], biological and technological heterogeneity cannot be modeled effectively using available oversimplified methods and tools. While testing current and established gene set research approaches, a crucial move is to synthesize datasets that maintain the true essence of gene expression data and gene set databases. Specifically, designing benchmark datasets representing the true existence of actual databases can be of significant significance in evaluating established and current approaches for gene analysis. There is currently no such standard, and we advocate creating these public metrics as potential studies. If publicly accessible, these benchmarks will promote the assessment of gene analysis strategies and facilitate the creation of new approaches. A significant factor in improving gene set analysis techniques is their capacity to cope with gene set overlaps, which has led to the loss of precision of such approaches [109]. Present methods aimed at resolving the vulnerability of gene clash sacrifice and thereby introduce false-negatives. Active research areas [110] are evolving methods to converge and hit high precision without losing sensitivity and continue to be an avenue for potential studies.

Even virtual gene set databases composed of nonoverlapping gene sets of equal sizes were used when evaluating gene set analysis techniques. Such an environment lacks the true existence of the gene set databases with multiple gene set overlaps and varying gene set volumes, stated to influence the effects of gene set studies [111]. Earlier, researchers have firmly opposed the usage of such artificial gene databases in order to test methods of gene set research in a practical sense [97]. If simulation expression databases are used, it is proposed that actual gene names/IDs are explored in the simulated expression details. This enables actual gene set databases to be used alongside virtual expression results. This small move could illustrate the actions of a system in the assessment of gene overlapping and various gene set sizes. Earlier, Tripathi and the team have demonstrated that competitive methods of gene set research are susceptible to the presence of unrelated genes [108]. We recommend that we adopt the guidance given by Tripathi et al. when implementing competitive gene analytical methods [108]. In addition, modern methods for the study of geneset should be developed to be resilient against shifts in the distribution of the history because of unrelated genes. In addition, the outcome of the gene set analysis approach can influence different up and down distributions in the gene sets, various gene sets, various differential expression levels, different sample sizes, and an imbalanced number of samples by each community [112]. Therefore, we suggest every effort to test or establish methods for the study of genes. Another path for potential studies is a comprehensive review for selecting a best fitting gene database before the gene set examination.

Conflict of Interest None

Additional Information Figs. 12.1–12.3 (CC BY 4.0) [82] have been used under the terms of the Creative Commons Attribution License.

References

1. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1–13.
2. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Nat Genet. 2000;25(1):25–9.
3. Ulgen E, Ozisik O, Sezerman OU. pathfindR: an R package for comprehensive identification of enriched pathways in omics data through active subnetworks. Front Genet [Internet]. 2019 [cited 2020 Dec 22];10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00858/full.
4. Paczkowska M, Barenboim J, Sintupisut N, Fox NS, Zhu H, Abd-Rabbo D, et al. Integrative pathway enrichment analysis of multivariate omics data. Nat Commun. 2020;11(1):735.
5. Khatri P, Draghici S, Ostermeier GC, Krawetz SA. Profiling gene expression using onto-express. Genomics. 2002;79(2):266–70.
6. Zeeberg BR, Feng W, Wang G, Wang MD, Fojo AT, Sunshine M, et al. GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol. 2003;4(4):R28.
7. Doniger SW, Salomonis N, Dahlquist KD, Vranizan K, Lawlor SC, Conklin BR. MAPPFinder: using gene ontology and GenMAPP to create a global gene-expression profile from microarray data. Genome Biol. 2003;4(1):R7.
8. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res. 2007;35:W169–75.
9. Castillo-Davis CI, Hartl DL. GeneMerge—post-genomic analysis, data mining, and hypothesis testing. Bioinformatics. 2003;19(7):891–2.
10. Robinson MD, Grigull J, Mohammad N, Hughes TR. FunSpec: a web-based cluster interpreter for yeast. BMC Bioinform. 2002;3(1):35.
11. Berriz GF, King OD, Bryant B, Sander C, Roth FP. Characterizing gene sets with FuncAssociate. Bioinformatics. 2003;19(18):2502–4.
12. Hong G, Zhang W, Li H, Shen X, Guo Z. Separate enrichment analysis of pathways for up- and downregulated genes. J R Soc Interface. 2014;11(92):20130950.
13. Curtis RK, Orešič M, Vidal-Puig A. Pathways to the analysis of microarray data. Trends Biotechnol. 2005;23(8):429–35.
14. Khatri P, Drăghici S. Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics. 2005;21(18):3587–95.
15. Gaudet P. The gene ontology. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of bioinformatics and computational biology. Oxford: Academic Press; 2019. p. 1–7. http://www.sciencedirect.com/science/article/pii/B9780128096338205001.

16. Holmans P. 7—Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. In: Dunlap JC, Moore JH, editors. Advances in genetics, Computational methods for genetics of complex traits, vol. 72. Oxford: Academic Press; 2010. p. 141–79. http://www.sciencedirect.com/science/article/pii/B9780123808622000072.
17. Hill DP, Smith B, McAndrews-Hill MS, Blake JA. Gene ontology annotations: what they mean and where they come from. BMC Bioinform. 2008;9(Suppl 5):S2.
18. Howes E, Pascall JC, Engel W, Jones R. Interactions between mouse ZP2 glycoprotein and proacrosin; a mechanism for secondary binding of sperm to the zona pellucida during fertilization. J Cell Sci. 2001;114(Pt 22):4127–36.
19. MacPhee DJ, Jones DH, Barr KJ, Betts DH, Watson AJ, Kidder GM. Differential involvement of Na(+),K(+)-ATPase isozymes in preimplantation development of the mouse. Dev Biol. 2000;222(2):486–98.
20. Myhre S, Tveit H, Mollestad T, Lægreid A. Additional gene ontology structure for improved biological reasoning. Bioinformatics. 2006;22(16):2020–7.
21. Binns D, Dimmer E, Huntley R, Barrell D, O'Donovan C, Apweiler R. QuickGO: a web-based tool for gene ontology searching. Bioinformatics. 2009;25(22):3045–6.
22. Supek F, Bošnjak M, Škunca N, Šmuc T. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One. 2011;6(7):e21800.
23. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. 2003;13(11):2498–504.
24. Sealfon RS, Hibbs MA, Huttenhower C, Myers CL, Troyanskaya OG. GOLEM: an interactive graph-based gene-ontology navigation and analysis tool. BMC Bioinform. 2006;7(1):443.
25. Bindea G, Mlecnik B, Hackl H, Charoentong P, Tosolini M, Kirilovsky A, et al. ClueGO: a Cytoscape plug-in to decipher functionally grouped gene ontology and pathway annotation networks. Bioinformatics. 2009;25(8):1091–3.
26. Wang J, Vasaikar S, Shi Z, Greer M, Zhang B. WebGestalt 2017: a more comprehensive, powerful, flexible and interactive gene set enrichment analysis toolkit. Nucleic Acids Res. 2017;45(W1):W130–7.
27. Pareja-Tobes P, Tobes R, Manrique M, Pareja E, Pareja-Tobes E. Bio4j: a high-performance cloud-enabled graph-based data platform. bioRxiv. 2015;016758.
28. Heberle H, Carazzolle MF, Telles GP, Meirelles GV, Minghim R. CellNetVis: a web tool for visualization of biological networks using force-directed layout constrained by cellular components. BMC Bioinform. 2017;18(10):395.
29. Merico D, Gfeller D, Bader GD. How to visually interpret biological data using networks. Nat Biotechnol. 2009;27(10):921–4.
30. van Ham F, Perer A. "Search, show context, expand on demand": supporting large graph exploration with degree-of-interest. IEEE Trans Vis Comput Graph. 2009;15(6):953–60.
31. Zhu J, Zhao Q, Katsevich E, Sabatti C. Exploratory gene ontology analysis with interactive visualization. Sci Rep. 2019;9(1):7793.
32. Wang K, Li M, Bucan M. Pathway-based approaches for analysis of Genomewide association studies. Am J Hum Genet. 2007;81(6):1278–83.
33. Holmans P, Green EK, Pahwa JS, Ferreira MAR, Purcell SM, Sklar P, et al. Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. Am J Hum Genet. 2009;85(1):13–24.
34. Tilford CA, Siemers NO. Gene set enrichment analysis. In: Nikolsky Y, Bryant J, editors. Protein networks and pathway analysis, Methods in molecular biology. Totowa: Humana Press; 2009. p. 99–121. https://doi.org/10.1007/978-1-60761-175-2_6.
35. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
36. Jia P, Wang L, Meltzer HY, Zhao Z. Common variants conferring risk of schizophrenia: a pathway analysis of GWAS data. Schizophr Res. 2010;122(1):38–42.

37. Yang W, Wang J, Liu L, Zhu X, Wang X, Liu Z, et al. Effect of high dietary copper on somatostatin and growth hormone-releasing hormone levels in the hypothalami of growing pigs. Biol Trace Elem Res. 2011;143(2):893–900.
38. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, et al. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. Nucleic Acids Res. 2009;37(Suppl_2):W340–4.
39. O'Dushlaine C, Kenny E, Heron EA, Segurado R, Gill M, Morris DW, et al. The SNP ratio test: pathway analysis of genome-wide association datasets. Bioinformatics. 2009;25 (20):2762–3.
40. Das S, McClain CJ, Rai SN. Fifteen years of gene set analysis for high-throughput genomic data: a review of statistical approaches and future challenges. Entropy. 2020;22(4):427.
41. Tian T, Liu Y, Yan H, You Q, Yi X, Du Z, et al. agriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res. 2017;45(W1):W122–9.
42. Dahlquist KD, Salomonis N, Vranizan K, Lawlor SC, Conklin BR. GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. Nat Genet. 2002;31(1):19–20.
43. Beißbarth T, Speed TP. GOstat: find statistically overrepresented gene ontologies within a group of genes. Bioinformatics. 2004;20(9):1464–5.
44. Martin D, Brun C, Remy E, Mouren P, Thieffry D, Jacq B. GOToolBox: functional analysis of gene datasets based on gene ontology. Genome Biol. 2004;5(12):R101.
45. Al-Shahrour F, Díaz-Uriarte R, Dopazo J. FatiGO: a web tool for finding significant associations of gene ontology terms with groups of genes. Bioinformatics. 2004;20(4):578–80.
46. Zheng Q, Wang X-J. GOEAST: a web-based software toolkit for Gene Ontology enrichment analysis. Nucleic Acids Res. 2008;36(Suppl_2):W358–63.
47. Martínez-Cruz LA, Rubio A, Martínez-Chantar ML, Labarga A, Barrio I, Podhorski A, et al. GARBAN: genomic analysis and rapid biological annotation of cDNA microarray and proteomic data. Bioinformatics. 2003;19(16):2158–60.
48. Boyle EI, Weng S, Gollub J, Jin H, Botstein D, Cherry JM, et al. GO::TermFinder—open source software for accessing gene ontology information and finding significantly enriched gene ontology terms associated with a list of genes. Bioinformatics. 2004;20(18):3710–5.
49. Sun H, Fang H, Chen T, Perkins R, Tong W. GOFFA: gene ontology for functional analysis—a FDA gene ontology tool for analysis of genomic and proteomic data. BMC Bioinform. 2006;7(Suppl 2):S23.
50. Ye J, Fang L, Zheng H, Zhang Y, Chen J, Zhang Z, et al. WEGO: a web tool for plotting GO annotations. Nucleic Acids Res. 2006;34(Suppl_2):W293–7.
51. Zhang B, Schmoyer D, Kirov S, Snoddy J. GOTree Machine (GOTM): a web-based platform for interpreting sets of interesting genes using gene ontology hierarchies. BMC Bioinform. 2004;5(1):16.
52. Das S, Rai A, Mishra DC, Rai SN. Statistical approach for gene set analysis with trait specific quantitative trait loci. Sci Rep. 2018;8(1):2391.
53. Luo W, Brouwer C. Pathview: an R/bioconductor package for pathway-based data integration and visualization. Bioinformatics. 2013;29(14):1830–1.
54. Yi M, Horton JD, Cohen JC, Hobbs HH, Stephens RM. WholePathwayScope: a comprehensive pathway-based analysis tool for high-throughput data. BMC Bioinform. 2006;7(1):30.
55. Ge SX, Jung D, Yao R. ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics. 2020;36(8):2628–9.
56. Barry WT, Nobel AB, Wright FA. Significance analysis of functional categories in gene expression studies: a structured permutation approach. Bioinformatics. 2005;21(9):1943–9.
57. Efron B, Tibshirani R. On testing the significance of sets of genes. Ann Appl Stat. 2007;1 (1):107–29.
58. Smyth GK. Limma: linear models for microarray data. In: Bioinformatics and computational biology solutions using R and bioconductor. New York: Springer; 2005. p. 397–420. https://link.springer.com/chapter/10.1007/0-387-29362-0_23.
59. Lai W, Tian L, Parkway P. SigPathway: pathway analysis with microarray data; 2013.

60. Goeman JJ, van de Geer SA, de Kort F, van Houwelingen HC. A global test for groups of genes: testing association with a clinical outcome. Bioinformatics. 2004;20(1):93–9.
61. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, et al. Improving gene set analysis of microarray data by SAM-GS. BMC Bioinform. 2007;8(1):242.
62. Breslin T, Edén P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinform. 2004;5:193.
63. Boorsma A, Foat BC, Vis D, Klis F, Bussemaker HJ. T-profiler: scoring the activity of predefined groups of genes using gene expression data. Nucleic Acids Res. 2005;33:W592–5.
64. Henegar C, Cancello R, Rome S, Vidal H, Clément K, Zucker J-D. Clustering biological annotations and gene expression data to identify putatively co-regulated biological processes. J Bioinforma Comput Biol. 2006;4(4):833–52.
65. Backes C, Keller A, Kuentzer J, Kneissl B, Comtesse N, Elnakady YA, et al. GeneTrail—advanced gene set enrichment analysis. Nucleic Acids Res. 2007;35(Suppl_2):W186–92.
66. Kim S-B, Yang S, Kim S-K, Kim SC, Woo HG, Volsky DJ, et al. GAzer: gene set analyzer. Bioinformatics. 2007;23(13):1697–9.
67. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40(17):e133.
68. Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinform. 2009;10(1):161.
69. Frost HR, Li Z, Moore JH. Spectral gene set enrichment (SGSE). BMC Bioinform. 2015;16 (1):70.
70. Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinform. 2005;6(1):144.
71. Rahmatallah Y, Emmert-Streib F, Glazko G. Gene sets net correlations analysis (GSNCA): a multivariate differential coexpression test for gene sets. Bioinformatics. 2014;30(3):360–8.
72. Hsueh H-M, Tsai C-A. Gene set analysis using sufficient dimension reduction. BMC Bioinform. 2016;17(1):74.
73. Reich M, Liefeld T, Gould J, Lerner J, Tamayo P, Mesirov JP. GenePattern 2.0. Nat Genet. 2006;38(5):500–1.
74. Yi X, Du Z, Su Z. PlantGSEA: a gene set enrichment analysis toolkit for plant community. Nucleic Acids Res. 2013;41(W1):W98–103.
75. Rahmatallah Y, Zybailov B, Emmert-Streib F, Glazko G. GSAR: bioconductor package for gene set analysis in R. BMC Bioinform. 2017;18(1):61.
76. Wu X, Hasan MA, Chen JY. Pathway and network analysis in proteomics. J Theor Biol. 2014;362:44–52.
77. Rahnenführer J, Domingues FS, Maydt J, Lengauer T. Calculating the statistical significance of changes in pathway activity from gene expression data. Statistical applications in genetics and molecular biology. 2004 [cited 2020 Dec 21];3(1). https://www.degruyter.com/view/journals/sagmb/3/1/article-sagmb.2004.3.1.1055.xml.xml.
78. Tarca AL, Draghici S, Khatri P, Hassan SS, Mittal P, Kim J-S, et al. A novel signaling pathway impact analysis. Bioinformatics. 2009;25(1):75–82.
79. Alexeyenko A, Lee W, Pernemalm M, Guegan J, Dessen P, Lazar V, et al. Network enrichment analysis: extension of gene-set enrichment analysis to gene networks. BMC Bioinform. 2012;13(1):226.
80. Glaab E, Baudot A, Krasnogor N, Valencia A. TopoGSA: network topological gene set analysis. Bioinformatics. 2010;26(9):1271–2.
81. Martini P, Sales G, Massa MS, Chiogna M, Romualdi C. Along signal paths: an empirical gene set approach exploiting pathway topology. Nucleic Acids Res. 2013;41(1):e19.
82. García-Campos MA, Espinal-Enríquez J, Hernández-Lemus E. Pathway analysis: state of the art. Front Physiol [Internet]. 2015 [cited 2020 Dec 20];6. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4681784/.

83. Pavlidis P, Qin J, Arango V, Mann JJ, Sibille E. Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. Neurochem Res. 2004;29(6):1213–22.
84. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
85. Barabási A-L, Oltvai ZN. Network biology: understanding the cell's functional organization. Nat Rev Genet. 2004;5(2):101–13.
86. Folger O, Jerby L, Frezza C, Gottlieb E, Ruppin E, Shlomi T. Predicting selective drug targets in cancer through metabolic networks. Mol Syst Biol. 2011;7:501.
87. Luo B, Cheung HW, Subramanian A, Sharifnia T, Okamoto M, Yang X, et al. Highly parallel identification of essential genes in cancer cells. Proc Natl Acad Sci U S A. 2008;105 (51):20380–5.
88. Draghici S, Khatri P, Tarca AL, Amin K, Done A, Voichita C, et al. A systems biology approach for pathway level analysis. Genome Res. 2007;17(10):1537–45.
89. Amaral LAN, Ottino JM. Complex networks. Eur Phys J B. 2004;38(2):147–62.
90. Emmert-Streib F, Dehmer M. Networks for systems biology: conceptual connection of data and function. IET Syst Biol. 2011;5(3):185–207.
91. Khatri P, Sellamuthu S, Malhotra P, Amin K, Done A, Draghici S. Recent additions and improvements to the Onto-Tools. Nucleic Acids Res. 2005;33:W762–5.
92. Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, et al. Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics. 2010;26(12):i237–45.
93. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. From molecular to modular cell biology. Nature. 1999;402(6761):C47–52.
94. Yaffe MB. Signaling networks and mathematics. Sci Signal. 2008;1(43):eg7.
95. Demir E, Cary MP, Paley S, Fukuda K, Lemer C, Vastrik I, et al. The BioPAX community standard for pathway data sharing. Nat Biotechnol. 2010;28(9):935–42.
96. Bar-Joseph Z, Gerber G, Simon I, Gifford DK, Jaakkola TS. Comparing the continuous representation of time-series expression profiles to identify differentially expressed genes. Proc Natl Acad Sci U S A. 2003;100(18):10146–51.
97. Maleki F, Ovens K, Hogan DJ, Kusalik AJ. Gene Set analysis: challenges, opportunities, and future research. Front Genet [Internet]. 2020 [cited 2020 Dec 20];11. Available from: https://www.frontiersin.org/articles/10.3389/fgene.2020.00654/full#h7.
98. Maleki F, Ovens K, McQuillan I, Kusalik AJ. Size matters: how sample size affects the reproducibility and specificity of gene set analysis. Hum Genomics [Internet]. 2019 [cited 2020 Dec 21];13(Suppl 1). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6805317/.
99. Tan SH, Swathi Y, Tan S, Goh J, Seishima R, Murakami K, et al. AQP5 enriches for stem cells and cancer origins in the distal stomach. Nature. 2020;578(7795):437–43.
100. Mathur R, Rotroff D, Ma J, Shojaie A, Motsinger-Reif A. Gene set analysis methods: a systematic comparison. BioData Mining. 2018;11(1):8.
101. Geistlinger L, Csaba G, Santarelli M, Ramos M, Schiffer L, Turaga N, et al. Toward a gold standard for benchmarking gene set enrichment analysis. Brief Bioinform [Internet]. 2020 [cited 2020 Dec 21];(bbz158). Available from: https://doi.org/10.1093/bib/bbz158.
102. Zyla J, Marczyk M, Weiner J, Polanska J. Ranking metrics in gene set enrichment analysis: do they matter? BMC Bioinform. 2017;18(1):256.
103. Ackermann M, Strimmer K. A general modular framework for gene set enrichment analysis. BMC Bioinform. 2009;10(1):47.
104. Nam D, Kim S-Y. Gene-set approach for expression pattern analysis. Brief Bioinform. 2008;9 (3):189–97.
105. Tamayo P, Steinhardt G, Liberzon A, Mesirov JP. The limitations of simple gene set enrichment analysis assuming gene independence. Stat Methods Med Res. 2016;25 (1):472–87.

106. Maleki F, Kusalik A. A synthetic kinome microarray data generator. Microarrays. 2015;4 (4):432–53.
107. Maciejewski H. Gene set analysis methods: statistical models and methodological differences. Brief Bioinform. 2014;15(4):504–18.
108. Tripathi S, Glazko GV, Emmert-Streib F. Ensuring the statistical soundness of competitive gene set approaches: gene filtering and genome-scale coverage are essential. Nucleic Acids Res. 2013;41(7):e82.
109. Wiebe DS, Omelyanchuk NA, Mukhin AM, Grosse I, Lashin SA, Zemlyanskaya EV, et al. Fold-change-specific enrichment analysis (FSEA): quantification of transcriptional response magnitude for functional gene groups. Genes. 2020;11(4):434.
110. Maleki F, Kusalik A. Gene set overlap: an impediment to achieving high specificity in over-representation analysis. In 2020 [cited 2020 Dec 21]. p. 182–93. Available from: https://www.scitepress.org/Link.aspx?doi=10.5220/0007376901820193.
111. Simillion C, Liechti R, Lischer HEL, Ioannidis V, Bruggmann R. Avoiding the pitfalls of gene set enrichment analysis with SetRank. BMC Bioinform. 2017;18(1):151.
112. Irizarry RA, Wang C, Zhou Y, Speed TP. Gene set enrichment analysis made simple. Stat Methods Med Res. 2009;18(6):565–75.

Part III

High Throughput Technologies

High-Throughput Sequencing Technologies 13

Elakkiya Elumalai and Krishna Kant Gupta

Abstract

Optical and biochemical methods determine the sequence of nucleotide bases in a DNA macromolecule. Since 2010, considerable progress has been made on DNA sequencing technology. There are various high-throughput sequencing (HTS) technologies, such as Roche 454, Illumina dye sequencing, PacBio's SMRT, Ion Torrent, Oxford Nanopore, SOLiD, and DNA nano-array sequencer, that have emerged with less cost and are time-saving. Sanger sequencing is the first-generation sequencing method. Subsequently, many next-generation sequencing (NGS) platforms are being used for genome sequencing. These DNA sequencing technologies have altered our view on understanding genomes and their analysis. This chapter presents a simple overview of the HTS technologies, their applications, and limitations. We aim to provide readers in the field with an easy and comprehensible description of HTS technologies to provide them with essential knowledge in full zeal.

Keywords

Long reads · Short reads · Next-generation sequencing technologies · DNA · High-throughput sequencing

E. Elumalai
Centre for Bioinformatics, Pondicherry University, Pondicherry, India

K. K. Gupta (✉)
School of Chemical and Biotechnology, SASTRA University, Thanjavur, Tamil Nadu, India
e-mail: krishnakant@scbt.sastra.ac.in

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_13

Abbreviations

ARE	Parallel analysis of RNA ends
BAM	Binary alignment map
CAGE	Cap analysis of gene expression
ChIA-PET	Chromatin interaction analysis by paired-end tag sequencing
DOE	Department of Energy
HTS	High-throughput sequencing
ISPs	Ion Sphere Particles
NGS	Next-Generation Sequencing
NHGRI	National Human Genome Research Institute
NIH	National Institute of Health
PCR	Polymerase chain reaction
RIP-chip	RNA immunoprecipitation chip
RNA-Map	RNA on a massively parallel array
SBS	Sequencing by synthesis
SMRT	Single molecule real-time
SNVs	Single nucleotide variants
SOLiD	Sequencing by sequential ligation of oligonucleotide probes
Svs	Structural variations
TADs	Topological Associated Domains
XIAP	X-lined inhibitor of apoptosis

13.1 Introduction

The discovery of the double helix structure was the landmark discovery that has led to the decoding of genomic sequences [1]. Sanger et al. developed the first sequencing technology and were awarded a noble prize in chemistry in 1980 [2]. Thereafter, a golden period came as it opened the door to develop efficient and faster sequencing technology. The emerging sequencing technologies have a significant role in genome analysis. These advanced sequencing technologies are known as "high-throughput sequencing technologies (HTS)" or "Next-Generation Sequencing (NGS) Technologies." It even led us to study the genetic code of living beings. HTS has steadily advanced over the last 15 years and advanced technologies are continuously being commercialized (Fig. 13.1). When the technology advances, coupled with expanded uses of fundamental and practical research, the effects of the technology advancement becoming more and more diverse and usable implementations [4]. In 2005, Roche's 454 technology was commercialized at a much lower cost and a very high throughput than the first sequencing technologies [5]. NGS technologies simultaneously perform parallel analysis from multiple samples at a much-reduced time and cost. Thus, the key aim of this chapter is to include a compendium of methods used for the study and evaluation of NGS reads.

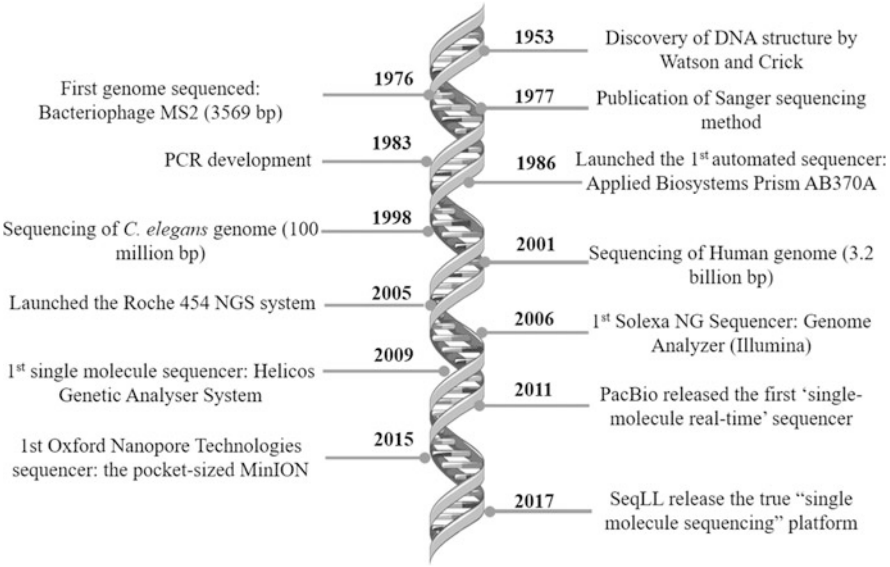

1953 Discovery of DNA structure by Watson and Crick
1976 First genome sequenced: Bacteriophage MS2 (3569 bp)
1977 Publication of Sanger sequencing method
1983 PCR development
1986 Launched the 1st automated sequencer: Applied Biosystems Prism AB370A
1998 Sequencing of C. elegans genome (100 million bp)
2001 Sequencing of Human genome (3.2 billion bp)
2005 Launched the Roche 454 NGS system
2006 1st Solexa NG Sequencer: Genome Analyzer (Illumina)
2009 1st single molecule sequencer: Helicos Genetic Analyser System
2011 PacBio released the first 'single-molecule real-time' sequencer
2015 1st Oxford Nanopore Technologies sequencer: the pocket-sized MinION
2017 SeqLL release the true "single molecule sequencing" platform

Fig. 13.1 DNA sequencing timeline. Some of the most revolutionary and remarkable events in DNA sequencing. *SeqLL* sequence the lower limit, *PCR* polymerase chain reaction, *NG* next generation, *SMS* single molecule sequencing. (Adapted from [3]) (CC BY 4.0)

Towards the end, we have also discussed the applications, the hindrance, and challenges of current HTS platforms.

13.2 Sequencing Technologies

Sequencing platforms can be broadly classified into three main groups: first generation, second generation, and third generation. Few important sequencing platforms have been described below (Table 13.1).

13.2.1 First Generation

The main techniques of DNA sequencing in the first generation to be taken into account are the Sanger dideoxy synthesis [2, 7] and Maxam-Gilbert chemical cleavage methods. The Maxam-Gilbert process includes chemical alteration of DNA and then cleaves the DNA backbone adjacent to the altered nucleotides. Sanger sequencing employs 3′-hydroxyl-free chain-terminating nucleotides (dideoxynucleotides). Therefore, DNA polymerase cannot form a phosphodiester bond at that site, resulting in the termination of the growing DNA chain. There are fluorescently or radioactively coded ddNTPs for identification in sequencing gels and devices, respectively. Though the methodology of the initial Maxam-Gilbert

Table 13.1 Summary of few important sequencing platforms, adapted from [6]

Platform Name	Illumina HiSeq 2500	Ion Torrent-Proton II	PacBio RS II	OxFord Nanopore Minion
Cost (USD)[a]	690 k	224 k	695 k	1 k[b]
Reagent cost per run/per GB	4126/45.84	1000/20.41	100/1111.11	900/1000
Reads per run	300 millions	280 millions	0.03 millions	0.1 millions
Average read length	2 × 150 bp	175 bp	14,000 bp	9000 bp
Run time	10 h	5 h	2 h	6 h
Major errors	Substitution	Indel	Indel	Deletion
Error rate (%)	0.1	1	1	4
Amplification	BridgePCR	emPCR	None, SMS	None, SMS
Advantage	Low cost per GB; high output	Low cost	Long reads, no amplification bias	Long reads, no amplification bias
Disadvantage	High cost	Homopolymer errors	Low throughput; high cost	High error rate

[a]Sources: http://www.molecularecologist.com/next-gen-table-3a-2014/
[b]Accessing fee. Sources: https://www.nanoporetech.com/products-services/minion-mki

procedure has also been changed to exclude potentially hazardous substances, SBS is still the sequencing norm. The Sanger sequencing system was first conceived in 1977. Though Sanger sequencing is relatively slow compared to other NGS methods, advancements in technology and commercialization also led to its current recognition as the best sequencing process. Several factors, including the production of automatic capillaries and "ultrafiltration" sample preparation methods, also led to improved accuracy and efficiency of the Sanger operation. Among the most important advances in Sanger sequencing are the following: (1) the creation of fluorescent (terminator) dyes, (2) the usage of thermal-cycle sequencing to reduce the amount of input DNA required, and thermostable polymerases that efficiently and accurately incorporate the terminator dyes into the growing DNA strands, (3) development of software to read and evaluate the sequence. Applied Biosystems (AB) is a pioneer in automated DNA sequencing. To improve identification, all commercially available AB sequencers employ fluorescent markers and capillary electrophoresis (CE) [4].

The ABI sequencing system, a platform based on sequential ligation of oligonucleotide probes, was introduced in autumn 2007. DNA is fragmented and ligated to adapter sequences [8]. It is then bound to beads. The amplification reagents are present in an emulsion (water droplet-oil), and the bead has only one fragment bounded. The DNA fragments are amplified on the beads using the emulsion PCR [9]. The further steps are briefly described as follows: (1) The beads are deposited onto a glass support surface after DNA denaturation. It follows primer annealing to the adapter. (2) The DNA fragments are allowed to hybridize with a mixture of oligonucleotide octamers in the presence of a ligation mixture. The doublet of the fourth and fifth bases of octamer is labeled with fluorescent dyes. Therefore bases fourth and fifth in the sequence are identified after detecting the fluorescence. (3) The octamer oligonucleotides are cut off after the fifth base, removing the fluorescent label. The hybridization and ligation cycles continue. This time 9th and 10th bases in the sequence are identified; in the next cycle bases 14th and 15th are determined, and so on. (4) In order to get all the remaining bases, the primer length is shortened by one base than the previous one. It allows us to determine, in the successive cycles, bases 3th and 4th, 8th and 9th, 13th and 14th. The achieved sequence reading length is at present about 35 bases. Because each base is determined with a different fluorescent label, error rate is reduced.

Applied Biosystems SOLiD 2.0 platform can sequence millions of bead clusters, increasing the instrument's output from 3 to 10 Gb per run. The Sanger sequencing technology tends to be beneficial in situations where we need not have high throughput. Many commercial DNA sequencing facilities and businesses offer Sanger sequencing services. The most popular uses are in DNA sequencing utilizing a particular oligonucleotide primer on a specific template, for example, to validate DNA constructs or PCR products [4].

13.2.1.1 Advantages and Limitation

First-generation methods are based on the chain termination method. Therefore, fragments of varied length are used for sequencing. This method can sequence read length from 800 to 1000 bp, but it is quite slow as only one fragment can be sequenced [10].

13.2.2 Second Generation

Second-generation sequencing approaches may be sequencing by synthesis (SBS) or sequencing by hybridization (SBH). The SBS method improves Sanger sequencing by eliminating dideoxy terminators and incorporating several synthesis cycles, imaging, and strategies for integrating new nucleotides into the developing chain. These new methods may seem costly at first sight, but since the reactions are stored in tiny chambers, the cost of DNA sequencing is nominal. Continued refinement at reducing the prices is still more demanding [4].

13.2.2.1 Sequencing by Hybridization

The initial technique used in the 1980s was based on utilizing oligonucleotides with known sequences to be used as DNA primers. It was possible to determine whether equal amounts of each labeled fragment are hybridized to each filter by hybridizing and removing nonhybridized DNA. As a result, it became important to assemble a contiguous series from the hybridization spots for the probes. Hybridization-based sequencing is primarily a technique for which particular probes are required, such as detecting disease-associated SNPs in particular genes or chromosome abnormalities (deletions, duplications, rearrangements, copy number variants) [4, 11–13].

13.2.2.2 Sequencing by Synthesis (SBS)

A variety of SBS techniques have taken different approaches. The second-generation approaches use solid supports of microchannels to facilitate the sequencing reactions. “Reversible” terminators may be used in several of the newer SBS technologies, allowing normal nucleotide integration reactions to occur when imagining the integrated nucleotides and then eliminating synthesis-blocking moieties from the marked DNA sequence to enable incorporation of another base in the sequences [4]. Recent SBS methods are distinguished from the initial sequencing method in that they depend on far shorter sequence reads. Compared to Sanger sequencing, they naturally have a higher error rate since they depend on millions or even billions of short reads of DNA sequence (50–300 nt) that achieve a consistent sequence. Certain architectures often suffer from sequence background failures, which are often unresolvable by which the number of reads increases. Homopolymer sequences, for example, include fragments of the same nucleic acid, viz., AAAAAAAAAA. Throughout this scenario, separate sequencing platforms have a slight advantage in terms of determining the sequential bases. Each platform has its specific collection of possible site errors, and users need to be informed of such site errors. We use various systems and different technology (discussed below) for circumventing these issues [4].

Roche 454

The technology for HTS was introduced by company 454 Life Sciences in 2005. It is the first proposed method for the NGS technology [14]. It includes the following steps:

Preparation of DNA Library

The DNA samples are broken down into small fragments of 300–800 bp by using the spray method. The different adapters are added at both ends of fragments. The DNA is denatured, followed by primer annealing. It is cloned using specific vectors and then designing a library of single-stranded DNA.

Emulsion PCR

Beads are taken and coated with streptavidin. These beads have primers that match the adapters used. Each bead is eluted with PCR reagents in a water-oil droplet (buffers, dNTPs, primer, DNA Polymerase) [15]. This is followed by numerous tiny water droplets wrapped in mineral oil formed during high-speed spinning. These individual oil droplets form an independent PCR reaction space having one DNA template and one bead. The double-stranded DNA with adapters are denatured at 95 °C, and single-stranded DNAs are attached to the beads. Reverse strand anneals to the forward primer and vice-versa. The incubation system contains PCR reagents so that PCR cycles continue.

Pyrosequencing

The DNA sequencing further needs a DNA polymerase and single-strand DNA binding protein [16]. Picotiter plate with special nanopores (44 μm diameters) is used to place these beads. The specialty of this plate is that only one bead can be placed at each nanopore. This setup is required for pyrosequencing. The principle of pyrosequencing is as follows: Place a bead into the nanopore. The amplified and fixed single-stranded DNAs on the bead provide a sequencing template. The pyrophosphate group will be released if a single dNTP pairs with the DNA template. Subsequently, as the released pyrophosphate group interacts with ATP sulfuric acid chemical enzymes, ATP is formed. The luciferase enzyme and ATP make the fluorescein molecule fluoresce, and a CCD camera captures it. Finally, computer software is used to process the data. Each dNTP exhibits a different fluorescence, allowing the DNA sequence to be observed. Finally, diphosphatase degrades ATP, and it leads to fluorescence quenching, and sequencing reaction continues. Roche 454 technology can sequence read lengths up to 400 bp [17].

Advantages and Limitations

Roche is a reliable, fast, and accurate technology which generates on an average 700 MB of data. In this technology, the read length is >700 bp. It does not need primers and labeled nucleotides. The limitations of this technology includes problems in homopolymer sequencing, and it is costlier than other NGS technologies [10].

Illumina

The first high-throughput Genome Analyzer (Solexa sequencer) was launched in 2006, and Illumina acquired it in early 2007. Many plants, animals, microbes, and human genomes have been sequenced with this technology. This technology holds the cluster generation. The occupancy reversible terminator technology is used for fast and accurate large-scale sequencing. It has broad applications in genomics, transcriptomics, and epigenomics. This technology allows researchers to sequence 1 gigabase (Gb) of genome data in a single run [4]. The steps are as follows:

Cluster Generation

Firstly, DNA templates are fixed on a proprietary flow cell surface to facilitate DNA replication enzyme's access. Solid-phase amplification creates many copies (~1000) of each template molecule nearby [18].

Sequencing by Synthesis

The four dNTPs (A, C, T, G) are labeled with different fluorescent dyes. These nucleotides are added to sequence the billions of clusters on the flow cell surface. It uses SBS technology. The nucleic acid template is sequenced by adding a single labeled deoxynucleoside triphosphate (dNTP), and it stops DNA polymerization. The fluorescent base is imaged and identified. Subsequently, it is enzymatically cleaved to incorporate the upcoming nucleotide. This whole process continues till it sequences the template [19].

Analysis Pipeline

This approach is built around a massive quantity of sequence reads in parallel. Each raw read base is associated with a score, and software applies a weighting factor in calling differences to get confidence scores. Therefore, deep sampling and uniform coverage allow weighted majority voting and statistical analysis to identify heterozygotes and homozygotes and identify sequencing errors. Users can align sequences to a reference genome in resequencing applications of Illumina data collection software [20].

Ion Torrent

In 2010, ThermoFisher company expanded its NGS portfolio with the Ion Personal Genome Machine (PGM) sequencer. This technology measures the H^+ ions liberate during base incorporation [21]. It follows the sequencing-by-synthesis method and emulsion PCR (emPCR). The chemically encoded information (A, C, G, T) is decoded into digital information (0, 1) on a semiconductor chip. This approach combines simple chemistry with proprietary semiconductor technology. A hydrogen ion is liberated as a byproduct when a nucleotide is added into a strand of DNA by a DNA polymerase in real time. The workflow consists of four significant steps: library construction, template preparation, sequencing, and analysis.

Library Construction

The process generally involves taking nucleic acid, its fragmentation (typically 200–400 bp), and then adding sequencing adapters.

Template Prep/Amplification

Similar to Roche 454, the DNA fragments generated are affixed to beads followed by its amplification using emulsion PCR. The complementary primers coat beads. It is mixed with a dilute aqueous solution containing PCR reagents and DNA fragments. An emulsion of microdroplets is formed when the solution is then mixed with oil. It is ensured that each microdroplet contains only one bead with fragments with a low concentration of beads and fragments. The clonal amplification of each fragment is done within the microdroplets. The amplified fragments are harvested by centrifugation and organic extraction. The glycerol gradient embeds those amplified beads.

Sequencing

The principle is based on standard pyrosequencing chemistry. The direct release of H + (protons) from the reaction is measured by the Ion Torrent system. The leads to the decrease in pH of liquid, and the sensing layer detects this pH change. It includes pH meters and other inexpensive sensors essentially. This liquid surrounds Ion Sphere Particles (ISPs). Further, it is converted to a voltage change. The software records the nucleotide. For example, we have two nucleotides in a row incorporated (two T's complementary to two A's)—double the hydrogen is released, which results in double the signal so that the software will record two T's in a row. It generally completes 200 bp reads in 2 h. The standard FASTQ file is generated and analyzed by using "Torrent Browser" software (https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/).

Advantages and Limitation

In this technology, a low concentration of DNA sample (~10 ng) is required for expression and mutation analysis. It produces shorter reads. It is widely used, as it is fast and affordable [10].

Complete Genomics DNA Nano-array Sequencer

DNA nano-array sequencing is often used to sequence an organism's whole genome [22]. This method creates DNBs (DNA nanoballs) by amplifying small fragments via rolling circle replication. Fluorescent nucleotides attach to complementary nucleotides, and the bound nucleotide's fluorescence is used to determine the base sequence. The steps are given below:

DNA Isolation, Fragmentation, and Size Capture

DNA is extracted using a standard protocol, and physical or enzymatic methods randomly fragment it. The ideal fragment length is selected by gel electrophoresis. The bead-based size selection is considered to be suitable for longer fragments.

Attaching Adapter Sequences

The fragments are ligated with adapter DNA sequences, which are parts of identified DNA flanking the unknown DNA. After PCR amplification, a splint oligo hybridizes to the ends of the ligated fragments to form a circle. The exonuclease digests the linear single- and double-stranded DNA components, resulting in a circular DNA template.

Rolling Circle Replication

The Phi 29 DNA polymerase recognizes and replicates a circular single-stranded DNA template. The newly synthesized long single-stranded DNA strand is cut off from the circular template. Self-assembly of the corresponding DNA results in a tight ball of nanoparticles (300 nm) in diameter. DNBs are injected with the fluidics system. Patterned Array chip is used for its loading [23]. The adapter region of DNB is hybridized with the sequencing primer. The sequencing reagents contain DNA polymerase and fluorescently labeled dNTP probes. These reagents are pumped into the system to begin sequencing reaction. The fluorescently labeled probes on the DNB are excited with lasers, and images are taken (CCD camera). The MGI's propriety software converts images into a digital signal. This information is then used to sequence the DNA of the sample.

Sequencing Data Format

The information collected by the DNA nanoballs is configured as standard FASTQ files which can be read by any data analysis pipeline that supports single- or paired-end FASTQ files.

13.2.3 Third-Generation Sequencing

Compared to second-generation sequencing approaches, third-generation sequencing methods seek to interpret longer DNA molecules sequentially. The pioneer in the current marketplace for this is Pacific Biosciences (PacBio) [24] and Oxford Nanopore Technology [25].

13.2.3.1 PacBio's SMRT (Single Molecule Real Time) Sequencing

In the year 2016, using long-read SMRT sequencing from Pacific Biosciences, scientists from the University of Washington, the McDonnell Genome Institute at Washington University in St. Louis, and other institutions reported the gorilla's strongest genome assembly. Pacific Biosciences of California, Inc., is the leader in long-read sequencing using its Single Molecule, Real-Time (SMRT$^{®}$). This is the third-generation sequencing technology, and it is widely used to sequence best-quality genome sequences [24].

The Principle of PacBio SMRT Sequencing

Zero-mode waveguides (ZMWs) are a sub-wavelength optical nanostructure contrived in a thin metal film, which is a useful analytical tool able to entrap an excitation volume in the attoliter range. It enables the isolation of fluorescently tagged biomolecules at physiologically appropriate concentrations for optical analysis. A DNA polymerase and a template DNA strand are covalently bound to the lower glass surface of the ZMW. The laser light passes across the lower portion of a ZMW without penetrating it. The dimensions of the ZMW are less than the wavelength of light. As a result, it enables the targeted excitation and detection of light released by base elongation nucleotides [10].

SMRT Steps

Library Construction

The library construction workflow includes the following steps: quality checks of genomic DNA (gDNA); gDNA fragmentation (Covaris); size selection and concentration adjustment; repairing of DNA damage and ends of DNA Fragments; DNA purification; and blunt-end ligation through blunt adapters. Subsequently, the ligating hairpin adapters are used to make circular single-stranded DNA. The adapters are attached to the terminal of target double-stranded DNA (dsDNA) molecules. This template is called a SMRTbell.

Sequencing

The template DNA is dispensed into the ZMW cell [26], and even the adapter is attached to the bottom-immobilized DNA polymerase. The four nucleotides have been labeled with a separate fluorescent dye (red, yellow, green, and blue for G, C, T, and A, respectively) to create distinct emission spectra. A fluorescently labeled sequence binds to the template in the polymerase's active site, and the fluorescence output of the color identifies the inserted nucleotide. To end the fluorescence pulse, the dye linker-pyrophosphate compound is separated from the DNA sequence. Subsequently, the polymerase is transferred to the next position, and the cycle continues. PacBio SMRT sequencing generates relatively long reads. The normal read length is between 8 and 15 KB and can reach 40–70 KB. It is possible to conduct bioinformatics analyses such as de novo assembly, reference genome annotation, genome mapping, gene feature annotation, SNP/InDel detection, comparative genomics study, evolutionary analysis, and divergence time estimation.

Advantages and Limitation

The enzyme kinetics and sequencing can be easily monitored in real time. In this technology, RNA molecules can be observed at a single molecule resolution. Here, de novo sequencing is possible due to longer reads and unbiased data. Its limitation includes high cost per base sequencing, high error rate, and less data is generated in each run [10].

13.2.4 Oxford Nanopore Technology

Nanopore sequencing comes under fourth-generation DNA sequencing technology. Nanopores are small holes through which DNA can travel, generating an electronic signal used to sequence the individual bases. Oxford Nanopore Technologies Ltd. Developed Nanopore sequencing technology, which can generate very long reads with a relatively lower error rate and inexpensive to own and operate [25]. The Oxford Nanopore system primarily consists of a nanopore embedded in an artificial membrane and a motor protein that moves DNA molecules from one side to another side of the membrane through the nanopore. An electric current is induced across the nanopore when a specific voltage is applied across the membrane. When DNA passes through the nanopore, it modifies the current, with different bases having slightly different effects. These differences are used to reconstruct the DNA sequence and, in some cases, base modifications as well. A single molecule of Nucleic Acid can be sequenced without the need for PCR amplification or chemical labeling of the sample using nanopore sequencing.

13.2.4.1 Advantages and Limitation

This platform is able to produce read length >882 bp, which helps in data assembly and alignment. It is easily portable and rapid. The major limitations of this technology include a high error rate [10].

13.3 The Applications of HTS

The HTS finds many applications such as transcript analysis, translation analysis, chromatin conformation deduction, RNA structure, RNA protein interactions analysis, microRNA target discovery, and enhancer assays. The decrease in price and good accessibility have enabled researchers to develop diverse HTS methods

[27]. There are some major projects such as ENCODE [28], Roadmap Epigenomics Project which characterizes the human genome, 1000 Genomes project [29] which studies human genetic variation, GTEx [30] which analyzes gene expression, and many other which are entirely based on HTS. These resources provide enormous data to the scientific community, and these consortia generally implement computational standards and robust experimental methods, ensuring high-quality data. The applications of HTS are briefly given below (Table 13.2):

13.3.1 Genome Sequencing and Variation

The first genome to be sequenced by the utility of HTS technologies is *Acinetobacter baumannii* [43]. The resequencing of exomes and human genomes was done as the HTS technologies improved. The mapping of reads to a reference genome was done, and variants were identified between the sample genome and the reference. Several genome sequencing projects started, and finally it was concluded that there are 3.5–4 million single nucleotide variants (SNVs) and thousands of short indels (insertion and deletion) relative to the reference genome. The 1000 Genomes Project Consortium, 2010 shed light on hundreds of variants that led to alterations in genes and loss of function. HTS has been used to identify the genome segments that have been rearranged (Structural Variations, Svs), duplicated, or deleted, but it is more challenging to determine Svs and indel in the short-read lengths. Basically, to identify Svs in the genome, four independent approaches are used. These approaches are depth of reading coverage, mapping of paired-end reads discordant from the reference genome, identifying split reads, and mapping breakpoint junctions. Each method has its pros and cons. No combination of them is conclusive; therefore SVs are never characterized in their entirety. The improvement in resolution over the array-based method has greatly increased our understanding of the prevalence of SVs throughout the genome and their contribution to disease. HTS is excessively used to sequence viral, prokaryotic, eukaryotic genomes, and exomes, yielding tremendous insight into human diversity and disease.

13.3.2 Genome Regulatory Information Mapping

The other most important HTS application includes high-resolution genome-wide mapping of DNA regulatory elements. The associated technology is ChIP-Seq, in which a transcription factor (TF) is associated with DNA, and it is thereafter immuno-selected, followed by HTS [44]. This sequenced DNA are mapped to the genome that mark bound regions or chromatin modifications. Therefore, it is a basic method for discovering many probable regulatory regions. Generally, the accessible regions of the genome are digested with DNase I, followed by sequencing of the ends fragments [45] The ENCODE Project Consortium, 2012 (https://www.encodeproject.org/) has provided a treasure of significant information regarding transcription factor binding networks, Epigenetic Maps, and transcript annotation.

Table 13.2 Bioinformatics tools for high-throughput sequencing

	Tools	Purpose	Free or paid	Languages used	References
Genome sequencing and variation	Pisces	It identifies variants in germline next-generation sequencing data	Free	Python	[31]
	BATCAVE	It identifies variants in somatic next-generation sequencing data	Free	R	[32]
Genome regulatory information mapping	Genetic Association Database	It provides information on genetic association in diseases.	Free	SQL, HTML, CSS, Javascript	[33]
	ENCODE	It gives information on regulatory elements in the human genome	Free	SQL, HTML, CSS, Javascript	[28]
Mapping the three-dimensional organization of the genome	3D-GNOME 2.0	It is a web service that provides 3D models of genomic structures from the 1000 Genomes Project phase	Free	SQL, HTML, CSS, PHP, Javascript	[34]
	GSDB	It provides a 3D genome structure framed by using HiC	Free	SQL, HTML, CSS, PHP, Javascript, Python, R	[35]
Characterizing the transcriptome	PdumBase	It is a repository for the early development of other metazoans	Free	SQL, HTML, CSS, PHP, Javascript,	[36]
Microbiome sequencing	iMAP	It is a tool for microbiome data analysis	Free	Bash, R 3.5, Perl, Python	[37]
Genome sequencing in diseases	Human Gene Mutation Database	It is a collection of mutated human genes involved in diseases	Paid	SQL, HTML, CSS, PHP, Javascript	[38]
Human Genome Project	ampliconDIVider	It identifies insertion and	Free	Python	[39]

(continued)

Table 13.2 (continued)

	Tools	Purpose	Free or paid	Languages used	References
		deletion in DNA amplicons			
Plant Genome Databases	PlantGDB, Ensembl Plants	These two databases provide genomic information of plants	Free	SQL, HTML, CSS, PHP, Javascript	[40, 41]
g-HTS in the coming era of precision medicine (PM)	PreMedKB	It is a knowledge base for precision medicine	Paid	SQL, HTML, CSS, PHP, Javascript	[42]

The Roadmap Epigenomics Consortium (http://www.roadmapepigenomics.org/) has reported that more than 3.5 million regulatory elements are identified in different cell types throughout the human genome.

13.3.3 Mapping the Three-Dimensional Organization of the Genome

The advancement of our understanding of chromosome's compartmentalization and the global organization has been increased severalfold by HTS technologies. The ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) and Hi-C assays help study 3D chromatin interactions [46]. These assays rely upon broken chromatin followed by sequencing to derive contact maps and proximity-based ligation of cross-linked. Hi-C was the first technique demonstrating the organization of the genome into topological associating domains (TADs). The combination of extremely deep sequencing (billions of reads per sample) and the Hi-C technique have released much higher resolution contact maps (~1 kb), which refine TAD domain size from 1 Mb to less than 200 kbp. Modeling of Hi-C data suggested the fractal globule chromatin state, a conformation that maximizes packing while preserving the flexibility to access any genomic locus [47].

13.3.4 Characterizing the Transcriptome

The advent of high-throughput sequencing has greatly increased our understanding of the diverse cellular roles of RNA. The HTS identifies different classes of RNA and characterizes genomic localization, RNA structure, and RNA-protein interactions. The transcriptome analysis depends on various HTS methods such as RNA-seq and Cap analysis of gene expression (CAGE). HTS and deep sequencing of RNA suggest that around 3/4 of the human genome is transcribed [48]. In transcriptome analysis, noncoding RNA, including lncRNAs (long, noncoding), snoRNAs (small, nucleolar), and microRNAs, have been systematically described RNA-seq and derivative techniques. RNA-seq data with ChIP-seq profiles characteristic of expressed genes identified a subset of lncRNAs.

Similarly, cDNA sequencing and tiling array experiments identified the lncRNAs and RNA editing. The HTS methods revealed the structure and biology of these newly discovered transcripts. The miRNA-mediated mRNA decay, using parallel analysis of RNA ends (PARE), has led to the microRNA-target discovery [49]. RNA immunoprecipitation chip (RIP-chip) and RIP-seq methods have shown that polycomb repressor complex 2 (PRC2), a chromatin-modifying complex, is associated with approximately 20% of the lncRNAs. The massively parallel array of RNA (RNA-MaP) helps in estimating RNA–protein interactions and, thus, plays a vital role in understanding human disease and normal cellular homeostasis. These analyses improved our understanding of RNA, which plays a vital role in human disease and normal cellular homeostasis.

13.3.5 Microbiome Sequencing

The Human Microbiome Project Consortium, 2012 has been identifying the many microbes residing in healthy human populations. The extensive analysis of metagenomic samples from the ocean, soil, and the human body provides insight into microbial species diversity [50]. The detailed species, gene composition, and phylogenetic relationships are predicted by using 16S rRNA gene sequencing. The advancement in HTS led us to a "personal microbiome." The relation between several human diseases and microbial diversity in a given niche has been established. For example, a decrease in gut microbes diversity causes obesity and inflammatory bowel diseases, whereas the rise in microbial diversity is associated with bacterial vaginosis. The transplant studies in mice have concluded that there is a direct relation between gut microbiome and metabolism. The detailed characterization of the dynamics of microbiomes is possible due to the huge advancements in HTS [51].

13.3.6 Genome Sequencing in Diseases

Comprehensive insight into the genetics of human disease has increased severalfolds due to the capacity to sequence transcriptomes, exomes, and genomes. There are more than 7800 Mendelian disorders reported in the Online Mendelian Inheritance in Man database. Only a little information on causative genes is known. Identifying causal alleles for various inherited diseases is performed by sequencing the exome of healthy and diseased family members [52]. For example, exome sequencing uncovered a mutation in the X-lined inhibitor of apoptosis (XIAP) that causes severe inflammatory bowel disease. This problem is alleviated by a bone marrow transplant. Exome sequencing is being extensively used to identify genetic defects [53]. The utility of HTS in cancer has revealed that tumors can differ significantly in terms of mutation quality and type. The exome and genome sequencing on many clinical samples by International Cancer Genome Consortium (ICGC) and the Cancer Genome Atlas (TCGA) detected the mutation rates in cancer driver genes [54]. Where a gene is mutated more often than predicted, gene expression and replication timing are critical. Numerous novel cancer mechanisms and driver genes were identified as a result of TCGA-led projects. Additionally, whole-genome sequencing (WGS) of cancer samples defined noncoding high-frequency mutations as a significant class of somatic variants. The mechanisms underlying drug resistance, clonal evolution, and tumor heterogeneity are well studied due to the scale and sensitivity of HTS. Single-cell sequencing identifies changes in the copy number of breast cancer cells [55]. Point mutations confer clonal variation, allowing the tumor to adapt to changing selective pressures. HTS has been used to form the foundation for the biomarker discovery and treatment of cancer.

13.3.7 Human Genome Project

The progress in DNA sequencing technologies has created a path for the human genome project. This project has significantly contributed to our understanding of the disease and human diversity. The sequencing of individual genomes is now a matter of a couple of days [56]. The Human Genome Project was completed with Sanger DNA sequencing. The finished human genome was released in 2004 by the International Human Genome Sequencing Consortium [57]. This was followed by the genome sequencing of several model organisms. The Human Genome Project has reported that there are around 20,500 human genes. It has in-depth information about the structure, function, and organization of the complete set of human genes. The 13-year project was coordinated by the U.S. Department of Energy (DOE) and the National Institutes of Health [58]. The significant goals of this project are: to report total genes in the human genome; to sequence the human genome; to store and analyze genomic data; to shift genome sequencing technologies to the private sectors; and to discourse the ethical, legal, and social issues (ELSI) related to genome project.

The National Human Genome Research Institute (NGHRI) initiated a 70 million dollar DNA sequencing technology to achieve a $1000 human genome in 10 years [59]. The improvements to traditional Sanger sequencing results in a 100-fold decrease in per-base cost. The present worth of Human Genome sequencing is less than $2000. Many commercial high-throughput sequencing (HTS) platforms follow a common paradigm: sample preparation, sample amplification by cloning, followed by massively parallel sequencing. The technique employed by each platform dictates the amount, consistency, and biases of the corresponding sequencing data. This breakthrough involved many HTS platforms and found its usefulness for particular applications [60].

13.3.8 Plant Genome Databases

The increasing availability of computational capacities made it possible to analyze and store huge data, which is ever-growing. This huge amount of data has led to the development of plant genome databases that can store, anatomize this data, and retrieve relevant information. Both species-specific and general plant databases are accessible on the web, such as PlantGDB (http://www.plantgdb.org/), Wheatgenome.info (http://www.wheatgenome.info/), Phytozome (http://www.phytozome.net), TreeGenes (http://dendrome.ucdavis.edu/treegenes/), and Legume Information System (http://legumeinfo.org).

13.3.9 g-HTS in the Coming Era of Precision Medicine (PM)

HTS has improved our understanding of cancer a lot. PM is a medical treatment to the individual characteristics such as lifestyle, environment, and genes. PM in cancer study involves the identification of all kinds of mutations in genomes predicting resistance or response to therapies. We have already switched from “one medicine for all” to a specific treatment of patients according to their clinical features, disease stage, and biomarkers [61]. HTS takes into account a wide set of patient features and the cancer mutational conditions to choose the best therapeutic approach in disease management. It is now possible to screen all sets of genes in one test with the help of HTS. It allows to get in-depth genetic information from a blood sample and identify predictive and prognostic factors.

13.4 Limitations of HTS Technologies

Thirteen years have elapsed following the commercialization of the first high-throughput sequencing tool from Life Sciences, the 454 GS FLX. The genome has also expanded into structural and functional genomics as well as structural and functional genetics. Besides, it provides for the development of concepts such as “omics” (transcriptomics, genomics, metabolomics, etc.), which add additional aspects to our understanding of all living species and how different organisms utilize biology and molecular genetics to survive and reproduce in both normal and diseased environments [3]. This knowledge is important to appraise your understanding of human health. NGS has helped in several ways in the fields of health care, agriculture, and other areas of research. But it has also brought fresh dilemmas. The first obstacle is the price for sequencing. Though it is accurate that the total costs for NGS are much better than those for sequencing studies, an NGS experiment is not inexpensive. The expense of the sequencing system differs considerably based on the model of machine obtained, on consumables, and on the particular reagents used. Costs involved with laboratory design, sampling, and sequencing should be taken into consideration. Additionally, the expense of designing sequencing pipelines, developing bioinformatical tools that update those pipelines and analyze sequencing results, and data storage are not included in the overall NGS costs. The BAM (binary alignment map) file containing the results of a single whole-exome sequencing encounter takes up to 30 Gb and maintaining and analyzing data of many patients needs higher computing capacity and storage, which have considerable costs. Expert bioinformaticians will also be expected to work with how to utilize and interpret the data from the sequenced genomes. These extra expenses must be included in the expense of genome sequencing [62, 63].

Privacy issues around data sharing and secrecy are often essential to remember while utilizing next-generation sequencing. Is it discussed whether genomic data should be exchanged by numerous stakeholders, including laboratory personnel, bioinformatics analysts, academics, clinicians, patients, and their families? This is a really critical concern for health practitioners [64]. When analyzing NGS results, it is important to be aware of the technology's limitations, as demonstrated in this essay, namely PCR amplification bias (a significant source of bias induced by random errors introduced during the amplification process) and sequencing errors. As a result, extensive coverage is needed to determine which variants are correct and which are the result of sequencing or PCR errors. Finally, shortcomings often arise in downstream research on the basis of read alignment/mapping issues, which can be particularly troublesome for indels. Certain alignment methods have poor detection capability or do not detect at all. Even in the absence of automated methods, manual checks of variants in the BAM file are often beneficial. Thus, it is important to consider the shortcomings of the in-house NGS framework and workflow to enhance the efficiency of variant detection.

13.5 Conclusion and Future Perspective

In conclusion, determining the order of nucleic acid residues in biological samples is a goal of many biomedical or agricultural studies. Over the last 50 years, numerous researchers worldwide have devoted themselves to the challenge of developing DNA sequencing technologies. At the start of this field, researchers concentrated on using RNA sequences that were only about a dozen and a hundred nucleotides in duration. Over the years, advances in sequencing protocols, molecular biology, and automation expanded the computational capacities of sequencing while decreasing the expense, enabling the reading of DNA hundreds of base pairs in duration, massively parallelized to generate gigabases of data in one phase. Researchers gradually switched from the laboratory to the machine while pouring over a gel to running code. Genomes were decoded, papers published, businesses started and combined, and repositories of DNA sequence data were generated and they grew. With regard to DNA sequencing, the area has an extensive history that relates to the latest advances in the field. A knowledge of this background will offer an awareness of recent success and insight for potential progress, as lessons gained in the previous generation guide future progress. In the end, these next-generation technologies are speeding up the process of personalized medicine and drug discovery for public health welfare. Now, it is not surprising to say that only the sky is the limit!

Conflict of Interest None

Other Information Fig. 13.1 (CC BY 4.0) [3] and Table 13.1 (CC BY 4.0) [6] have been used under the terms of the Creative Commons Attribution License.

References

1. Padilla-Sanchez V, Gao S, Kim HR, Kihara D, Sun L, Rossmann MG, et al. Structure-function analysis of the DNA translocating portal of the bacteriophage T4 packaging machine. J Mol Biol. 2014;426(5):1019–38.
2. Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463–7. https://doi.org/10.1073/pnas.74.12.5463. PMID: 271968; PMCID: PMC431765.
3. Pereira R, Oliveira J, Sousa M. Bioinformatics and computational tools for next-generation sequencing analysis in clinical genetics. J Clin Med. 2020;9(1):132.
4. Slatko BE, Gardner AF, Ausubel FM. Overview of next-generation sequencing technologies. Curr Protoc Mol Biol. 2018;122(1):e59.
5. Harrington CT, Lin EI, Olson MT, Eshleman JR. Fundamentals of pyrosequencing. Arch Pathol Lab Med. 2013;137(9):1296–303.
6. Ye H, Meehan J, Tong W, Hong H. Alignment of short reads: a crucial step for application of next-generation sequencing data in precision medicine. Pharmaceutics. 2015;7(4):523–41.
7. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. 1975;94(3):441–8.
8. Chung CAB, Boyd VL, McKernan KJ, Fu Y, Monighetti C, Peckham HE, et al. Whole methylome analysis by ultra-deep sequencing using two-base encoding. PLoS One. 2010;5 (2):e9320.
9. Pu D, Chen J, Bai Y, Tu J, Xie H, Wang W, et al. Sequencing-by-ligation using oligonucleotide probes with 3′-thio-deoxyinosine. J Biomed Nanotechnol. 2014;10(5):751–9.
10. Gupta N, Verma VK. Next-generation sequencing and its application: empowering in public health beyond reality. In: Arora PK, editor. Microbial technology for the welfare of society, Microorganisms for sustainability. Singapore: Springer; 2019. p. 313–41. https://doi.org/10.1007/978-981-13-8844-6_15.
11. Mirzabekov AD. DNA sequencing by hybridization—a megasequencing method and a diagnostic tool? Trends Biotechnol. 1994;12(1):27–32.
12. Drmanac R, Drmanac S, Chui G, Diaz R, Hou A, Jin H, et al. Sequencing by hybridization (SBH): advantages, achievements, and opportunities. Adv Biochem Eng Biotechnol. 2002;77:75–101.
13. Qin Y, Schneider TM, Brenner MP. Sequencing by hybridization of long targets. PLoS One. 2012;7(5):e35819.
14. Luo C, Tsementzi D, Kyrpides N, Read T, Konstantinidis KT. Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS One. 2012;7(2):e30087.
15. Verma V, Gupta A, Chaudhary VK. Emulsion PCR made easy. BioTechniques. 2020;69 (1):421–6.
16. Nyrén P. The history of pyrosequencing(®). Methods Mol Biol. 2015;1315:3–15.
17. Buermans HPJ, den Dunnen JT. Next generation sequencing technology: advances and applications. Biochim Biophys Acta. 2014;1842(10):1932–41.
18. Heather JM, Chain B. The sequence of sequencers: the history of sequencing DNA. Genomics. 2016;107(1):1–8.
19. Ambardar S, Gupta R, Trakroo D, Lal R, Vakhlu J. High throughput sequencing: an overview of sequencing chemistry. Indian J Microbiol. 2016;56(4):394–404.

20. García-Chequer AJ, Méndez-Tenorio A, Olguín-López G, Sánchez-Vallejo C, Isa P, Arias CF, et al. Illumina next generation sequencing data and expression microarrays data from retinoblastoma and medulloblastoma tissues. Data Brief. 2016;6:908–16.
21. Salipante SJ, Kawashima T, Rosenthal C, Hoogestraat DR, Cummings LA, Sengupta DJ, et al. Performance comparison of Illumina and ion torrent next-generation sequencing platforms for 16S rRNA-based bacterial community profiling. Appl Environ Microbiol. 2014;80 (24):7583–91.
22. Xu Y, Lin Z, Tang C, Tang Y, Cai Y, Zhong H, et al. A new massively parallel nanoball sequencing platform for whole exome research. BMC Bioinform. 2019;20(1):153.
23. Porreca GJ. Genome sequencing on nanoballs. Nat Biotechnol. 2010;28(1):43–4.
24. Nakano K, Shiroma A, Shimoji M, Tamotsu H, Ashimine N, Ohki S, et al. Advantages of genome sequencing by long-read sequencer using SMRT technology in medical area. Hum Cell. 2017;30(3):149–61.
25. Petersen LM, Martin IW, Moschetti WE, Kershaw CM, Tsongalis GJ. Third-generation sequencing in the clinical laboratory: exploring the advantages and challenges of nanopore sequencing. J Clin Microbiol. 2019;58(1):e01315–9.
26. Jadhav V, Hoogerheide DP, Korlach J, Wanunu M. Porous zero-mode waveguides for picogram-level DNA capture. Nano Lett. 2019;19(2):921–9.
27. Mardis ER. Next-generation sequencing platforms. Annu Rev Anal Chem (Palo Alto, Calif). 2013;6:287–303.
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
29. Siva N. 1000 genomes project. Nat Biotechnol. 2008;26(3):256.
30. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013;45 (6):580–5.
31. Chen J, Li X, Zhong H, Meng Y, Du H. Systematic comparison of germline variant calling pipelines cross multiple next-generation sequencers. Sci Rep. 2019;9(1):9345.
32. Saunders CT, Wong WSW, Swamy S, Becq J, Murray LJ, Cheetham RK. Strelka: accurate somatic small-variant calling from sequenced tumor-normal sample pairs. Bioinformatics. 2012;28(14):1811–7.
33. Becker KG, Barnes KC, Bright TJ, Wang SA. The genetic association database. Nat Genet. 2004;36(5):431–2.
34. Wlasnowolski M, Sadowski M, Czarnota T, Jodkowska K, Szalaj P, Tang Z, et al. 3D-GNOME 2.0: a three-dimensional genome modeling engine for predicting structural variation-driven alterations of chromatin spatial structure in the human genome. Nucleic Acids Res. 2020;48 (W1):W170–6.
35. Keen G, Burton J, Crowley D, Dickinson E, Espinosa-Lujan A, Franks E, et al. The Genome Sequence DataBase (GSDB): meeting the challenge of genomic sequencing. Nucleic Acids Res. 1996;24(1):13–6.
36. Chou H-C, Acevedo-Luna N, Kuhlman JA, Schneider SQ. PdumBase: a transcriptome database and research tool for Platynereis dumerilii and early development of other metazoans. BMC Genomics. 2018;19(1):618.
37. Buza TM, Tonui T, Stomeo F, Tiambo C, Katani R, Schilling M, et al. IMAP: an integrated bioinformatics and visualization pipeline for microbiome data analysis. BMC Bioinform. 2019;20(1):374.
38. Krawczak M, Ball EV, Fenton I, Stenson PD, Abeysinghe S, Thomas N, et al. Human gene mutation database-a biomedical information and research resource. Hum Mutat. 2000;15 (1):45–51.
39. Varshney GK, Carrington B, Pei W, Bishop K, Chen Z, Fan C, et al. A high-throughput functional genomics workflow based on CRISPR/Cas9-mediated targeted mutagenesis in zebrafish. Nat Protoc. 2016;11(12):2357–75.
40. Dong Q, Schlueter SD, Brendel V. PlantGDB, plant genome database and analysis tools. Nucleic Acids Res. 2004;32(Database issue):D354–9.

41. Bolser DM, Staines DM, Perry E, Kersey PJ. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. In: van Dijk ADJ, editor. Plant genomics databases: methods and protocols, Methods in molecular biology. New York: Springer; 2017. p. 1–31. https://doi.org/10.1007/978-1-4939-6658-5_1.
42. Yu Y, Wang Y, Xia Z, Zhang X, Jin K, Yang J, et al. PreMedKB: an integrated precision medicine knowledgebase for interpreting relationships between diseases, genes, variants and drugs. Nucleic Acids Res. 2019;47(D1):D1090–101.
43. Fang Y, Quan J, Hua X, Feng Y, Li X, Wang J, et al. Complete genome sequence of Acinetobacter baumannii XH386 (ST208), a multi-drug resistant bacteria isolated from pediatric hospital in China. Genom Data. 2016;7:269–74.
44. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle. 2014;13(18):2847–52.
45. Boyle AP, Davis S, Shulha HP, Meltzer P, Margulies EH, Weng Z, et al. High-resolution mapping and characterization of open chromatin across the genome. Cell. 2008;132(2):311–22.
46. Li G, Cai L, Chang H, Hong P, Zhou Q, Kulakova EV, et al. Chromatin Interaction analysis with paired-end tag (ChIA-PET) sequencing technology and application. BMC Genomics. 2014;15(Suppl 12):S11.
47. Mirny LA. The fractal globule as a model of chromatin architecture in the cell. Chromosom Res. 2011;19(1):37–51.
48. Lowe R, Shirley N, Bleackley M, Dolan S, Shafee T. Transcriptomics technologies. PLoS Comput Biol. 2017;13(5):e1005457.
49. Yu L, Shao C, Ye X, Meng Y, Zhou Y, Chen M. miRNA Digger: a comprehensive pipeline for genome-wide novel miRNA mining. Sci Rep. 2016;6:18901.
50. Hills RD, Pontefract BA, Mishcon HR, Black CA, Sutton SC, Theberge CR. Gut microbiome: profound implications for diet and disease. Nutrients [Internet]. 2019 [cited 2021 Jan 5]; 11(7). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6682904/.
51. Arnold JW, Roach J, Azcarate-Peril MA. Emerging technologies for gut microbiome research. Trends Microbiol. 2016;24(11):887–901.
52. Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. Online Mendelian Inheritance in Man (OMIM), a knowledge base of human genes and genetic disorders. Nucleic Acids Res. 2005;33(Suppl_1):D514–7.
53. Cifaldi C, Chiriaco M, Matteo GD, Cesare SD, Alessia S, Angelis PD, et al. Novel X-linked inhibitor of apoptosis mutation in very early-onset inflammatory bowel disease child successfully treated with HLA-haploidentical hemapoietic stem cells transplant after removal of αβT and B cells. Front Immunol [Internet]. 2017 [cited 2021 Jan 5]; 8. Available from: https://moh-it.pure.elsevier.com/en/publications/novel-x-linked-inhibitor-of-apoptosis-mutation-in-very-early-onse-3.
54. ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. Nature. 2020;578(7793):82–93.
55. Lim Z-F, Ma PC. Emerging insights of tumor heterogeneity and drug resistance mechanisms in lung cancer targeted therapy. J Hematol Oncol. 2019;12(1):134.
56. Green ED, Watson JD, Collins FS. Human genome project: twenty-five years of big biology. Nature. 2015;526(7571):29–31.
57. Collins FS, Morgan M, Patrinos A. The human genome project: lessons from large-scale biology. Science. 2003;300(5617):286–90.
58. Jones KM, Ankeny RA, Cook-Deegan R. The Bermuda triangle: the pragmatics, policies, and principles for data sharing in the history of the human genome project. J Hist Biol. 2018;51 (4):693–805.
59. McEwen JE, Boyer JT, Sun KY, Rothenberg KH, Lockhart NC, Guyer MS. The Ethical, Legal, and Social Implications Program of the National Human Genome Research Institute: reflections on an ongoing experiment. Annu Rev Genomics Hum Genet. 2014;15:481–505.

60. Sekse C, Holst-Jensen A, Dobrindt U, Johannessen GS, Li W, Spilsberg B, et al. High throughput sequencing for detection of foodborne pathogens. Front Microbiol [Internet]. 2017 [cited 2021 Jan 5]; 8. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2017.02029/full.
61. Dong L, Wang W, Li A, Kansal R, Chen Y, Chen H, et al. Clinical next generation sequencing for precision medicine in cancer. Curr Genomics. 2015;16(4):253–63.
62. Sboner A, Mu XJ, Greenbaum D, Auerbach RK, Gerstein MB. The real cost of sequencing: higher than you think! Genome Biol. 2011;12(8):125.
63. Mardis ER. The $1,000 genome, the $100,000 analysis? Genome Med. 2010;2(11):84.
64. Moorthie S, Hall A, Wright CF. Informatics and clinical genome sequencing: opening the black box. Genet Med. 2013;15(3):165–71.

Mapping Algorithms in High-Throughput Sequencing

14

Manoj Kumar Gupta, Gayatri Gouda, S. Sabarinathan, Ravindra Donde, Ramakrishna Vadde, and Lambodar Behera

Abstract

The launch of high-throughput sequencing led to the production of billions of DNA fragments of several organisms from the vast array of the biological specimen in one run. Because of the significant rise in the sequences number, most of the analytical time that was earlier expended processing biological information is now devoted to identifying where the reads in the study come from. More reads are being mapped back to the reference sites, thus disclosing the type, quantity, and composition of DNA sequences present within the biological specimen. This stage, which includes the correct mapping of reads into a reference sequence, is vital since it decides how good the downstream analyses are. Thus, in this chapter, the authors attempt to understand the underlying mechanism of mapping algorithm and how they are employed in biological studies. Information retrieved suggested that the algorithms employed for mapping are known as mappers, whose key job is to locate the true position of each sequence/sequence alignments from a theoretically massive quantity of reference data while allowing for anomalies and structural variance. Based on index property, alignment

M. K. Gupta · G. Gouda · R. Donde · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_14

algorithms can be broadly categorized into three categories, i.e., algorithms that are focused on different index features, those that are based on various suffix trees, and eventually those that are based on merge sorting. These algorithms allow investigators to answer relevant science concerns, such as which genes are differentially expressed between environments.

Keywords

High-throughput sequencing · Next-generation sequencing · Hash table algorithms · Index-based algorithms

Abbreviations

BWT	Burrows-Wheeler Transform
HTS	High-throughput sequencing
NGS	Next-generation sequencing

14.1 Introduction

The launch of high-throughput sequencing (HTS), also known as deep sequencing or next-generation sequencing (NGS) during early 2000, offered an impetus for the production of billions of DNA fragments of several organisms from a vast array of the biological specimen in one run. Modern microarrays, by comparison, contain just a few hundred or even thousands of information per trial. Because of the significant rise in the sequences number, most of the analytical time that was early expended processing biological information is now devoted to identifying where the reads in the study come from. More reads are being mapped back to the reference sites, thus disclosing the type, quantity, and composition of DNA sequences present within the biological specimen [1]. This stage, which includes the correct mapping of reads into a reference sequence, is vital since it decides how good the downstream analyses are. The algorithms employed for this step are known as mappers. Mappers must be receptive and precise, thus resisting being too computationally taxing. They should be able to evaluate the true alignment of a read on a reference genome and differentiate between a sequencing error and a specific difference [2]. This, in turn, allows investigators to answer relevant science concerns, such as which genes are differentially expressed between environments [3], or how a specific genome varies from a reference genome [1]. These essential characteristics facilitate the ultra-deep sequencing technologies to be broadly employed in the field of biology and medical science [3].

In the initial days of NGS as well as sequencing through synthesis, researchers found that sequence alignment software tools common at the time were clearly not effective enough to evaluate NGS-scale data sets and were also not engineered for the challenge. Since then, a vast number of publications have identified new approaches to strengthen this issue. These methods started in numerous fields of research, including computer science, statistics, and mathematics. Few methods are oriented towards particular sequencing systems, which are not as common as other methods. These models fix various issues, such as performance, scalability, precision, and interpretability [1].

The popular mapping tools (Table 14.1 and Fig. 14.1) mainly focus on the hash table or index-based algorithms that are accurate and effective [4]. Index-based aligners are sluggish in operation, but they precisely align long gaps. On the other hand, heuristic-based algorithms are fast, use less power, and can be utilized for quick reads. While these algorithms have different fundamental structures, they are achieving almost similar performance [5]. Thus, selecting a correct mapping method before any downstream study is important since it would have an effect on the final outcome. A variety of benchmarking analyses exist to assist users in selecting aligners. For instance, Shang et al. tested the link between genome size and processing time [6]. Martin and the team checked the success of aligners on metagenomics results and established the shortcomings of the aligners with respect to the genome size and reference organism within a population such that each genus may be defined separately [7]. Thus, in this chapter, the authors attempt to examine the developments in different mapping or alignments algorithms and how they can be employed in various biological research. In this chapter, we summarize this extensive body of work, emphasizing crucial points that are relevant to researchers and practitioners. We also try to discuss alignment approaches that often tackle the statistical issue of read mapping.

14.2 Overview of Mapping Algorithm

"The HTS data mapping problem can be generally stated as follows: given a set of sequences Q (produced by an HTS technology), a set of reference sequences R, a possible set of constraints and a distance threshold k, find all substrings m of R that respect the constraints and that are within a distance k to a sequence q in Q, i.e. d(q, m) $<K$, where $d()$ is some distance function. The occurrences m in R are called *matches*. The constraints imposed can vary depending on the HTS application and data type (e.g. whether the data generated are single- or paired-end reads)" [8]. A mapper's key job is to locate the real position of every sequence/sequence alignments q from a theoretically massive amount of reference information while permitting for structural variance and anomalies. For ensuring the right matches, the correct position should be estimated. The distance metrics are usually employed for estimating the number of mismatches and indels to compensate for inconsistencies as well as sequence variance; however, they can often include different sizes or possibilities involved with the reads [8].

The most apparent limitation of selecting a mapper is the data form that it is built for or is appropriate for managing (miRNA, RNA, DNA, or bisulphite). Another significant aspect to remember is the sequencing platform used for data production.

Table 14.1 Features of few important aligners (Adapted from [6])

Aligners	Operate system	Programming language	Input Format[a]? (Fasta and Fastq)	Output format	Multithread?	Gapped alignment?	Paired-end alignment?	Trimming alignment?	Bisulfite alignment?	Note
Bowtie	★	C++	√	SAM	√		√	√		Maximum allowed mismatches ≤3
BWA	◎	C++	√	SAM	√	√	√			BWA-short: 200 bp; BWA-SW: 100 kbp
BOAT	◎	C	√	*	√	√				Maximum allowed mismatches ≤3
GASSST	◎	C++	Fasta	SAM	√	√				Merely Fasta format required for reads
Gnumap	◎	C	√ (prb)	SAM	√	√		√	√	Maximum read length <1000 bp
GenomeMapper	◎	C	√	BED	√	√				Maximum read length <2000 bp
mrFAST	★	C	√	SAM		√	√			Maximum read length <300 bp
mrsFAST	★	C	√	SAM			√		√	Maximum read length <200 bp
MAQ	◎	C++	Fastq	Map			√			Maximum read length ≤128 bp

NovoAlign	●	C++	√	SAM	√	√	√	√	√	Restrictions for academic version
PASS	※	C++	√ (sff)	GFF3	√	√	√			Maximum read length <1000 bp
PerM	※	C++	√	SAM	√		√	√		Maximum read length ≤128 bp
RazerS	★	C++	√ (prb)	Eland, GFF		√	√	√		Arbitrary read length
RMAP	◎	C++	√	BED			√		√	Fixed-length reads required
SeqMap	★	C++	Fasta	Eland		√				Maximum allowed mismatches ≤5
SOAPv2	◎	C++	√	*	√	√	√			Maximum read length <1000 bp
SHRiMAP2	◎	Python	Fasta	SAM	√	√	√			Parallel computing supported
Segemehl	◎	C	Fasta	*	√	√	√	√	√	Large memory usage required
SSAHA2	●	NA	√	GFF, SAM			√			For long reads mapping

※ Windows, Linux, or Unix operating system
★ Windows, Linux, Unix, or Mac X operating system
● Linux, Unix, or Mac X operating system
◎ Linux or Unix operating system

* The short-read aligning algorithms' own output format

[a]We consider only short-reads input format here

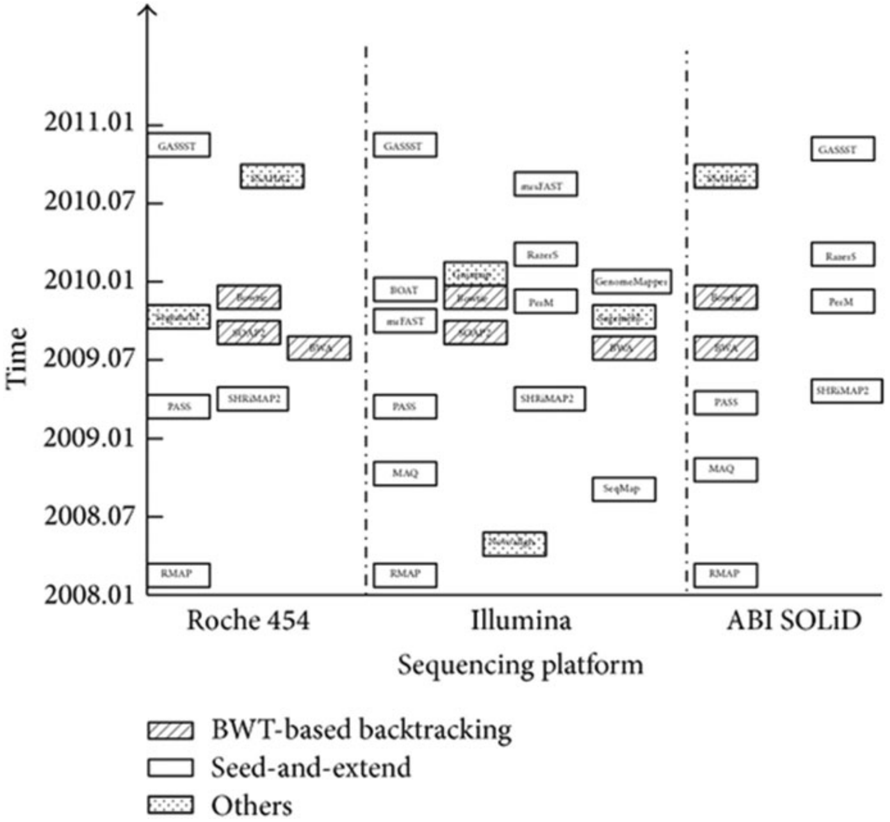


Fig. 14.1 Aligners based on algorithms classification across different NGS platforms. Rectangles with different grey scales represent hash table-based algorithm, BWT-based backtracking algorithm, and other algorithms, individually. Aligners for specific types of data generated by different sequencing platforms are separately shown in three columns, namely, Roche 454, Illumina, and ABI SOLiD. (Adapted from [6])

Sequence alignment systems, like Mummer, Exonerate, and BLAT, may match all sequences from any source (e.g., Protein DNA, or RNA) and irrelevant to the source form. The slider was developed for use with Illumina data and is better used in the light of Illumina's raw data output. In comparison, MapReads, SOCS, and RNA-Mate are better adapted for aligning reads that are colour-coded. Few mappers can manipulate particular prejudices while designing their own job. The Illumina sequencing platform has a range of limitations, which means that less precise base calls are generated towards each read's 3′-end. Some alignment algorithms, such as SOAP, can, therefore, trim the 3′-end of reads in order to resolve this issue [8].

Eukaryotic genes are comprised of several exons and can be fused altogether for producing several transcript sequences. As a consequence, RNA-seq sequences would be matched against a reference genome, with reads that cover several exons providing gaps within the alignment conforming intronic sequence. Earlier studies have suggested that the identification of splice junctions is rendered either de novo or from user-defined junction positions library. Using de novo splicing identification, mappers can identify splice junctions easily without any current annotation. An alternative solution is to construct exon junction libraries that involve regions covering established or expected splicing junctions. Some mappers build these libraries throughout execution utilizing splice junction knowledge supplied by the user, while some require that the user supplies the library. And, hybrid methods that integrate fresh information with existing results are also feasible. In the first stage, QPALMA distinguishes clusters of mapped reads correlated with exons through aligning the reads to their genomic position. Next possible exonic junctions over a certain span around the putative exons are enumerated. Sequences flanking possible junctions are mapped to the unmapped readings, rendering it easier to find new junctions. Similar to TopHat, QPALMA differs by training a support vector machine-like algorithm using the known genome splice junctions. Therefore, a collection of recognized junctions from the reference are also needed [8]. Thus, most of the fast alignment algorithms need auxiliary data structures, termed indices, for the read sequences or reference sequences, or often both. Based on index property, alignment algorithms can be broadly categorized into three categories, i.e., algorithms that are focused on different index features, those that are based on various suffix trees, and eventually those that are based on merge sorting [9].

14.2.1 Algorithms Based on Hash Tables

Hash-based genome-searching programs utilize hash tables (Fig. 14.2) [10]. A hash table is an index in which data is sorted by keys. Thus, a hash table is a form of an associative list. Two separate approaches for mapping short reads onto genomes are explored in the chapter. One stores short reads' subsequence, while the other stores subsequences of the genome as well as its location within a hash table. As there is no important distinction in their usage of hash, we discussed in detail in three different approaches (i–iii). A hash-based approach prepares a table in which the locations of target subsequences are the keys, and the target genome positions are the corresponding values. A short DNA sequence is mapped onto the genome where the resulting key is used to index a hash table [10].

Three approaches exist for identifying the *n*-mismatch genome location of the subsequence of length l with the hash table.

14.2.1.1 Refer to all *n*-Mismatch Subsequences

Prepare a hash table whose key length is *l*, and use the subsequence and its *n*-mismatch subsequences as keys to refer to the table. It requires $\sum_{i-1}^{n} l^{C_n 3l}$ hash references to find all the *n*-mismatch genome positions.

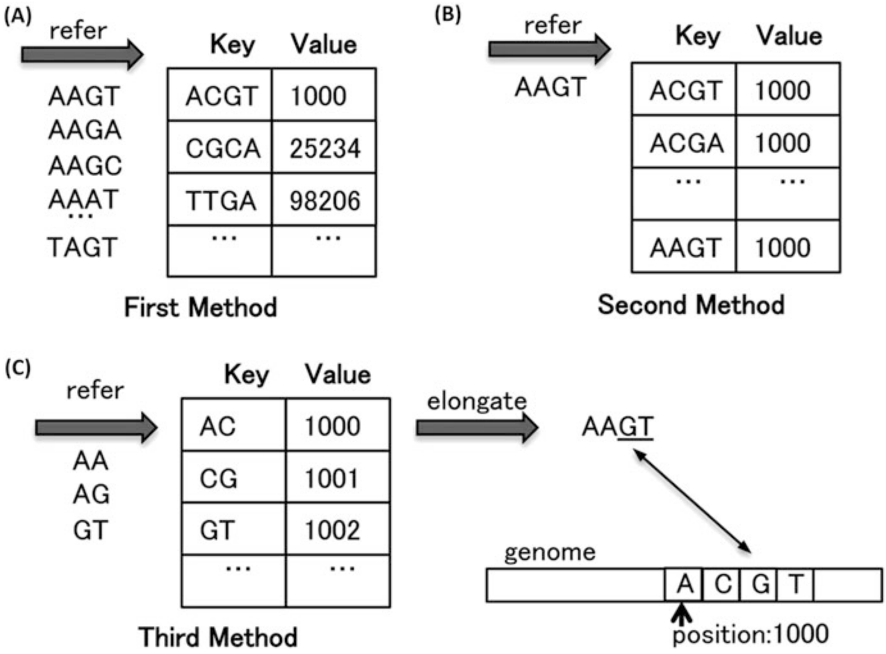


Fig. 14.2 Three methods to find genome positions of 1-mismatch from the subsequence AAGT. Genome position 1000 is ACGT, which is the 1-mismatch of the subsequence. (**a**) The first method refers to the hash table 16 times. (**b**) The second method refers to the table just once, but the table is 16-fold larger. (**c**) The third method refers to the table three times. After getting position 1002 from the hash table, the method elongates the alignment toward the front of the sequence. (Adapted from [10])

14.2.1.2 Store *n*-Mismatch Positions in the Hash Table

For each position of the subsequence of the genome, store the position $\sum_{i-1}^{n} l^{C_n 3l}$ times. The hash keys are the subsequence and its *n*-mismatch subsequences.

14.2.1.3 Use Pigeonhole Principle; Combine Hash Table and Another Method

Generate a hash table whose key length is $\lfloor l/n \rfloor$. After getting the perfect-match genome position of length $\lfloor l/n \rfloor$ by referring to the hash table, find *n*-mismatch sequences by another method, such as dynamic programming or BWT [10].

The first and second approaches require so many hashes and a large hash table, respectively. The third approach is the strongest; however, when *n* increases, the capacity to reduce the genome region down becomes weak, and thus the pressure of the post-processing to identify *n* mismatches rises. To resolve these difficulties and boost the efficacy of using hash tables for genome mapping, technological breakthroughs were required. In 2011, Takenaka and the team introduced a method for finding the genomic locations of two or more mismatches within a hash table without expanding the scale of the hash table [10]. To carry out the process, 4-ary perfect Hamming code is employed. This is a hash-based approach for genome mapping that decreases the amount of hash references for detecting mismatches without maximizing the hash table scale. The approach includes defining DNA subsequences as terms on a Galois extension field GF(2^2) and assembling them into a complete, unambiguous Hamming code. The perfect Hamming code specifies what equivalent subsequences of nucleotides are. They proposed "a hash-based method for genome mapping that reduces the number of hash references for finding mismatches without increasing the size of the hash table. The method regards DNA subsequences as words on Galois extension field $GF(2^2)$ and each word is encoded to a code word of a perfect Hamming code. The perfect Hamming code defines equivalence classes of DNA subsequences. Each equivalence class includes subsequence whose corresponding words on GF(2^2) are encoded to a corresponding code word. The code word is used as a hash key to store these subsequences in a hash table. Specifically, it reduces by about 70% the number of hash keys necessary for searching the genome positions of all 2-mismatches of 21-base-long DNA subsequence" [10].

14.2.2 Algorithms Based on Suffix/Prefix Tries

In this group, all algorithms effectively decreases the inaccurate matching issue to the exact matching problem and indirectly require two steps: detection of exact matches and creation of inaccurate alignments backed by correct matches. For locating identical match, these algorithms depend on some representations of suffix/prefix trie, like enhanced suffix array, suffix tree, and FM-index [9]. The benefit of utilizing a trie is that several identical copies alignment of a substring within the reference is only expected to be achieved once since these identical copies fall on a common path in the trie, while for a traditional hash table index, an alignment must be conducted for each replica. The optimal choice of data structures is not based on the system used to locate inexact matches. This form of algorithm will also probably work for suffix tree index in theory [9].

14.2.2.1 Trie, Prefix/Suffix Tree and FM-Index

A suffix tree is a data structure that holds the complete set of suffixes of a sequence, which enables quick scans of a string [9]. To create the relationship amongst a trie as well as an FM-index, a data structure dependent on Burrows-Wheeler Transform (BWT), we concentrate on prefix trie that would be the reverse sequence trie. Complete algorithms on a prefix trie can be smoothly added to the related prefix trie. "The time complexity of determining if a query has an exact match against a trie is linear in the length of the query, independent of the length of the reference sequence. However, a trie takes $O(L^2)$ space where L is the length of the reference. It is impractical to build a trie even for a bacterial genome. Several data structures are proposed to reduce the space. Among these data structures, a suffix tree is most widely used. It achieves linear space while allowing linear-time searching. Although it is possible in theory to represent a suffix tree in $L \log_2 L + O(L)$ bits using rank-selection operations, even the most space efficient implementation of bioinformatics tools requires 12–17 bytes per nucleotide, making it impractical to hold the suffix tree of the human genome in memory" [9]. To address these issues, Abouelhoda and the team [11] built an improved suffix array that is an expanded suffix array and an array of auxiliary arrays, consuming 6.25 bytes per nucleotide. It can be considered as an implied depiction of the suffix tree, which has an identical time complexity to suffix tree in seeking exact matches, better than what Manber and Myers [12] originally created.

Another enhancement to memory is accomplished by Ferragina and Manzini, who introduced the FM-index and noticed that finding a child of a parent node mostly in prefix trie can be performed within constant time employing a backwards scan on this data structure [9]. Therefore, the time complexity to locate exact matches in a hash is the same as that in a trie. The FM-index is a compact data structure constructed, so the index size is smaller than the initial string where there are repetitive characters in the string (equivalently, has small entropy). The FM-index is most definitely not compact in accordance with since-DNA sequences have a limited alphabet. The memory footprint for an FM-index is usually 0.5–2 bytes per nucleotide, based on the implementation and software parameters. The human genome only makes up 2–8 GB of memory.

14.2.2.2 Identifying Inexact Matches Employing a Suffix/Prefix Trie

Of the numerous available algorithms for query-reference alignment, SOAP2 [13], Segemehl [14], BWA [15], OASIS [16], Vmatch [11], BWT-SW [17], MUMmer [16], Bowtie [18], and BWA-SW [19] are the most common. When constructed correctly, one graph of a trie may be conveniently converted into another. The FM-index is more often used since its memory footprint is smaller. The algorithms for inexact matching are based on maximal matches, maximal unique matches, exact matches, or maximal repeats, and these have been joined with gap-dependent alignment. Likewise, Segemehl activates the alignment having the longest prefix match from each suffix, and may even illustrate mismatches as well as gaps at some locations of the query for reducing false alignments [9]. With OASIS and BWT-SW, they are able to efficiently extract substrings of the reference via a top-down traversal on the trie and match these substrings against the query via dynamic programming. BWA-SW goes beyond BWT-SW by describing the demand as a directed word graph (DAWG) [20], which often permits it for deploying heuristics to accelerate alignment. BWA and Bowtie generally sample the reference's short substrings; however, rather than doing dynamic programming, they evaluate the query and the sampled substrings to allow just a few variations. In conjunction, since they enable the whole read to be matched, the traversal of the trie should be bounded because it is wasteful to descend deeper throughout the trie if it can be expected that doing so contributes to an alignment with substantial mismatches as well as gaps. Conversely, BWA and Bowtie should be considered to explicate all combinations of potential mismatches as well as gaps within the query sequence so that the modified query can be precisely matched [9].

14.3 Choice of Mapper

The sequencing platforms differ in terms of the reads generated per run and mean length of the read [21]. For instance, the SOLiD & Illumina systems generate shorter reads with higher throughput. The Roche/454 and PacBio systems produce longer reads with lower throughput. The throughput and read length are the two key specification estimates in the ads for HTS instruments, and the connection between them is stressed. In ideal situations, we would prefer to execute long reads in massive concentrations, but current technology allows for either long reads, which are short in number, or short reads, which are high in number. Long reads are not necessarily appropriate if the experiment cannot effectively sequence sufficient fragments. For instance, CHiP-Seq only returns small size pieces of DNA (~100 nt), so to recognize peaks of mapped read clusters, which reflect the binding sites, high throughput is needed [21].

The length and the throughput of the sequencing system also have to be addressed when evaluating the samples. In addition to the prejudices introduced through the experiment at hand, unique biases by the laboratory design and sequencing procedure must also be recognized. The three popular forms of error during NGS analysis are insertions, deletions, and miscalling of bases within the read [22]. Each high-throughput sequencer has a particular error profile, and this is due to the techniques involved. However, the systems will measure the likelihood of errors while interpreting the base sequences, which can be used in the FASTQ format. Illumina sequencers have a very less indel error rate, but here substitution errors are the most common form of error. For most sequencing platforms, errors escalate as the duration of the reads grows [23, 24]. In reads generated from the Roche/454 FLX Titanium, the frequency of indel errors is almost nearly ten times higher in comparison to substitution errors and errors that generate near the read end [25]. One method for reducing the error rate is to eliminate the read end that has poor base qualities. While this approach decreases the length of the read; it increases the sequence data accuracy and decreases the number of mismatches required for mapping a read. Compared with other technologies such as shotgun sequencing, PacBio reads demonstrate a more stable distribution of error, as well as the most dominant errors, are insertions [26, 27]. The Ion PGM system has a higher incidence of indel errors compared to replacement, and the coverage dropped for AT-rich regions as well [24]. It is observed that there is a trend among Illumina sequencing to have higher coverage of GC-rich areas [28]. The experimental approach may also add variation into a sequencing experiment, for instance, cross-linking within PAR-CLIP can incorporate T to C translation within the reads. Other important factor is the ligated adapters, which will be part of the reads if the length of the DNA fragment is longer than that of the read length. The adapters must be discarded prior to mapping completion, which may be problematic if the sequenced reads have been changed by sequencing errors [21].

Thus, the selection of a mapping software depends on a lot of considerations (Fig. 14.3 and Table 14.2), such as sequencing technologies, the volume of data to be mapped, and the form of experiment. A big problem is that systems are developed

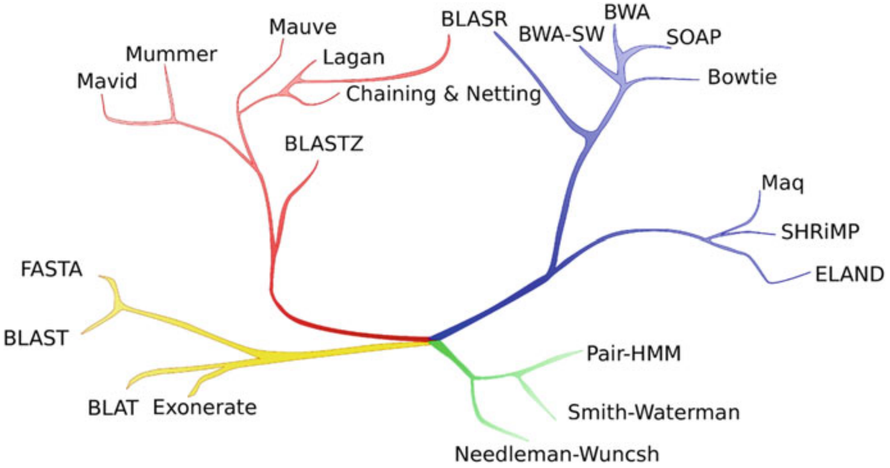


Fig. 14.3 "An illustration of relationships between alignment methods. The applications/corresponding computational restrictions shown are (green) short pairwise alignment/detailed edit model; (yellow) database search/divergent homology detection; (red) whole genome alignment/alignment of long sequences with structural rearrangements; and (blue) short read mapping/rapid alignment of massive numbers of short sequences. Although solely illustrative, methods with more similar data structures or algorithmic approaches are on closer branches. The BLASR method combines data structures from short read alignment with optimization methods from whole genome alignment." (Adapted from [27])

for mapping partly long reads or entire genome sequences. Since the length of the read is associated with the read number, fast read mappers must rely on speed instead of precision in order to make the most of the query reads [21]. Common methods for shorter reads (~200 nt) are Bowtie, BWA, and Novoalign [29]. In the last 10 years, further mapping programs have been released to the public. One must be mindful that there may be significant variations within the mapped reads, as well as the mapping positions amongst programs. It is also suggestive of attempting numerous programs with a given reads data. When analyzing lengthy readings, BWA-SW and BLAT may be employed. BWA-MEM substituted BWA-SW [30]. Nevertheless, BWA-MEM, BWA, and BWA-SW are all available through the same app kit. Pacific Biosciences notes that they support the BLASRmapper program long read produced via RS2 & RS instruments [27]. Few other probabilistic mapping were also reported, where the quality scores of per base are being employed within a probabilistic model for measuring the likelihood of correctly mapping, e.g. in MAQ [31], Stampy [32], and LAST [33].

Table 14.2 Assessment and comparative study of various important aligners (Adapted from [6])

	Computational speed				Memory usage	Accuracy				
Aligners	Speed with single thread	Speed with multithread	Key factor impacting speed (genome size or read count)	Overall evaluation	Key factor impacting memory (genome size or read count)	Memory usage with multithread	Sensitivity	Precision	% of multimapped	% Corrected multi-mapped
Bowtiel	Fast	↑	Genome size	Low	Genome size	≡	High	–	–	
BWA	Fast	↑	Both	Low	Genome size	≡				
BOAT	Slow	↑↑	Genome size	Low	Read count	↑↑	High	–	–	Low
GASSST	–	↑	Genome size	High[a]	Genome size	≡	Low	High	–	
Gnumap	Slow	↓	Genome size	High[a]	Genome size	≡				
GenomeMapper	Slow	≡	Genome size	Low[b]	Genome size	≡	High	–	–	
mrFAST	Slow	×	Genome size	High[a]	Read count	×	High	–	–	
mrsFAST	–	×	Genome size	Low	Read count	×	High	–	–	
MAQ	–	×	Genome size	High[a]	Read count	×				

(continued)

Table 14.2 (continued)

	Computational speed				Memory usage	Accuracy				
Aligners	Speed with single thread	Speed with multithread	Key factor impacting speed (genome size or read count)	Overall evaluation	Key factor impacting memory (genome size or read count)	Memory usage with multithread	Sensitivity	Precision	% of multimapped	% Corrected multi-mapped
NovoAlign[c]	–	/	Read count	Low[b]	Genome size	/	High	High	Low	Low
PASS	–	↑	Genome size	Low[b]	Genome size	↑	High	High	Low	Low
PerM[d]	Fast		Genome size	Low[b]	Genome size	/	Ind: low	–	Low	
RazerS	Slow	×	Genome size	High[a]	Read count	×	High	–	–	
RMAP	–	×	Genome size	High[a]	Genome size	×	Mis: low	High	Low	
SeqMap	–	×	Genome size	High[a]	Read count	×	High	–	–	
SOAPv2	Fast	↑	Genome size	Low	Genome size	≡	High	High	Low	
SHRiMAP2	Slow	↑	Genome size	High[a]	Genome size	↑	High	Low	High	

Segemehl	–	↑	Both	High[a]	Genome size	≡	High	–		

For computational speed, we defined the aligners which are extremely faster than others as fast, while we defined the ones which are extremely slower as slow

×: without multithread function. — represents medium level remark. ≡ means there is no obvious change

[a] For memory usage, we evaluated the aligners as follows: among the seven datasets, the maximum memory usage ≤4 GB, low; the maximum memory usage ≥32 GB, high

[b] Low represents that the maximum memory usage will have an extreme increase with *H. sapiens* datasets (≥4 GB)

[c] Novoalign could support multithread only for commercial version

[d] PerM could adjust the threads automatically during running process

14.4 Conclusion and Future Perspective

In conclusion, we need reliable and fast aligners to manage the exponentially growing amount of sequencing data. In this chapter, we discussed in detail about the various algorithms for aligning short reads and long reads which achieve high accuracy in both cases. The advancement of modern sequencing technologies significantly increases the scale and precision of many biological applications, including the scanning of genome-wide heterogeneity, the discovery of protein binding sites, the quantification of the transcriptome, the detection of genome-wide methylation sequence, and the assembly of new genome or transcriptome. Given the stunning successes of NGS in genomics and post-genomics, three significant obstacles that are faced via mapping or alignment algorithms and have hindered the advancement of these technologies into full maturity are the computational challenge, the operational challenge, and the cross-platform unification problem [3]. The rising difference between the computing capacity of next-generation sequencing and the tools available to interpret the data is a critical problem requiring immediate attention. The method of aligning more and more reads against a broad genomic sequence is very popular in today's genomic studies. However, machines that can accommodate the relentless computing demands are not cost-effective for any person. Time control is an inevitable obstacle when doing NGS jobs. Therefore, an incredibly accurate algorithm is needed to minimize computer costs. Parallelization techniques, including BWT algorithm implemented via Bowtie, BWA, and SOAP2, have been suggested and attempted to enable aligners to speed up their execution time and reduce their machine memory requirement with uncompromising results precision [3].

Since NGS technologies are constantly evolving, developers of quick reads mapping and assembly applications must keep up with the ever-changing cutting-edge technologies involved. To hold up or even overtake Sanger sequencers in terms of reading time, which is a very significant parameter that counts for detecting split mapping signatures as well as de novo sequencing, NGS sequencing devices all aim to generate longer readings [3]. This ensures that quick reads or next-generation sequencers with larger genomes can be designed to be consistent with longer reads. Furthermore, unknown data formats generated from the so-called next–next-generation sequencers, like Helicos HeliscopeTM & Pacific Biosciences SMRT, an enormous mass of various experiments, and varied scale of study all call for more stable and effective algorithms in automatically redressing parameters for particular demands [3]. Another key difficulty faced by developers of NGS mappers and assemblers arises from the specification's inconformity in size of inserts across mates, error profiles and "true match" benchmarks across different NGS platforms. Different types of inserts, which are popular in variant NGS platforms, often provide different effectiveness in detecting variations. Shorter insert sizes, contrasted with long inserts (which give benefits in detecting larger incidents), improve the exposure of smaller events [34–36]. Therefore, a mixture of several libraries of differing insert sizes would be a reasonable option in potential studies [35, 37]. Additionally, since various platforms generate reads of different error models and often isolate "real alignment" from several potential matches of their own criterions, authors are sometimes disappointed as they investigate the data from many platforms. Therefore, a standardized standard for deciding authentic matches should be defined, and quality management of the data should be controlled [38]. In addition, considering that "NGS users are always puzzled by a complicated maze of base calling, alignment, assembly, and analysis tools with often incomplete documentation and providing no ideas on how to compare and validate the outputs, Paul Medvedev and the team [34] recommended that new methods should combine the previous approaches and possess different types of signatures to support an event". Challenges will constantly occur for further creation of NGS. Efforts need to be given to mapping and assembly as well as downstream analysis, like metagenomics, small RNA detection, and transcriptome analysis. New and still unexamined considerations, as well as questions, will keep emerging, and thereby novel programs must evolve to keep up with the pace of NGS and modifications in the adoption of these techniques.

Conflicts of Interest None

Additional Information Fig. 14.1 (CC BY 3.0) [6], Fig. 14.2 (CC BY 2.0) [10], Fig. 14.3 (CC BY 2.0) [27], Tables 14.1 and 14.2 (CC BY 3.0) [6] have been used under the terms of the Creative Commons Attribution License.

References

1. Reinert K, Langmead B, Weese D, Evers DJ. Alignment of next-generation sequencing reads. Annu Rev Genomics Hum Genet. 2015;16(1):133–51.
2. Caboche S, Audebert C, Lemoine Y, Hot D. Comparison of mapping algorithms used in high-throughput sequencing: application to ion torrent data. BMC Genomics. 2014;15(1):264.
3. Bao S, Jiang R, Kwan W, Wang B, Ma X, Song Y-Q. Evaluation of next-generation sequencing software in mapping and assembly. J Hum Genet. 2011;56(6):406–14.
4. Lindner R, Friedel CC. A comprehensive evaluation of alignment algorithms in the context of RNA-Seq. PLoS One. 2012;7(12):e52403.
5. Hatem A, Bozdağ D, Toland AE, Çatalyürek ÜV. Benchmarking short sequence mapping tools. BMC Bioinform. 2013;14(1):184.
6. Shang J, Zhu F, Vongsangnak W, Tang Y, Zhang W, Shen B. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. BioMed Res Int. Hindawi; 2014 [cited 2021 Jan 10]. p. e309650. Available from: https://www.hindawi.com/journals/bmri/2014/309650/.
7. Martin J, Sykes S, Young S, Kota K, Sanka R, Sheth N, et al. Optimizing read mapping to reference genomes to determine composition and species prevalence in microbial communities. PLoS One. 2012;7(6):e36427.
8. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012;28(24):3169–77.
9. Li H, Homer N. A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform. 2010;11(5):473–83.
10. Takenaka Y, Seno S, Matsuda H. Perfect hamming code with a hash table for faster genome mapping. BMC Genomics. 2011;12(3):S8.

11. Abouelhoda MI, Kurtz S, Ohlebusch E. Replacing suffix trees with enhanced suffix arrays. J Discrete Algorithms. 2004;2(1):53–86.
12. Manber U, Myers G. Suffix arrays: a new method for on-line string searches. SIAM J Comput. 1993;22(5):935–48.
13. Li R, Yu C, Li Y, Lam T-W, Yiu S-M, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966–7.
14. Hoffmann S, Otto C, Kurtz S, Sharma CM, Khaitovich P, Vogel J, et al. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. PLoS Comput Biol. 2009;5(9):e1000502.
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25(14):1754–60.
16. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12.
17. Lam TW, Sung WK, Tam SL, Wong CK, Yiu SM. Compressed indexing and local alignment of DNA. Bioinformatics. 2008;24(6):791–7.
18. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. 2009;10(3):R25.
19. Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010;26(5):589–95.
20. Blumer A, Blumer J, Haussler D, Ehrenfeucht A, Chen MT, Seiferas J. The smallest automation recognizing the subwords of a text. Theor Comput Sci. 1985;40:31–55.
21. Frellsen J, Menzel P, Krogh A. 6.03—Algorithms for mapping high-throughput DNA sequences Jes Frellsen and Peter Menzel contributed equally. In: Brahme A, editor. Comprehensive biomedical physics. Oxford: Elsevier; 2014 [cited 2021 Jan 10]. p. 41–50. Available from: http://www.sciencedirect.com/science/article/pii/B9780444536327011035.
22. Cock PJA, Fields CJ, Goto N, Heuer ML, Rice PM. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res. 2010;38(6):1767–71.
23. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012. Hindawi; 2012 [cited 2021 Jan 10]. p. e251364. Available from: https://www.hindawi.com/journals/bmri/2012/251364/?utm_source=google&utm_medium=cpc&utm_campaign=HDW_MRKT_GBL_SUB_ADWO_PAI_DYNA_JOUR_X_PJ_GROUP3&gclid=EAIaIQobChMIrda9rNmQ7gIVwdmyCh0HkgrEEAAYBCAAEgLQV_D_BwE.
24. Quail MA, Smith M, Coupland P, Otto TD, Harris SR, Connor TR, et al. A tale of three next generation sequencing platforms: comparison of ion torrent, Pacific Biosciences and Illumina MiSeq sequencers. BMC Genomics. 2012;13(1):341.
25. Gilles A, Meglécz E, Pech N, Ferreira S, Malausa T, Martin J-F. Accuracy and quality assessment of 454 GS-FLX titanium pyrosequencing. BMC Genomics. 2011;12(1):245.
26. Carneiro MO, Russ C, Ross MG, Gabriel SB, Nusbaum C, DePristo MA. Pacific biosciences sequencing technology for genotyping and variation discovery in human data. BMC Genomics. 2012;13(1):375.
27. Chaisson MJ, Tesler G. Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory. BMC Bioinform. 2012;13 (1):238.
28. Benjamini Y, Speed TP. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. 2012;40(10):e72.
29. Menzel P, Frellsen J, Plass M, Rasmussen SH, Krogh A. On the accuracy of short read mapping. In: Shomron N, editor. Deep sequencing data analysis [internet]. Methods in molecular biology. Totowa, NJ: Humana press; 2013 [cited 2021 Jan 10]. p. 39–59. Available from: https://doi.org/10.1007/978-1-62703-514-9_3.

30. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv13033997 Q-Bio [Internet]. 2013 [cited 2021 Jan 10]; Available from: http://arxiv.org/abs/1303.3997.
31. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res. 2008;18(11):1851–8.
32. Lunter G, Goodson M. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res. 2011;21(6):936–9.
33. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462 (7271):315–22.
34. Medvedev P, Stanciu M, Brudno M. Computational methods for discovering structural variation with next-generation sequencing. Nat Methods. 2009;6(11):S13–20.
35. Amlot PL, Grennan D, Humphrey JH. Splenic dependence of the antibody response to thymus-independent (TI-2) antigens. Eur J Immunol. 1985;15(5):508–12.
36. Bashir A, Volik S, Collins C, Bafna V, Raphael BJ. Evaluation of paired-end sequencing strategies for detection of genome rearrangements in cancer. PLoS Comput Biol. 2008;4(4): e1000051.
37. Campbell PJ, Stephens PJ, Pleasance ED, O'Meara S, Li H, Santarius T, et al. Identification of somatically acquired rearrangements in cancer using genome-wide massively parallel paired-end sequencing. Nat Genet. 2008;40(6):722–9.
38. Pop M, Salzberg SL. Bioinformatics challenges of new sequencing technology. Trends Genet TIG. 2008;24(3):142–9.

DNA–Protein Interaction Analysis 15

Piyali Goswami

Abstract

DNA–protein interaction study has always been a topic of interest for scientists as it is crucial for the cell to survive, differentiate, and divide. It helps the cell to conduct different major activities like replication, transcription, translation, DNA repair and recombination, and RNA processing. A lot of research has already been conducted to understand how the DNA–protein interaction occurs inside the cell. In this chapter, we discuss the major requirements for DNA–protein interaction to occur, several detection methods for DNA–protein interaction, their limitations, and advantages. This study will give a better understanding of the vital biological processes occurring inside the cell to understand the process of cell growth, development, and disease.

Keywords

DNA · Protein · Sequence-specific interaction · Non-specific interaction · DNA-binding motifs · EMSA · DNA footprinting · X-ray · NMR · CHIP · DNA pull-down

Abbreviations

bZIP	Basic leucine zipper
CHIP	Chromatin immunoprecipitation assays
EMSA	DNA electrophoretic mobility shift assays
LEF1	Lymphoid enhancer binding factor 1

P. Goswami (✉)
Department of Biotechnology, IIT Kharagpur, Kharagpur, West Bengal, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_15

Myt1	Myelin transcription factor
ST18	Suppressor of tumourigenicity protein18

15.1 Introduction

DNA and protein are both vital components of the cell. DNA contains all the genetic information of the cell, which are transmitted from one cell to the other through replication. DNA–protein interaction controls many crucial biological processes in the cell apart from replication, such as transcription, translation, and DNA repair. These DNA–protein interactions may be specific or non-specific in nature. There are various factors which determine the specificity of the interaction like site specification, recognition, affinity, and equilibrium selection [1].

Specific interactions involve the protein binding at specific DNA binding sites known as DNA binding domains and motifs which are discussed later in this chapter. This type of interaction is generally mediated by various transcription factors through hydrogen bond, hydrophobic interaction, ionic interaction, and Van der Waals force [2, 3]. The major groove of DNA is generally involved in the sequence-specific binding of the protein as it provides a larger area to be accessed by the DNA-binding proteins [4].

Non-specific interactions are carried out by histones and are not at all dependent on the nucleotide sequence. Here, interactions occur through the functional groups of the protein with the sugar-phosphate backbone of the DNA through ionic bonds.

DNA–protein interaction studies date back to the nineteenth century when the scientists identified an interaction between the DNA and protein through microscopic studies. Since then, a lot of research has been conducted to understand the association of proteins with DNA and to identify the proteins involved in this process, as it controls the structure and function of DNA.

15.2 DNA Binding Domains and Motifs

Domains are conserved sequences which can fold, evolve, and act independently irrespective of the rest of the protein sequence, whereas motifs are super secondary structures made up of sequentially arranged secondary structures [5, 6]. Motifs may have distinct function or may be a part of the protein domain.

To conduct the protein–DNA interaction, the protein identifies "specific segments" in the DNA as the binding region. The regions of the proteins responsible for binding these "specific segments" are known as DNA binding domains. The DNA binding domains must have at least one motif which has the ability to recognize both single-stranded and double-stranded DNA. The DNA binding may be either specific like transcriptional factors or non-specific like histones. The major

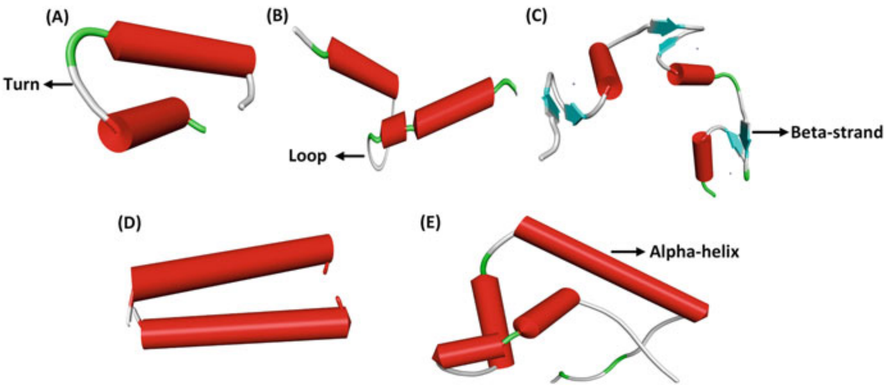


Fig. 15.1 (**a**) Helix-turn-helix, (**b**) helix-loop-helix, (**c**) zinc finger, (**d**) leucine zipper, and (e) HMG-box

motifs (Fig. 15.1) responsible for specific recognition of DNA sequence has been discussed below:

15.2.1 Helix-Turn-Helix

Helix-turn-helix is a common DNA binding motif of both prokaryotes and eukaryotes. It is about 20 amino acids in length with 2 alpha helices interconnected by a short linker region of 3 amino acids. In majority of the cases, the second alpha helix near the C-terminal region serves as the "recognition helix" which binds the major groove of DNA through Van der Waals forces and hydrogen bonds. The other helix near the N-terminal end helps to position and stabilize the interaction between the recognition helix and the DNA. Helix-turn-helix can be classified into several types depending upon their structure and spatial arrangement, namely di-helical, tri-helical, tetra- helical, winged helix-turn-helix, and other modified helix-turn-helix motifs. These types of motifs can be found in several regulatory proteins like Cro, CAP, and λ repressor [7–9].

15.2.2 Helix-Loop-Helix

Helix-loop-helix are found in eukaryotic proteins ranging from yeast to humans. It is about 50–60 amino acids in length, made up of 2 alpha helices interconnected by a loop of amino acid variable in length. Generally, one helix is smaller than the other which folds back to the larger one and forms a dimer due to the flexibility of the interconnecting loop. These types of motifs are characteristic features of dimeric transcription factors. They can form both homo- and heterodimers. The helix-loop-helix motif was first identified by Murre et al., in two murine transcription factors E12 and E47 [10]. The DNA-binding region comprises of about 13 highly basic amino acids. Proteins with this motif typically bind to the major groove of DNA at a consensus hexanucleotide sequence known as the E box, CANNTG [11]. Proteins with this motif can be divided into six major categories: A, B, C, D, E, and F [12, 13]. Examples of proteins containing helix-loop-helix motifs are C-Myc, N-Myc, MyoD, Myf5, AHR, and ARNT.

15.2.3 Zinc Finger

Zinc finger proteins are one of the oldest and most abundant proteins found in both eukaryotes and prokaryotes. It was first identified in *Xenopus laevis* transcription factor IIIA [14, 15]. Zinc fingers have small compact structure with about 30 amino acids forming a ββα configuration [16]. The domain structure is stabilized by coordination of one or more zinc ions, which is a characteristic feature of these sort of proteins. There are about 30 types of zinc finger proteins classified by the HUGO Gene Nomenclature Committee determined by the zinc finger structure of the domain [17]. Cys2 His2 zinc finger proteins are the abundant forms in most species. The zinc ion coordinates the two histidines on the alpha helix with the two cystines on the β sheet. Zinc finger proteins bind to specific sites on the major groove of the DNA with the help of residues on the alpha helix [18, 19]. Examples of proteins containing zinc finger domains include Myelin transcription factor (Myt1), MYST family histone acetyltransferases, and Suppressor of tumourigenicity protein18 (ST18).

15.2.4 Leucine Zipper

Leucine zipper are found in both eukaryotic and prokaryotic proteins, but is a characteristic feature of the eukaryotes. It was first identified by Landschulz and collaborators [20]. Leucine zippers are 60–80-amino acid-long motifs consisting of 2 alpha helices which dimerize with the help of leucine residues present at a periodic interval of 7 residues in both the helices forming a coiled coil structure. The leucine being hydrophobic in nature interacts via hydrophobic interaction. At the end of the two alpha helices is the DNA binding region made up of basic residues like arginine and lysine which interacts with the DNA in the major groove via ionic interaction. The DNA binding region is a conserved region consisting of 16–25 basic amino acids commonly referred to as basic leucine zipper (bZIP). The overall domain structure gives the appearance of a Y-shaped zip which has been proved by many biochemical studies [21–23]. Examples of proteins containing leucine zipper motif include many transcription factors like Jun, Fos, GCN4, and HSF.

15.2.5 HMG-Box

HMG-box or high mobility group box domain may be present as single or multiple copies in HMG box factors or proteins which are responsible for DNA binding. They are usually found in eukaryotes and bind to the minor groove of DNA of non-B type DNA [24]. HMG-box proteins mediate either sequence-specific (transcription factors) or sequence non-specific binding (chromatin-remodeling complexes) [25]. HMG-box proteins with multiple domains generally bind non-specifically to DNA and act as chromatin factors, whereas HMG-box with single domains generally act as transcription factors which bind specifically to DNA. HMG-box motif is made up of three alpha helices connected by loops in an irregular fashion [25]. Examples of proteins containing HMG-box motifs include lymphoid enhancer binding factor 1 (LEF1), yeast transcription factors like Rox1, and Ixr1 [26, 27].

15.3 Detection Methods for Protein-DNA Interaction

There are several methods available for the detection of protein-DNA interaction. The widely used methods are discussed below.

15.3.1 DNA Electrophoretic Mobility Shift Assay (EMSA)

DNA electrophoretic mobility shift assay or EMSA is a very sensitive technique to analyze DNA–protein interaction. It is based on the principle that protein–DNA complexes have lesser mobility than free nucleic acids when run on a polyacrylamide or agarose gel. It is also known as gel shift or gel retardation assay due to the difference in migration pattern. The DNA oligonucleotides that are used for EMSA are generally radiolabelled with ^{32}P, which helps it to be detected by autoradiography after electrophoresis. But the probe can also be fluorescent tagged or biotin labelled for detection [28–30].

Advantages

- It is able to detect proteins directly from the crude lysate, which are present in limited amount making the assay highly sensitive.
- It can be used to identify the binding site in the upstream regulatory region of the gene by conducting mutations with the probe yielding a variety of configurations of the probe and testing it with the same lysate.

Limitations

- Difficult to quantitate.
- Detection of proteins from complexes might be difficult owing to weak interaction or rapid dissociation during electrophoresis.

15.3.2 DNA Pull-Down Assay

In this method, the interacting protein is detected by using a DNA probe with an affinity tag such as biotin. The DNA probe binds to the protein from the cell lysate which is then purified using agarose or magnetic beads. The bound protein is detected using either western blot or mass spectrometry [31]. Alternatively, the protein can also be affinity tagged and can be used for detecting any DNA–protein interaction with the protein-specific antibody. The unknown DNA sequence can be deciphered using PCR or southern blotting.

Advantages

- Detection of scanty targets.
- Isolation of entire DNA–protein complex is possible.

Limitations

- Assay must be conducted in vitro.
- Long DNA probes may result in non-specific binding.

15.3.3 DNA Footprinting

DNA Footprinting, which is a modification of Maxem-Gilbert sequencing technique, is one of the oldest methods for the detection of DNA–protein interaction. It was first developed by Galas and Schmitz [32]. It is based on the principle that proteins bound to DNA protect the DNA from getting cleaved by enzymes.

The DNA fragment of interest is first PCR amplified with ^{32}P 5′ labelled primer producing DNA fragments radiolabeled on one end. The DNA is mixed with DNA binding protein and then subjected to cleavage by DNase and the fragments are separated by running a PAGE. The gel is then visualized by autoradiography. A comparative analysis is done by running both the samples with and without the DNA binding protein. The sample without the protein should produce a uniform ladder of bands, whereas the sample with the DNA binding protein should have gaps in between the bands. These gaps indicate the DNA sequence where the protein is bound [33].

Advantages

- Can identify DNA–protein interaction from crude lysates.

Limitations

- Protein titrations need to be done to fully saturate the DNA probe with DNAse titrations also for proper cutting of DNA.
- Single nucleotide differences is difficult to be figured out by this assay.

15.3.4 Chromatin Immunoprecipitation (CHIP) Assays

Chromatin Immunoprecipitation or the CHIP assays are currently one of the widely used assays to identify DNA–protein interactions. It helps to identify the interactions that takes place in the living cell. CHIP requires the prior knowledge of the target protein and DNA sequence that is to be analyzed. The process involves first fixing the cells with formaldehyde, which is a reversible crosslinker that helps to crosslink the chromatin with the associated proteins. The cells are then lysed and the chromatin is sheared using sonication or enzymatic cleavage. Immunoprecipitation is then carried out with protein-specific antibody to precipitate the protein associated with the chromatin. The crosslink is then reversed and the DNA can be analyzed using various enrichment procedures like qPCR with primers flanking the gene of interest [34].

Advantages

- In vivo interaction can be studied through CHIP without any modifications by fixing the cell at the current instant.

Limitations

- Difficult to be performed on a large scale.
- Antibodies should be highly specific in nature.

15.3.5 Reporter Assay

Reporter assays also help to find out DNA–protein interactions in a living cell. It gives a real-time in vivo read out of the translational activity of the promoter of interest. Reporter genes are fusions of target promoter gene sequence and a reporter gene DNA sequence. The reporter gene codes for enzymes like firefly luciferase or alkaline phosphatase which catalyzes a substrate to produce either chemiluminescence or a colorimetric change. The enzyme is produced when the promoter is activated or it is bound by a DNA binding protein. The colorimetry or chemiluminescence gives an idea about the DNA–protein interaction [35, 36].

Advantages

- The assay can be performed in vivo.
- Mutational studies of promoter binding can be carried out with ease.

Limitations

- Exogenous DNA is used for study.

15.3.6 Biophysical Assays

Biophysical methods like X-ray crystallography and NMR provide gainful insights on DNA–protein interaction. X-ray studies help us to understand the DNA–protein interaction at the atomic level. It identifies the amino acids required for the DNA–protein interaction. The major drawback of X-ray crystallography is it is very hard to crystallize the DNA–protein complex structure [37–39].

Advantages

- Interaction can be studied at the atomic level.

Limitations

- Proteins are very hard to crystallize.

15.4 Conclusion

DNA–protein interaction is very important for the cell to conduct day-to-day activities. A small mistake in the recognition process can lead to huge disorders in the cell resulting in several diseases. Though several studies have been conducted to understand the DNA–protein interaction inside the cell, it only provides a vague idea on how the DNA–protein interaction occurs, as every process has one or the other limitation. This study has highlighted the limitations of the different assay methods so that further studies can be conducted to help in better understanding the process in the future.

Conflict of Interest None

References

1. von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. PNAS. 1986;83 (6):1608–12.
2. Luscombe NM, Laskowski RA, Thornton JM. Amino acid–base interactions: a three-dimensional analysis of protein–DNA interactions at an atomic level. Nucleic Acids Res. 2001;29(13):2860–74.
3. Lin M, Guo J. New insights into protein–DNA binding specificity from hydrogen bond based comparative study. Nucleic Acids Res. 2019;47(21):11103–13.
4. Seeman NC, Rosenberg JM, Rich A. Sequence-specific recognition of double helical nucleic acids by proteins. PNAS. 1976;73(3):804–8.
5. Phillips DC. The three-dimensional structure of an enzyme molecule. Sci Am. 1966;215 (5):78–93.

6. Chiang Y-S, Gelfand TI, Kister AE, Gelfand IM. New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. Proteins. 2007;68(4):915–21.
7. Brennan RG, Matthews BW. The helix-turn-helix DNA binding motif. J Biol Chem. 1989;264 (4):1903–6.
8. Anderson WF. The helix-turn-helix motif and the cro repressor. In: Taylor WR, editor. Patterns in protein sequence and structure, Springer series in biophysics. Berlin: Springer; 1992. p. 85–95.
9. Wintjens R, Rooman M. Structural classification of HTH DNA-binding domains and protein–DNA interaction modes. J Mol Biol. 1996;262(2):294–313.
10. Murre C, McCaw PS, Baltimore D. A new DNA binding and dimerization motif in immunoglobulin enhancer binding, daughterless, MyoD, and myc proteins. Cell. 1989;56(5):777–83.
11. Ellenberger T, Fass D, Arnaud M, Harrison SC. Crystal structure of transcription factor E47: E-box recognition by a basic region helix-loop-helix dimer. Genes Dev. 1994;8(8):970–80.
12. Atchley WR, Fitch WM. A natural classification of the basic helix–loop–helix class of transcription factors. PNAS. 1997;94(10):5172–6.
13. Ledent V, Paquet O, Vervoort M. Phylogenetic analysis of the human basic helix-loop-helix proteins. Genome Biol. 2002;3(6):research0030.1.
14. Brown RS, Sander C, Argos P. The primary structure of transcription factor TFIIIA has 12 consecutive repeats. FEBS Lett. 1985;186(2):271–4.
15. Miller J, McLachlan AD, Klug A. Repetitive zinc-binding domains in the protein transcription factor IIIA from Xenopus oocytes. EMBO J. 1985;4(6):1609–14.
16. Zhang W, Xu C, Bian C, Tempel W, Crombet L, MacKenzie F, et al. Crystal structure of the Cys2His2-type zinc finger domain of human DPF2. Biochem Biophys Res Commun. 2011;413 (1):58–61.
17. Gray KA, Yates B, Seal RL, Wright MW, Bruford EA. Genenames.org: the HGNC resources in 2015. Nucleic Acids Res. 2015;43(D1):D1079–85.
18. Elrod-Erickson M, Rould MA, Nekludova L, Pabo CO. Zif268 protein–DNA complex refined at 1.6å: a model system for understanding zinc finger–DNA interactions. Structure. 1996;4 (10):1171–80.
19. Pavletich NP, Pabo CO. Zinc finger-DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 A. Science. 1991;252(5007):809–17.
20. Landschulz WH, Johnson PF, McKnight SL. The leucine zipper: a hypothetical structure common to a new class of DNA binding proteins. Science. 1988;240(4860):1759–64.
21. Ellenberger TE, Brandl CJ, Struhl K, Harrison SC. The GCN4 basic region leucine zipper binds DNA as a dimer of uninterrupted α helices: crystal structure of the protein-DNA complex. Cell. 1992;71(7):1223–37.
22. Glover JNM, Harrison SC. Crystal structure of the heterodimeric bZIP transcription factor c-Fos–c-Jun bound to DNA. Nature. 1995;373(6511):257–61.
23. Fujii Y, Shimizu T, Toda T, Yanagida M, Hakoshima T. Structural basis for the diversity of DNA recognition by bZIP transcription factors. Nat Struct Biol. 2000;7(10):889–93.
24. Štros M, Launholt D, Grasser KD. The HMG-box: a versatile protein domain occurring in a wide variety of DNA-binding proteins. Cell Mol Life Sci. 2007;64(19):2590.
25. Thomas JO. HMG1 and 2: architectural DNA-binding proteins. Biochem Soc Trans. 2001;29 (4):395–401.
26. Deckert J, Khalaf RA, Hwang S-M, Zitomer RS. Characterization of the DNA binding and bending HMG domain of the yeast hypoxic repressor Rox1. Nucleic Acids Res. 1999;27 (17):3518–26.
27. Vizoso-Vázquez A, Lamas-Maceiras M, Fernández-Leiro R, Rico-Díaz A, Becerra M, Cerdán ME. Dual function of Ixr1 in transcriptional regulation and recognition of cisplatin-DNA adducts is caused by differential binding through its two HMG-boxes. Biochim Biophys Acta. 2017;1860(2):256–69.

28. Garner MM, Revzin A. The use of gel electrophoresis to detect and study nucleic acid–protein interactions. Trends Biochem Sci. 1986;11(10):395–6.
29. Carey J. Gel retardation. Methods Enzymol. 1991;208:103–17.
30. Hellman LM, Fried MG. Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions. Nat Protoc. 2007;2(8):1849–61.
31. Murarka P, Srivastava P. An improved method for the isolation and identification of unknown proteins that bind to known DNA sequences by affinity capture and mass spectrometry. PLoS One. 2018;13(8):e0202602.
32. Galas DJ, Schmitz A. DNAase footprinting a simple method for the detection of protein-DNA binding specificity. Nucleic Acids Res. 1978;5(9):3157–70.
33. Brenowitz M, Senear DF, Kingston RE. DNase I footprint analysis of protein-DNA binding. Curr Protoc Mol Biol. 1989;7(1):12.4.1–12.4.16.
34. Gade P, Kalvakolanu DV. Chromatin immunoprecipitation assay as a tool for analyzing transcription factor activity. In: Vancura A, editor. Transcriptional regulation: methods and protocols [Internet]. Methods in molecular biology. New York, NY: Springer; 2012 [cited 2021 Apr 14]. p. 85–104. Available from: https://doi.org/10.1007/978-1-61779-376-9_6.
35. Solberg N, Krauss S. Luciferase assay to study the activity of a cloned promoter DNA fragment. In: Bina M, editor. Gene regulation: methods and protocols [Internet]. Methods in molecular biology. Totowa, NJ: Humana Press; 2013 [cited 2021 Apr 14]. p. 65–78. Available from: https://doi.org/10.1007/978-1-62703-284-1_6.
36. Jugder B-E, Welch J, Braidy N, Marquis CP. Construction and use of a Cupriavidus necator H16 soluble hydrogenase promoter (PSH) fusion to gfp (green fluorescent protein). PeerJ. 2016;4:e2269.
37. Smyth MS, Martin JHJ. X ray crystallography. Mol Pathol. 2000;53(1):8–14.
38. Pervushin K, Riek R, Wider G, Wüthrich K. Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. PNAS. 1997;94(23):12366–71.
39. Salzmann M, Pervushin K, Wider G, Senn H, Wüthrich K. TROSY in triple-resonance experiments: new perspectives for sequential NMR assignment of large proteins. PNAS. 1998;95(23):13585–90.

RNA–Protein Interaction Analysis

16

Sushil Kumar Rathore and Pallabi Pati

Abstract

RNA-binding proteins (RBPs) play very crucial role in various physiological and biochemical functions of cells. It is a well-known fact that proteins are the key component of cell through which various functions of cells are regulated as they are receptors, adaptors, and enzymes. However, their activities are also regulated by different types of RNA. There are different types of noncoding RNA in addition to coding RNAs like mRNA, tRNA, and rRNA. The noncoding RNA bind with protein in a very specific manner identifying specific sites, motifs, or its structure. The interactions of RNAs with proteins are the main cause of various biological functions of cells. In this chapter, various aspects of RNA–protein interactions have been discussed, like structures of RBPs, types of RPI, functions of RBP, and various approaches to understand RNA–protein interaction.

Keywords

RBP · Protein · RNA · RNA–protein interaction

Abbreviations

CSD	Cold shock domain
DDX	Dead box proteins
dsRNA	Double-stranded RNA

S. K. Rathore (✉)
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_16

FMRP	Fragile X mental retardation protein
KH	K-homology
LINE	Long interspersed element
mRNA	Messenger RNA
ncRNA	Noncoding RNA
PUM-HD	PUM-homology domains
RBP	RNA-binding proteins
RNP	Ribonucleoprotein
RPI	RNA–protein interaction
RRM	RNA recognition motif
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
XBP1	X-box-binding protein 1
YTH	YT521-B homology
ZKD	Zinc knuckle domain

16.1 Introduction

RNA molecules exhibit a wide range of form and function. RNAs have been categorized based on their coding ability into two major groups: protein coding messenger RNAs (mRNAs) and noncoding RNAs (ncRNA) [1]. To initiate protein synthesis, mRNA molecules act as scaffolds for additional details. ncRNAs are classified depending on the sequences, intracellular localizations, structures, and functions, as follows: rRNAs and tRNAs, that are core elements of the translation system; [2, 3] small nuclear RNAs (snRNAs) and small nucleolar RNAs (snoRNAs) are involved in splicing of RNA and its modification [4]. Further, developments in deep sequencing have demonstrated that at least 80% of mammalian genomes produce RNAs, and scores of new ncRNAs have been discovered in living organisms that play undefined roles [5, 6]. However, the underlying mechanisms of these roles have remained elusive. RNA-binding proteins (RBPs) play a significant role in the RNA life cycle like its synthesis, function, and turnover. During all three phases of the RNA life cycle, such roles are always accompanied by involvement with RNA-binding proteins, including synthesis, function, and turnover [7]. RBPs bind directly to RNA sequences and/or structures with its RNA-binding domains in order to make decisions about RNA fate and function.

Interactions between proteins and RNA are the basis of various functions like organization and protein complexes stabilization, mRNA processing and maturation for trafficking and silencing and stabilization of matured mRNA. RBP could recognize single-stranded RNA, double-stranded RNA, structural characteristics of folded RNAs, or may not interact RNA explicitly unlike DNA binding proteins that usually bind double-stranded DNA [8].

RNA–protein interactions (RPI) regulates essential biological processes such as DNA replication, transcription, tolerance to pathogens, viral replication, and gene expression regulation at the posttranscriptional level. Recent high-throughput research has indicated various cellular RNA-binding proteins and are recognizing and characterizing pairs of proteins and RNAs that are involved in RPIs. However, our knowledge regarding RNA-binding proteins is far less in comparison to regulatory DNA-binding proteins, like replication factors and transcription factors. Most computational studies have dealt with the problem of predicting the positions amino acid residues present in a protein that may bind to an RNA.

Till date, there are very limited studies that have focused on the issue of partner prediction, i.e., characterization of specific RNA for an already known RNA-binding protein or protein-binding partner(s) required for nontranslating RNAs. Although many studies like as RIP-Chip, RNA compete, PAR-CLIP, and HITS-CLIP may offer critical information on RNA–protein interaction, they are limited by their high cost and labor-intensive nature. Computational techniques are thus required to correctly predict RPIs and design networks of RNA–protein interaction. It would be particularly helpful to establish sequence-based approaches that can be employed to recognize potential RNA–protein partners without the need for any experimental interactions, because there are only a small number of known RNA protein complexes in the PDB [9].

16.2 About RNA-Binding Proteins: Structure, Diversity, and Evolution

The majority of RBPs are proteins with a globular RNA-binding domain that binds RNA, which modifies the fate or function of the bound RNA. Some assume that unique and high-affinity RBPs are more likely to possess biological functions. This popular conception of RBPs, though, assumes that they seek to modify the outcome or functionality of RNA. The RBPs are identified as “the mRNA’s clothes.” This makes sure that the 5′ and 3′ UTRs and the coding region are in separate states: one time hidden, the next time exposed, enabling the mRNA to pass through different life stages [7]. Ribonucleoprotein (RNP) complexes that are primarily involved in gene expression consist of a traditional RNA-binding protein (RBP). For RBP function, it utilizes well-defined RNA-binding domains such as the RNA recognition motif (RRM), KH domain, or DEAD-box helicase domain. Additionally, complex protein–RNA interactions can be found in various unconventional RBP types, such as those that employ RNA-binding domains [7]. Four main RNA–protein interactions have been proposed on the basis of fundamental features of RNAs like structure, sequence, modification, and target engagement, as well as the recognition mechanism of RBPs [10].

16.2.1 RNA–Protein Interactions Based on RNA Motif

RNA motifs are short sequences which regulate the fate of RNA and cellular processes. Interaction of RNA and proteins usually involves modular combination of one or more RBD like RNA recognition motifs (RRMs), hnRNP K-homology (KHs), PUM-homology domains (PUM-HD), and dead box proteins (DDXs). One of the best examples including this principle is the RRM domains of RBFOX2 that bind to a UGCAUG motif [11], while PUM2 take the help of PUM-HD to bind with UGUANAUA [11, 12].

The discovery of an increasing number of RBP-binding motifs has also exposed the intricacies of RNA–protein interaction that depends on RNA motifs. A single RBP possess a variety of binding motifs as in the case of LIN28 where, N-terminal cold shock domain (CSD) and the C-terminal zinc knuckle domain (ZKD) play a part in the binding of two different RNA motifs, namely the 'GGAG' motif and the '(U) GAU' motif [13]. In posttranscriptional regulations, LIN28 impedes the biosynthesis of let-7 miRNAs, regulating production and impacting various disease states [13]. Also, the "insulin-like growth factor 2 mRNA-binding protein 1" (IGF2BP1) is one of such RBPs that could bind several motifs. IGF2B P1 is more complex protein than LIN28 as it contains four domains of hnRNPK homology (KH) and two motifs for RNA recognition.

In addition to the number as well as sequence-specific RBP-binding motifs, RNA motif-based RNA–protein interactions occur along with the motif's flanking sequences. RNA motifs are especially well-suited for RBP-specific interactions, where RNA motif-dependent RNA–protein interactions often allow the use of motif contexts and other RBP-specific interactions.

16.2.2 RNA–Protein Interactions Based on RNA Structure

Typically, RBPs bind to small sequences of single-stranded RNA, but some RBPs perform their biological activities by interacting on the basis of their common structural characteristics with groups of RNAs, including secondary and tertiary structural characteristics [14]. RNA sequences could fold into various secondary structures, including long stems with bulges or hairpins through base pairing. After complementary base pairing, double-stranded RNA (dsRNA) can fold into various structures such as hairpins and long stems with bulges, known as classic secondary structures. dsRNA is essential in multiple biological functions, including transport of mRNA, editing of RNA, innate immune response, and RNA interference [15]. The detection and operation of RBPs are necessary for all of the process mentioned above. "Double-stranded RBPs" (dsRBPs) are the proteins that bind to dsRNAs and are characterized by the availability of minimum one "double-stranded RBD" (dsRBD).

The ADAR family, which includes dsRBPs of various sizes, all of which possess conserved modular domain organization carrying a catalytic domain at C-terminal, possesses various dsRBD [16, 17]. Though they usually focus sequences with fewer interruptions and under certain sequence constraints, ADAR proteins search out and process dsRNAs with any given sequence [18]. ADAR1/ADAR 2 bind to mRNA and miRNA precursors to promote adenosine to inosine conversion [19, 20]. Mostly conversion from adenosine to inosine occurs in noncoding sequences of mRNAs, like 5′ and 3′ UTRs and retrotransposon elements of introns, such as long interspersed elements (LINEs) and Alu elements. It is also important to point out that multiple biological changes can be caused by A-to-I editing, which can include the possibility to edit pre-mRNA splicing patterns, and thus create new isoforms [21] Many editing sites are present in miRNAs where some of the sites influence synthesis and function of miRNA [22].

dsRNAs and dsRBPs also mediate translation, mRNA, splicing, stability, and degradation of mRNA. STAU1 is a dsRBP that is localized to the rough endoplasmic reticulum. In order to analyze STAU1-bound RNA structures in human cells, researchers used hiCLIP technology to look for structures formed by STAU1 within these samples and found STAU1 to bind mainly to intramolecular RNA duplexes. An RNA duplex that spans 858 nucleotides in the X-box-binding protein 1 (XBP1) mRNA was discovered, which controls splicing and stability in cytoplasm [23]. Depending on their particular three-dimensional tertiary structures, multiple RNAs have important regulatory roles in diverse biological processes.

A helix internal loop helix motif is formed in the double-stranded region by the Kink-turn (K-turn) RNA structure which consists of a three-base loop surrounded by a noncanonical stem (NC-stem) and a canonical stem (C-stem) that starts with a tandem base pair of GA/AG [24, 25]. There are various RNA structures that include the K-turn motif, including box C/D snRNAs, snoRNAs, mRNAs, and rRNAs. Some K-turn motifs are different, but they have the same three-dimensional distinctive shape.

In addition to organized RBDs, there are amino acid sequences in proteins that are not self-structural and need an external molecule to attain secondary structure. These are termed intrinsically disordered regions (IDRs). IDRs may promote RNA–protein interactions [26]. While certain structural characteristics support specific interactions, the RGG/RG motifs of IDR bind RNA through weak multivalent interaction. The fragile X mental retardation protein (FMRP) binds with the secondary structure of the G4 RNA by utilizing the RGG/RG motifs present in an IDR [27]. The interplay of the G4 and FMRP IDR is important for the attachment of several mRNAs and regulates translation control and alternative splicing [28, 29]. Disordered sequences are observed in one-third of the RBPs, with many of these have missing canonical RBDs [30] demonstrating the major role of IDRs in the ability to bind RNA. The recent advancements of RNA structurome and RBDs with respect to variety, dynamics, and expansion have indicated that various facets of regulation in gene expression may be discovered from protein interactions with RNA structure-dependent RNA.

16.2.3 RNA–Protein Interactions Based on RNA Modification

There are approximately 160 RNA variations that have been discovered to date [31]. A new layer of RNA stability and functional control is provided by the use of nucleotide-base chemical modifications in RNA [32]. Researchers also discovered several RNA mutations associated with human disease, such as cancer and neurological disorders [33]. RNA and protein interaction also occurs by posttranscriptional modifications such as 5-methylcytosine (m5C) and N6-methyladenosine (m6A). There have been further m6A and m5C studies suggesting that these modifications are indispensable in various biological processes. M6A is the most prevalent and reversible RNA modification, which is involved in a number of RNA functions including mRNA polyadenylation, splicing, transport, translation, and degradation. M6A modification is a complex process, and after cellular stress, m6A levels go through a wide-ranging redistribution of the transcriptome. RNA–protein interaction is mediated by m6A methylation of RNA. The newly altered RNA following methylation of m6A, acts as a reactants for m6A-specific interactors, including m6A readers and erasers. YTHDC1–2 and YTHDF1–3 are well known m6A readers that include YT521-B homology (YTH) domain-containing proteins. All of these m6A readers recognize m6A via a non-motif-specific process. Typically, YTHDC2 and YTHDF1–3 are found in the cytoplasm. YTHDF1 is a cytoplasmic protein containing two domains, a C-terminal YTH-binding domain and an N-terminal domain that promotes recruitment of complex 3 (eIF3) translation initiation factor, all of which enables cap-independent translation. The terminal YTH-binding domain of YTHDF2 interacts with m6A mRNA and the CCR4-NOT deadenylase complex is recruited by its N-terminal domain enhances the deadenylation and degradation of mRNA modified at m6A.

Recent research indicates that m6A-modified mRNAs decay and translations are facilitated by YTHDF3. The nuclear reader YTHDC1 recruits and suppresses a pre-mRNA splicing factor called SRSF3. The nuclear reader YTHDC1 recruits the pre-mRNA splicing factor SRSF3, which inhibits accessibility of SRSF10 to m6A-altered mRNAs. This subsequently enables inclusion of exon in specific mRNAs and governs slicing of mRNA. YTHDC1 also communicates with SR SF3, CPSF6, and SRSF7 in the oocyte nucleus to regulate pre-mRNAs and affect fetal growth. M6A modification created on chromatin-associated RNAs are mediated by METTL3 and recognized by YTHDC1, facilitating degradation of these m6A-modified RNAs [10].

16.2.4 RNA Guide-Bas Ed RNA–Protein Interactions

There are various kinds of small noncoding RNAs, including snRNAs, piRNAs miRNAs, snoRNAs, crRNAs, and other ncRNAs, that help facilitate protein–RNA interactions. In addition to regulating diverse life processes, this mode of RNA–protein interaction helps to control disease growth. However, despite there being some continuity in these interactions, there is notable variation in the structures and roles that various ncRNAs use.

16.2.4.1 Role of miRNA in RNA–Protein Interaction

miRNAs are one of the small noncoding RNA molecules found in plants, animals, and viruses [34, 35]. Drosha, DGC R8, Dicer, and TRBP are some of the dsRBPs used in the biogenesis of miRNAs [34]. miRNA is inserted into the RISC and binds to the core sequences of the target mRNAs, thus inducing translational repression [36]. siRNAs are made in a way close to that of miRNAs [37]. Dicer cut the DsRNAs or hairpin RNAs into small fragments [38]. At this stage, the guide strand is anchored to AGO2 and other proteins, and RISC is synthesized, which takes mRNA substrates that have a complementary sequence and starts to degrade them [39].

Genes are regulated by miRNAs by base-pairing with mRNAs while preserving complementarity to the seed region of miRNA (2–8 nucleotides) [40, 41]. Two kinds of miRNA–mRNA interactions can be found: canonical and atypical.

miRNAs make base pairs fully with target mRNA during both atypical and canonical matching even if the seed region is located at the 5′ end of miRNA. The mRNA repression is different for these matching process. In one case, endonucleolytic cleavage is activated by key constituents of RISC-AGO2 when miRNAs have a significant complementary matching with the coding sequence or UTR of mRNA targets. While in other cases, proteins directed by miRNA can cause translation inhibition or deadenylation of mRNA, if mismatches between miRNAs and their targets are observed [42, 43].

16.2.4.2 RNA–Protein Interactions Guided by piRNA

A new category of small noncoding RNAs known as piRNAs has been discovered in the male gametes of animals [44]. piRNAs are 30 nucleotides long (26–31 nucleotides). Murine PIWI (MIWI), which includes Aub, AGO3, and piwi [45], are also linked with PiRNAs of the PIWI subfamily, and piRNA guides the PIWI proteins to play a critical role in the silencement of transcriptional and posttranscriptional transposons and to defend themselves against the regeneration of viral stem cells [45]. Almost every species relies on this mechanism to prevent transposons from being expressed in their genome of gametes. Additionally, piRNA-directed nuclear PIWI proteins associate with nascent transposon transcripts to produce heterochromatin by DNA or histone methylation, ultimately leading to transcriptional silencing [46, 47]. Mosquitoes mount an antiviral response based on piRNA whenever they are infected with positive sense ssRNA virus. Piwi5 and Ago3 are precursors of piRNAs, and the heterotypical ping pong system synthesizes piRNAs. Thus, as the number of piRNAs increases, RNA virus replication is suppressed, achieving the antiviral response target [48]. The entire process is supervised by piRNA. Additionally, piRNAs participate in the metabolic activities of PIWI and facilitate its degradation [49]. Recently published research indicates that piRNAs obtained from transposons and pseudogenes can degrade specific mRNAs as well as lncRNAs through interaction with PIWIL1L [50]. In addition, degradome sequencing [50] also provides a systematic method of analyzing RNA degradation patterns mediated by piRNA and has significantly expanded insight into the interaction of universal piRNA-guided RNA–protein.

16.2.4.3 RNA–Protein Interactions Based on SnoRNA Guide

SnoRNAs are a group of highly expressed ncRNAs present in archaeans and eukaryotes, mainly located inside nucleolus. They are derived from pre-mRNA introns having a size of 60–300 nt. SnoRNAs can be classified as box H/ACA or box C/D snoRNAs based on their conserved sequence. The motifs of box C (RUGAUGA) and D (CUGA) are combined with less conserved box C and box D motifs to form a stem-internal loop-stem structure. The folding of the box H/ACA snoRNAs results in a distinctive hairpin-hinge-hairpin-tail arrangement, with box H (ANANNA) situated amongst the two hairpins and ACA motifs near the 3′ end. A subclass of snoRNA named Cajal body-specific RNAs (scaRNAs) have been found extensively in Cajal bodies where both C/D box and a H/ACA box domain are present. snoRNA performs various types of functions which include guidance of chemical modification in rRNAs and snRNAs in sequence-specific manner. Box C/D snoRNAs mediate 20-O-methylation inside SNORD-ribonucleoprotein (RNP) complexes, 20-O-methylation in ribose present in snRNA, and rRNA is capable of affecting its production and function, which could have an effect in cellular processes and diseases [10].

SNORA-RNP complexes are created by a combination of box H/ACA snoRNA with DKC1, NHP2, GAR1, and NOP10 that catalyse the conversion of uridine to pseudouridine located at 15 nt upstream of boxH/ACA. Box H/ACA snoRNAs instruct rSNORA-RNPs to modify uridine residues on snRNAs necessary for RNA splicing as well as uridine residues on rRNAs. Apart from directing RNA modification, SnoRNA has been used to facilitate pre-rRNA and pre-mRNA alternative splicing processing [10].

16.2.4.4 Spliceosome Assembly and Function Using snRNA Guides

The spliceosome is constructed stepwise from components such as pre-mRNAs, proteins, and snRNAs. Specifically, snRNAs act as guides, leading each snRNP to its final destination. There are five distinct types of RNA–protein interactions relying on snRNA guide, as per the type of snRNA involved in RNA splicing: U1 snRNP:: 50-splicing site (50SS) interacting ions, U2 snRNP:: branch point sequence (BPS) interactions, At the 5′ and 3′ splice sites, U6 snRNP:: 50SS interactions, U6 snRNP:: U2 snRNP interactions, and U5 snRNP:: exonsequence interactions [10]. U1 snRNP, the first snRNP to bind to precursors of splicing, identifies mRNA precursors with high specificity through base pairing between 50SS and U1 snRNA bases 3–10. In eukaryotes, the interaction of pre-mRNAs and snRNPs led by U1 is extremely conserved and necessary for splicing. Recent studies, however, have identified U1 as a unique mutated gene in chronic lymphocytic leukemia, hepatocellular carcinoma (HCC), and hedgehog medulloblastoma. The first base of the U1 50SS recognition sequence contains significant mutations A > G and A > C, implying the splicing patterns of different cancer pathways [10]. After recognition by U1 snRNP, U2 snRNP binds to the BPS of a pre-mRNA through a base-pairing interaction between the U2 snRNA and BPS. The tri-snRNP U4/U 6.U5 then participates in spliceosome assembly and the substitution of U1 snRNPs. Finally, the U6 snRNP interacts via base pairing with the 5′ end of the intron and the U6 snRNA. Additionally, the U5 snRNA binds to the exon sequence at the 5′ and 3′ splice sites and is involved in trans-esterification reactions [10].

16.2.4.5 RNA Targeting by the Clustered Regularly Interspaced Short Palindromic Repeat (CRISPR)-Cas System Based on RNA

CRISPRs are bacteria and archaea-specific repetitive sequences that play a crucial role in prokaryotes' RNA-based adaptive immune systems. They were first used in research on DNA and genome editing. System of CRISPR/Cas9 and novel CRISPR/Cas have been developed to achieve accurate RNA targeting, restriction, monitoring, and editing in mammalian cells. As in case of CRISPR/SpyCas9 (*Streptococcus pyogenes* Cas9), specially engineered PAMmers can be used to direct Cas9 to selectively bind or cut RNA targets while avoiding matching sequences of DNA. Additionally, the integration of PAMmers and deactivated Cas9 (dCas9) allows monitoring of RNA in living cells without the use of genetically programmed tags, avoiding the use of microsatellite repeat RNA expansion sequences. Besides SpyCas9, some Cas9 homologs derived from other bacterial organisms, such as SauCas9, NmeCas9, and CjCas9, are capable of attaching and breaking intracellular RNA in a PAM-independent manner. Cas3a, Cas13b, and CasRx, all Class 2 type VI CRISPR-Cas effectors, are customizable singular RNA-targeting RNases directed by RNA.

Cas13a has been engineered to target and monitor endogenous RNAs in plant and mammalian cells. When compared to RNA interference, the CasRx ribonuclease effector derived from *Ruminococcus flavefaciens* XPD3002 exhibits high specificity and efficiency against a wide variety of endogenous transcripts. Its inactive type (d CasRx) can be used to modulate alternative splicing and relieve dysregulated tau isoform ratios in a neuronal model of frontotemporal dementia. REPAIR and RESCUE, both based on Cas13b, were also developed and used to modify RNA from A to I and C to U. CRISPR/Cas inspired RNA targeting system (CIRTS) is a new RNA engineering toolkit that was recently developed by researchers. It is composed of a tri-domain protein with a single-strand RNA-binding domain, a hairpin RNA-binding domain, and an effector domain, as well as a designed gRNA with a hairpin and a single strand. The discovery of the CRISPR-gRNA system provided new insights into ncR NA-mediated RNA–protein interactions. Along with protein engineering, the CRISPR-gRNA system has enormous potential for research and gene editing, especially for gene therapy [10].

16.3 Functional Roles of RBPs

16.3.1 mRNA Localization

Genes can be regulated by localization of mRNA to various subcellular locations [51, 52]. The efficiency and temporal resolution of protein synthesis is enhanced by mRNA trafficking, triggered by cellular signals. Additionally, it facilitates the synthesis of protein complexes by increasing the localized concentration of particular mRNA.

Localization of mRNAs involves three different mechanisms [53, 54]: (1) mRNA-directed transport, (2) local selective stabilization, and (3) local trapping. Different RBPs are required to recognize separate localized signals in the mRNAs. Signals of localization for active and direct transport usually seem as synergistic clustered secondary structure repeats [55–57], whereas some similar signals seem to be available in the primary sequence [58, 59]. Various localizing RBPs interact with the UTRs of localized mRNA separately with low specificity and affinity [60]. Multiple RBPs interacting cooperatively is considered important [61]. The effect of RBP-mediated defense on a single cellular position results in selective stabilization.

The well-studied example is Hsp83, whose deadenylation and degradation in Drosophila is controlled by the 3′ untranslated regions (3′ UTRs)-bound Smaug RBP with the exception of the posterior pole, where it localizes embryos. Diffusion and local trapping are used by the third mechanism. However, due to its moderate efficiency to limit mRNAs spatially, selective stabilization normally occurs, similar to the localization of Nanos mRNA at the posterior pole in Drosophila embryos [62].

16.3.2 Translation of mRNA

Regulation of translation can take place by changes to the translational machinery, or it can specifically target specific mRNAs. RBP-based modulation, an intriguing regulatory mechanism, enables mRNA-specific control of the basic translational machinery [63]. For example, mRNA-specific RBPs can obstruct the interaction between the mRNA and the ribosome 43S complex by physical blockage in a cap-dependent pathway [64] or arrest by 43S scanning in a cap-independent pathway, as observed in Drosophila msl-2 mRNA by SXL [65–67]. On the other hand, specific mRNAs are suppressed by global eIF4E structural adaptors, as seen in the case of Bruno and Smaug RBPs, which promote the blockage of Cup and Maskin eIF4E adaptors on nanos, oskar, and poly (A)-tailed mRNAs [68–70]. RBPs can also regulate translation at a later phase of initiation steps, by prohibiting the linking of ribosomal subunits [71], or after initiation stages, as demonstrated by the hnRNP E1 RBP, which inhibits ribosomal subunits [71] Dab2 and ILEI at the extension phase by attaching to the 3′ UTR [72, 73].

A group of RBPs recognize aberrant mRNAs as opposed to normal mRNAs in the translation-dependent quality control process, which is coupled with a degradation mechanism to turn on the translation machinery. Cytoplasmic polyadenylation [74] is another effective mechanism for regulating translation. RBPs are thought to serve as “place-markers” in the assembly of catalytic complexes on the poly (A) dynamic combinatorial code in several models.

16.3.2.1 Degradation of mRNA

In addition to RNA maturation, several different degradation mechanisms, RNA maturation, and regulated mRNA turnover are all involved in quality surveillance. RBP protects nuclear RNA quality by exporting and degrading abnormal RNA in the cytoplasm or adenylation through nuclear TRAMP and exosome-mediated 3′-5′ decay [75, 76].

Surveillance of cytoplasm is either achieved by “nonsense-mediated decay” (NMD) when aberrant stop codons are found or by “ribosome extension-mediated decay” (REMD) when translation extends beyond the stop codons. NMD, for example, includes the “exon-junction complex” (EJC), “poly-A binding protein 1” (PABPC1), and HRP1 to identify regulatory sites in mRNA decay substrates. RBPs may also function as adaptors, as evidenced by Upf1, which is involved in the development of the SURF complex and subsequent association with EJC [77]. Additional RBPs, such as Pub1 [78], the “APOBEC1–ACF editing complex” [79], and several 3′ UTR helicases or chaperones [80] provide selective control of decay performance. REMD decay recognizes the role of 3′ UTRs by designating the correct space between the terminating codon and the polyadenylation region [80]. Some key factors in quality surveillance mechanisms are frequently used in conditionally regulated degradation pathways that depend on mRNA-specific RBPs such as Staufen1 [81] and SLBP [82].

16.3.2.2 Editing of mRNA

RNA editing that occurs posttranscriptionally involves covalently altering RNA sequences by inserting adenosines or cytidines into uridines or inosines, respectively (C-to-U editing). Adenosines that readily localize to the double-stranded portion of viral RNAs, cellular pre-mRNAs, and noncoding RNAs are affected by adenosine-to-inosine (A-to-I) editing. Adenosine deaminase enzymes acting on the RNA (ADAR) family catalyze A-to-I editing. dsRNA-binding motifs (dsRBMs) are located in amino(N)-terminal ADAR regions, while ADAR portions of carboxy-terminal have a conserved domain with catalytic activity. ADARs can act on any double-stranded RNA sequence, but they prefer nucleotides that are close together. The 5′ nearest neighbor is the most powerful to bring about editing of adenosine in both ADAR1 and ADAR2. Since the catalytic domain is primarily responsible for nearest neighbor preferences, dsRBM helps human hADAR2 discern adenosines with a 3′ G. Also, the nucleotides outside of the nearest neighbor have an effect on ADAR preferences. Various factors like length of the dsRNA and presence of loops, bulges, and mismatches determine the number of adenosines to be edited [8].

Adenosine-to-inosine conversion has been suggested to take part in a number of processes, including regulation of neuronal signaling, formation of higher brain function, RNAi activity shaping, and regulation of microRNA synthetic pathway. Cytidine editing to uridine is carried out by the enzyme family of AID–APOBEC.

Following the discovery of cytidine editing to uridine in mRNA of apoB, detailed investigations into the possible target sites of APOBEC1 revealed that such editings are mostly restricted to 3′ UTRs. The proof of localization for editing sites at 3′ UTRs is a presence of cytidine surrounded on either side by uridine or adenosine and accompanied by a properly separated sequence motif (WCWN2-4WRAUYANUAU). Nonetheless, the consensus sequences of these motifs were not a target site when available in translating sequences, except ApoB. Editing of 3′ UTRs which is-mediated by 'APOBEC1' can affect posttranscriptional processes such as stability of transcripts, polyadenylation, subcellular localization, and translational output. The passing on of information of nucleotide sequence from DNA to RNA is a crucial operation, as shown by adenosine-to-inosine and cytidine-to-uridine editing. Along these lines, one study paper reported an unusual degree of changes in bases from DNA to RNA that cannot be explained by classical editing, and the reason behind the mechanisms are unknown [8].

16.3.2.3 Stability of a Specific mRNA Species

RBPs that interact with "adenine/uridine-rich elements" (AREs) are preferentially located within 3′ UTRs of mRNA, including TTP, AUF1, and Hu family members. The stability of a particular mRNA is determined by the interaction of many RBPs that both stabilize and destabilize it. The effect of RBP binding to be cooperative or antagonistic is affected by the spatial interaction and variance in affinity within the UTR between their regulatory sites. The effect of RBP binding is also influenced by the comparative quantity of such RBPs in the cellular condition and its confinement where the binding takes place. Furthermore, microRNAs and RBPs can also join together and their structural stability can be affected by RBPs and microRNAs [8].

16.3.2.4 Role in Diseases

Due to the fact that RBPs are engaged in almost all aspect of RNA metabolism, any mutation or disturbance of RBP function can result in a number of diseases. In cancer, overexpression of RBP or genetic variation can lead to inaccurate or extensive RNA binding at different phases of RNA metabolism, which can have a significant impact on cancer cells. During the development of the nervous system, gene expression is subject to strict dynamical regulation. RBPs involved in normal neuron growth and functioning were identified by Deschenes-Furry and colleagues. Lukon et al. have identified that many illnesses are caused by inhibition of function or overactivity of RBP. CGG triplet expansion on FMR1's 5′ UTR is linked to Fragile X syndrome, resulting in FMR1 function loss required for normal neuronal development. In autoimmune disorders like paraneoplastic neurologic syndromes (PNSs), RBPs like Nova proteins and Hu family are targeted by autoantibodies causing loss of function in RBP. The neuronal-specific Nova protein family mediates alternative splicing of their target pre-mRNAs present in the regions of CNS like the hindbrain and ventral spinal cord.

Numerous trinucleotide disorders are caused by defective RBPs. "Myotonic dystrophy type 1" (DM1) has several repetitions in the 3′ UTR region of the DMPK gene, whereas myotonic dystrophy type 2 (DM2) has significantly longer repetitions of the tetra-nucleotide CCTG, resulting in toxic mutant RNAs. A GCG repeat extension in the PABPN1 exon gene results in the development of a PABPN1 variant in oculo-pharyngeal muscular dystrophy (OPMD), a degenerative disease which starts during adulthood. After that, the mutant gene induces the continuation of its poly (A) tails to the size of a nascent mRNA. Transcripts with a lengthy poly (A) tail accumulate in the nuclei of skeletal muscle, resulting in the development of muscular dystrophy. ASF/SF2 and eIF4E are two additional cancer-related RBPs that have been studied. EIF4E is a particularly overexpressed oncogene in breast cancer that is correlated with a poor prognosis. Various cancers also overexpress ASF/SF2. ASF/SF2 overexpression has the potential to alter the splicing of important cell cycle regulators and tumor suppressor genes, making it an attractive target for cancer therapy. Mutations in the consumer regions of RNA operators, the master regulators of co-expressed genes, may result in the loss of one or more mRNA targets. Two SNPs in the FGF20 gene's 3′ UTR region have been linked to Parkinson's disease. Similarly, RBP function may be lost as a result of SNPs on mRNAs in miRNA genes or their target sites [8].

16.4 Investigative Methods for Interactions of RBP–RNA

This section describes the conceptual structure for experiments designed to classify RNA species bound by RBPs or, alternatively, subsets of RBPs bound to particular RNAs. This section is divided into four. In the first chapter, in vitro methods for studying protein–RNA interactions are discussed, as well as the basic concepts of these experimental protocols. In addition, newly developed techniques that complement in vivo approaches will be considered. The second section shows how to examine large in vivo transcriptomes, and the third section offers a few examples of structural approaches for studying protein–RNA interactions.

16.4.1 In Vitro Identification of RNA–Protein Interactions

In vitro methodologies usually use one of the two approaches to understanding interactions between RNA and RBP. An established RBP can be used as a starting point for identifying RNAs that interact with it. Traditional "electrophoretic mobility shift assays" (EMSA) or supershift assays are frequently used to illustrate that protein incubation in the presence or absence of an antibody specific for RBP disrupts RNA movement in PAGE. The second strategy entails finding any RBPs that are bound to the target RNA. To attach an antisense oligonucleotide to a matrix, affinity chromatography can be used. The oligonucleotide attaches to any RBPs or related proteins after the cell lysate flows through the matrix. One of in vitro methodologies' flaws is their inability to differentiate between physiologically important and nonphysiologically important interactions. Interactions between RNA and RBP must be measured in vivo in order to understand their biological significance [8].

16.4.1.1 Systematic Evolution of Ligands by Exponential Enrichment (SELEX)

SELEX had aided in our knowledge of the molecular mechanism by which proteins interact with RNA. To execute in vitro selection, a DNA pool containing a random and mutant sequence segment surrounded on both ends by a conserved sequence and maybe a promoter of T7 RNA polymerase is being chemically synthesized. Following many PCR cycles, the DNA is amplified and then in vitro transcribed to generate the RNA pool. According to their capacity to bind to a protein, RNAs are classified as binders or nonbinders. The RNAs are obtained, reverse transcribed, amplified by PCR, and transcribed again. With each round of filtering, the ratio of high- to low-affinity sequences increases until the pool is populated by the RNA species with highest-affinity. It is possible to detect sequences with a wide range of affinities when the sequence pool is at an intermediate stage of selection. Each sequence's relative concentration is proportional to its affinity, with a lower concentration suggesting a greater affinity [83].

16.4.1.2 RNA Compete

The RNA compete technique is used to determine the binding specificity of RBPs. This approach is based on an RNA library that contains all potential 8-base sequences identified a minimum of 12 times in unorganized RNAs, as well as all possible 6- and 7-nucleotide loop sequences (and about 60% of 8-base loops) within RNA hairpins of RNA with special 10-base pair stems. These sequences are utilized to generate ssDNA using a microarray, which is subsequently converted to dsDNA and amplified by polymerase chain reaction. Ultimately, an in vitro transcription step is used to create the ssRNA library from dsDNA. Thus, after the generation of the RNA library, a single drive of RNA target sequences employing a tagged RBP of interest is conducted. Then, RNA sequences selected by RBP are tagged and hybridized to a microarray of the same form as the RNA library. The richness of the specified RNAs from the start library is determined using computational analysis.

RNA compete provides a detailed estimation of RBP-binding tendencies to small RNAs spanning the entire k-mer range in both structured and nonstructured conformation. RNA can be employed to validate and assess in vivo approaches to understand protein–RNA interactions. Furthermore, positional weight matrices (PWMs) and consensus motifs are supported. In a broad sense, RNA compete includes the following three steps:: (1) the construction of an RNA pool from a collection of RNA sequences and structures; (2) a single pull-down of RNAs associated with a labeled RBP of interest; and (3) hybridization of the microarray and computational analysis of the proportional enrichment of the bound percentage with respect to the initial pool of RNAs [84].

16.4.2 In Vivo Identification of Protein–RNA Interactions

In vivo protein–RNA interaction methods may be used to characterize either the RBPs that bind to specific RNAs or the RNAs that bind to specific RBPs to and complement each other with a previously identified RBP. In the following segment, we will go through these two distinct but complementary approaches.

16.4.2.1 RIP-Chip

In this technique, immunoprecipitation is used to assay RNA–protein binding in vivo. The RIP-Chip employs antibodies to bind unique RBPs and enrich RNA fragments bound to these RBPs. When hybridized to a microarray, the associated RNA fragments are classified, allowing for genome-wide analysis of RNA–protein interactions. The RIP Chip has some drawbacks, including the likelihood of co-immunoprecipitation of additional RBPs alongside the RBP of interest. Furthermore, RBP–RNA associations sometimes fail to accurately reflect in vivo associations due to RBP and RNA re-association after cell lysis. Furthermore, RBP binding sites could not be identified within the specified RNA fragments with this technique. Hence motif analysis is also required to ascertain RNA binding preferences [85].

16.4.2.2 Cross-Linking and Immunoprecipitation (CLIP) and HITS-CLIP

Ultraviolet (UV) radiation CLIP enables the stringent in vivo purification of both RBPs and small RNA fragments that could be used for amplification and sequencing. UV-induced crosslinking of RBPs and RNAs is performed in vivo prior to protein purification in order to boost the performance of conventional immunoprecipitation methods. For example, photocrosslinking inhibits in vitro RNA–protein reassociation and co-immunoprecipitation. UV cross-linking helps in easy purification of protein complexes ensuring more stringent purification schemes to be employed. This results in high pure protein–RNA complexes and binding sites are identified by incomplete proteinase K digestion. In some cases, the reverse transcriptase (RT) that is used to prepare samples was shown to effectively transcribe via cross-linked regions. Cross-linked sites with reverse transcription errors may be used to precisely localize protein–RNA interface (such as by the iCLIP method).

"High-throughput Sequencing CLIP" (HITS-CLIP) is a technique that blends regular CLIP with HITS-CLIP. CLIP-based quantification of high-throughput sequencing of DNA (HTS/NGS) enhances the sensitivity, and RBP binding sites have a spatial resolution. CLIP suffers from HTS technique limitations, including high error rates in sequencing, uneven CLIP tag alignments, and also the description of acceptable context CLIP tag distributions for evaluating the statistical significance of RBP binding sites. Additionally, variations in CLIP analysis procedures might affect the RBP's assumed specificity. Some RNAases are employed to degrade unbound RNA, unattached RNA exhibit sequence selectivity, which may have an effect on CLIP-tagged RBP-binding sites [86]. Additionally, although the CLIP cross-linking protocol is more sensitive, it may have a lower specificity [87].

16.4.2.3 Photo-Activatable Ribonucleoside-Enhanced Cross-Linking and Immunoprecipitation (PAR-CLIP)

The PAR-CLIP method is a variation of the cross-linking and immune precipitation technique in which photo-activated nucleosides are applied to the medium, followed by cell absorption and protein–RNA crosslinking. This improvement has a number of advantages over conventional CLIP. To begin, PAR-CLIP recovers 100–1000 times more cross-linked RNA when intensities of radiation are equivalent. The second advantage is that UV radiation induces T-to-C mutations, which are typical in cross-linked nucleoside analog-containing sites. PAR-CLIP leverages mutation analysis to enhance the detection of RBP attachment site locations or footprints [8].

16.4.2.4 Individual-Nucleotide Resolution Ultraviolet Cross-Linking and Immunoprecipitation (iCLIP)

Although all other CLIP techniques operate in the same way, iCLIP is a version that focuses on the RNA–protein interaction detection during sample preparation and the formation of crosslinking sites. iCLIP accomplishes this by taking advantage of reverse transcription's natural tendency to terminate before cross-bound nucleotides owing to the remaining amino acids. After circularization and linearization, the circularized and linearized cDNAs are PCR-amplified and then HTS-analyzed. The location may be used in place of the adaptor sequence utilized in the circularized PCR amplification to identify the RBP-binding site [8].

16.4.2.5 Finding the Proteins Bound to RNAs

Although studying protein components of RNA protein complexes in vivo can be challenging, some strategies have been developed. This problem is addressed by integrating and improving magnetic bead-based assays and crosslinking of protein-nucleic acid induced by UV radiation, as well as improving the PNA-assisted RBP identification method. The use of PNA oligonucleotides linked to peptides and peptide-PNA-linked oligonucleotides that can bind RNAs with greater specificity and selectivity than complementary RNA or DNA, as well as targeting of oligonucleotides to living cells efficiently, are among the method's unique features. PNAs hybridize with their RNA cognates once within the cell, and UV light is used to crosslink the targeted RNAs. After magnetic beads have been used to separate the RBP–PNA complexes, they are combined with an antisense PNA oligo and characterized using mass spectrometry techniques. Many protein–RNA complexes discovered by protein capture methods are severely misidentified, according to researchers. In contrast, quantitative mass spectrometry [88] aids in the differentiation of proteins particularly bound to the RNA of choice from other compounds with similar binding affinity.

16.4.2.6 Protein–RNA Interactions: Structural Analysis

CLAMP (crosslinking and mapping the protein domain) allows the mapping of RNA-binding domains that are cross-linked to unique nucleotides in the RNA within RBPs. This method is particularly useful when dealing with RNA-binding domains and protein–RNA interactions. The chromophore must be inserted into the site, photochemical protein–RNA crosslinking must be added, and a site-specific chemical protein cleavage is required for CLAMP to function.

16.4.2.7 Online Resources for Experimental Protein–RNA Interactions

Only a few resources were utilized to record the data provided by the given technologies about protein–RNA interactions. The RNA-binding protein database can be found at http://rbpdb.com, while the CLIPZ database can be found at http://www.clipz.unibas.ch [89]. RBPDB might be a good place to start if someone wants to learn more about manually collected RNA-binding interactions and/or regions for a particular RNA-binding protein The RBPDB contains experimental associations identified in vitro (e.g., RNA compete) or in vivo (e.g., RIP-Chip, CLIP) (human, mouse, fly and worm). RBPDB extends the capabilities of searching for motifs in an input RNA sequence by adding the ability to retrieve probable binding sites annotated by PWM ratings CLIPZ, in comparison to RBPDB, seems to be a more structured database of RNA-binding sites developed by the HITS–CLIP approach that enables display and study of the data collected using this approach. Using motif enrichment review, RBP binding sequence motifs can be predicted. The statistical significance of putative binding site motifs is also restored. Other methods are also capable of assessing spatial relationships between RBPs.

http://pridb.gdcb.astate.edu/index.php is a database including interactions between proteins and RNA. The Protein Data Bank (PDB) has a database of complex-derived protein–RNA interactions. It makes it easier to find and visualization of covalently linked amino acids and ribonucleotides in the primary sequences of the proteins and RNA chains involved. PRIDB uses both a distance-based criterion and the ENTANGLE algorithm to characterize interfaces [90]. Additionally, PRIDB searches for ProSite [91] and FR3D [92] motifs, respectively.

The Atlas of UTR Regulatory Behavior (AURA) is a manually compiled Catalog of Human UTRs and UTR Regulatory Annotations that can be found at http://aura.science.unit.it (AURA). A simple, interactive online interface gives complete access to a vast amount of data on UTRs, including information on phylogenetic preservation, RNA sequence and structure data, single nucleotide variation, gene expression, and functional descriptions of genes. It has also taken into account interactions between RBPs and miRs that have been experimentally determined to be nonredundant, as well as their effects on human UTRs [93].

16.5 Computational Inference of RBP-Binding Sites

There are a variety of analytical methods for identifying RNA sequence elements that operate as RBP binding sites. These techniques will be explored briefly in this section.

16.5.1 Binding Site Search

PWMs are often used to summarize the statistical features of observed binding sites. PWMs denote the odds of each nucleic acid occurring at each position. PWMs are used to scan RNA sequences for potential RBP binding sites. This search can be carried out using regulatory sequence analysis methods like RSAT (http://rsat.ulb.ac.be/rsat/). The accuracy of this RNA-binding specificity representation is on the basis of a large fraction of experimental data.

16.5.2 Models of Binding Sites

When introducing the most up-to-date techniques for modeling RBP attachment sites, models of transcription factor attachment site provide valuable guidance for the solution of pattern prediction and discovery. New techniques or modifications of existing techniques are used to model the binding elements of RBPs. Due to their distinctiveness and commonalities, several strategies for identifying RBP binding sites are described here in comparison to discovery of DNA attachment site. As with transcription factors, RBP attachment sites are modeled using both unsupervised and supervised (regression) methods. There may be two models of RBP-binding sites: one that ignores RNA structure and another that does not, because RNA structure may affect binding. RBP attachment sites are distinct from binding sites of transcription factor in that they allow for the binding of RNA structure. Therefore, as a result, models can be classified into those that neglect the structure of the RNA and those that do give importance to the RNA structure. The methods that consider RNA structure can be divided into two groups: First model predict the structure of RNA and second model is about the structure of RNA in its structural context.

The unsupervised methods take collection of RNA sequences as inputs that are optimized for a given RBP's attachment sites (obtained, for example, via a SELEX procedure) and a standard model of usual composition of RNA sequence. Techniques of transcription factor techniques could be used directly with minimal adjustments (i.e., replacing Us with Ts) in case the impact of RNA structure is overlooked. For example, Multiple Expectation Maximization for Motif Elicitation (MEME) [94] maximizes the probability of the observed sequence set fitting a position-specific scoring matrix (PSSM) motif model using the expectation-maximization (EM) algorithm. Centered on the assumption of nucleotide independence, the PSSM model describes a product multinomial distribution over bound k-mers. It is interesting to note that MEME does not allow gaps in sequence pattern, which may present a problem when RNA-binding domains such as RRMs bind to randomly separated and very short RNA sequence. Another well-known example of a structure-naive approach used for RBP binding site modeling is the assignment of a conservation index to all possible k-mers (RNA words of length k) in order to perform an independent genome-wide search for k-mers retained in $3'$ UTRs [94]. These k-mers can serve as regulators.

MEMERIS is an upgraded version of the MEME algorithm that incorporates RNAfold-derived probabilities for base-pairing when fitting PSSM motifs [94]. The probabilities of base-pairing constrain the search space for an RBP-binding site's initial position. MEMERIS looks for a motif that is important to a specific sense of the RNA structure (i.e., unpaired regions); this technique is distinct from those that focus on sequence-specific structural elements (e.g., stem-loops).

Three types of motif-finding algorithms can be used to model RNA structure. The first group employs co-variation to arrive at a consensus structure for all aligned sequences following RNA sequence alignments. The efficacy of such approaches is largely dependent on the alignment accuracy, which requires a high level of homology between the input RNA sequences, which is an uncommon occurrence. The chances of such event become more common when searching for shared local patterns by multiple mRNAs attached by the very same RBP within long 3′ UTRs. Alternatively, methods such as RNAProfile [95] estimate the minimum free energy folds for every sequence before looking for particular folds. The primary issues here are accurately predicting folds and representing an entire set of folds using a single fold having the minimum free energy fold. The third method uses dynamic programming to match and fold two RNA sequences simultaneously, with the usual secondary structure anticipated utilizing energy-based factors, culminating in a structure-based alignment [96]. This pair-wise alignment is then extended using a variety of heuristics to multiple alignments. Since the secondary structure of an RNA sequence is frequently defined by algorithmic assumptions, the analysis of noisy inputs is essential. Probabilistic covariance models, such as CMfinder [97], are more effective at capturing observable difference in the sequence and structure of RNA patterns. RNApromo [98] was recently used to model co-regulated RBP sequence preferences across a range of RNA sequences.

RBP binding models are used in supervised approaches as part of regression models designed to forecast quantitative estimates of RBP binding, as well as RNA binding. Due to the difficulty of obtaining the required input data in the past, these approaches have been limited to RBP binding data. Earlier efforts in this field were either structure-naïve [99] or relied on simplistic stem-loop models [100]. Examples from more recent years include ATS [101] and RNAcontext [101]. In vitro assays, RNAcompete, and RNAcontext provide information on RBP binding affinity, and that information is used to learn the RNAcompete, for example, by setting a physical model to information of RBP attachment affinity and sequence of RBP. In vitro assays, RNAcompete, and RNAcontext provide information on RBP binding affinity, and that information is used to learn the RNA-protein interaction. RNAcontext is fascinating for two reasons: it is capable of modeling RBP preferences for sequences based on their structural contexts, and it makes extensive use of high-throughput quantitative data to evaluate different parameters of model. RNAcontext operates in three steps, beginning with the input of a series of sequences and their corresponding affinity measurements. The first step calculates the probability that a word of length k contains an RBP binding site using the product of two terms. The first term denotes the inferred RBP sequence's priorities (in the form of a positional weight matrix), while the second term denotes the relative structural priorities of RBPs in different structural contexts. The second step is to estimate a sequence affinity based on the affinities assigned to each phrase by the previous motif model. The third step is to determine which array of parameters reduces the amount of squared differences between measured and expected input affinities when the sequence score function is modelled as a linear function. ATS is comparable to RNAcontext, except that it employs a selfish search strategy and considers only one structural background at a time while attempting to locate a degenerate consensus sequence motif. ATS, on the other hand, is a better fit for in vivo binding assays than RNAcontext, as the former's sequence scoring function is optimal for the longer RNA sequences associated with these assays.

16.6 Conclusion and Future Perspectives

With the introduction of efficient high-throughput technique capable of analyzing whole transcriptome and proteome, it is estimated that the number of RBP and types of its interaction with RNA is more than expected. The integration of structural data, defining site of molecular contacts, and high-throughput sequencing method that unfold RNA sequence specificity could allow for the determination of predictive model for specific RBPs. The growing experimental data of transcriptome should facilitate development in computational methods for prediction of RNA–protein interaction and for modeling regulatory pathway of RPI. According to genome-wide study, SNPs found in the RBP-binding region were associated with diseases. Disease susceptibility is influenced by genetic variation in RPI and interference with normal function. Further investigation on association of genetic variation and investigation will give better understanding of RPI.

Conflicts of Interest None

References

1. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014;15:423–37.
2. Khade P, Joseph S. Functional interactions by transfer RNAs in the ribosome. FEBS Lett. 2010;584:420–6.
3. Sauert M, Temmel H, Moll I. Heterogeneity of the translational machinery: variations on a common theme. Biochimie. 2015;114:39–47.
4. Esteller M. Non-coding RNAs in human disease. Nat Rev Genet. 2011;12:861–74.
5. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, et al. Landscape of transcription in human cells. Nature. 2012;489:101–8.
6. Uszczynska-Ratajczak B, Lagarde J, Frankish A, Guigó R, Johnson R. Towards a complete map of the human long non-coding RNA transcriptome. Nat Rev Genet. 2018;19:535–48.
7. Hentze MW, Castello A, Schwarzl T, Preiss T. A brave new world of RNA-binding proteins. Nat Rev Mol Cell Biol. 2018;19:327–41.
8. Re A, Joshi T, Kulberkyte E, Morris Q, Workman CT. RNA–protein interactions: an overview. In: Gorodkin J, Ruzzo WL, editors. RNA sequence, structure, and function: computational and bioinformatic methods [Internet]. Totowa, NJ: Humana Press; 2014 [cited 2021 May 2]. p. 491–521. Available from: https://doi.org/10.1007/978-1-62703-709-9_23.
9. Muppirala UK, Honavar VG, Dobbs D. Predicting RNA-protein interactions using only sequence information. BMC Bioinform. 2011;12:489.
10. Liu S, Li B, Liang Q, Liu A, Qu L, Yang J. Classification and function of RNA–protein interactions. WIREs RNA. 2020;11:e1601.
11. Wang X, McLachlan J, Zamore PD, Hall TMT. Modular recognition of RNA by a human pumilio-homology domain. Cell. 2002;110:501–12.

12. Chao JA, Patskovsky Y, Patel V, Levy M, Almo SC, Singer RH. ZBP1 recognition of β-actin zipcode induces RNA looping. Genes Dev. 2010;24:148–58.
13. Ustianenko D, Chiu H-S, Treiber T, Weyn-Vanhentenryck SM, Treiber N, Meister G, et al. LIN28 selectively modulates a subclass of Let-7 microRNAs. Mol Cell. 2018;71:271–283.e5.
14. Zhang M, Perelson AS, Tung C-S. RNA structural motifs. eLS [Internet]. American Cancer Society; 2011 [cited 2021 May 2]. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0003132.pub2.
15. Poynter SJ, DeWitte-Orr SJ. Understanding viral dsRNA-mediated innate immune responses at the cellular level using a rainbow trout model. Front Immunol [Internet]. Frontiers; 2018 [cited 2021 May 2]; 9. Available from: https://www.frontiersin.org/articles/10.3389/fimmu.2018.00829/full.
16. Eisenberg E, Levanon EY. A-to-I RNA editing—immune protector and transcriptome diversifier. Nat Rev Genet. 2018;19:473–90.
17. Nishikura K. Functions and regulation of RNA editing by ADAR deaminases. Annu Rev Biochem. 2010;79:321–49.
18. Eggington JM, Greene T, Bass BL. Predicting sites of ADAR editing in double-stranded RNA. Nat Commun. 2011;2:319.
19. Luciano DJ, Mirsky H, Vendetti NJ, Maas S. RNA editing of a miRNA precursor. RNA. 2004;10:1174–7.
20. Palladino MJ, Keegan LP, O'Connell MA, Reenan RA. A-to-I pre-mRNA editing in Drosophila is primarily involved in adult nervous system function and integrity. Cell. 2000;102:437–49.
21. Lev-Maor G, Sorek R, Levanon EY, Paz N, Eisenberg E, Ast G. RNA-editing-mediated exon evolution. Genome Biol. 2007;8:R29.
22. Pinto Y, Buchumenski I, Levanon EY, Eisenberg E. Human cancer tissues exhibit reduced A-to-I editing of miRNAs coupled with elevated editing of their targets. Nucleic Acids Res. 2018;46:71–82.
23. Sugimoto Y, Vigilante A, Darbo E, Zirra A, Militti C, D'Ambrogio A, et al. hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1. Nature. 2015;519:491–4.
24. Klein DJ, Schmeing TM, Moore PB, Steitz TA. The kink-turn: a new RNA secondary structure motif. EMBO J. 2001;20:4214–21.
25. Wang J, Daldrop P, Huang L, Lilley DMJ. The k-junction motif in RNA structure. Nucleic Acids Res. 2014;42:5322–31.
26. Calabretta S, Richard S. Emerging roles of disordered sequences in RNA-binding proteins. Trends Biochem Sci. 2015;40:662–72.
27. Phan AT, Kuryavyi V, Darnell JC, Serganov A, Majumdar A, Ilin S, et al. Structure-function studies of FMRP RGG peptide recognition of an RNA duplex-quadruplex junction. Nat Struct Mol Biol. 2011;18:796–804.
28. Didiot M-C, Tian Z, Schaeffer C, Subramanian M, Mandel J-L, Moine H. The G-quartet containing FMRP binding site in FMR1 mRNA is a potent exonic splicing enhancer. Nucleic Acids Res. 2008;36:4902–12.
29. Schaeffer C, Bardoni B, Mandel J-L, Ehresmann B, Ehresmann C, Moine H. The fragile X mental retardation protein binds specifically to its mRNA via a purine quartet motif. EMBO J. 2001;20:4803–13.
30. Neelamraju Y, Hashemikhabir S, Janga SC. The human RBPome: from genes and proteins to human disease. J Proteome. 2015;127:61–70.
31. Boccaletto P, Machnicka MA, Purta E, Piatkowski P, Baginski B, Wirecki TK, et al. MODOMICS: a database of RNA modification pathways. 2017 update. Nucleic Acids Res. 2018;46:D303–7.
32. Song J, Yi C. Chemical modifications to RNA: a new layer of gene expression regulation. ACS Chem Biol. 2017;12:316–25.

33. Kadumuri RV, Janga SC. Epitranscriptomic code and its alterations in human disease. Trends Mol Med. 2018;24:886–903.
34. Bartel DP. MicroRNAs: genomics, biogenesis, mechanism, and function. Cell. 2004;116:281–97.
35. Krol J, Loedige I, Filipowicz W. The widespread regulation of microRNA biogenesis, function and decay. Nat Rev Genet. 2010;11:597–610.
36. Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. Nat Rev Drug Discov. 2013;12:847–65.
37. Carthew RW, Sontheimer EJ. Origins and mechanisms of miRNAs and siRNAs. Cell. 2009;136:642–55.
38. Ramaswamy G, Slack FJ. siRNA: a guide for RNA silencing. Chem Biol. 2002;9:1053–5.
39. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 2001;409:363–6.
40. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120:15–20.
41. Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003;115:787–98.
42. Filipowicz W, Bhattacharyya SN, Sonenberg N. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat Rev Genet. 2008;9:102–14.
43. Oliveto S, Mancino M, Manfrini N, Biffo S. Role of microRNAs in translation regulation and cancer. World J Biol Chem. 2017;8:45–56.
44. Girard A, Sachidanandam R, Hannon GJ, Carmell MA. A germline-specific class of small RNAs binds mammalian Piwi proteins. Nature. 2006;442:199–202.
45. Ku H-Y, Lin H. PIWI proteins and their interactors in piRNA biogenesis, germline development and gene expression. Natl Sci Rev. 2014;1:205–18.
46. Aravin AA, Sachidanandam R, Bourc'his D, Schaefer C, Pezic D, Fejes Toth K, et al. A piRNA pathway primed by individual transposons is linked to de novo DNA methylation in mice. Mol Cell. 2008;31:785–99.
47. Vodovar N, Bronkhorst AW, van Cleef KWR, Miesen P, Blanc H, van Rij RP, et al. Arbovirus-derived piRNAs exhibit a ping-pong signature in mosquito cells. PLoS One. 2012;7:e30861.
48. Miesen P, Girardi E, van Rij RP. Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. Nucleic Acids Res. 2015;43:6545–56.
49. Zhao S, Gou L-T, Zhang M, Zu L-D, Hua M-M, Hua Y, et al. piRNA-triggered MIWI ubiquitination and removal by APC/C in late spermatogenesis. Dev Cell. 2013;24:13–25.
50. German MA, Luo S, Schroth G, Meyers BC, Green PJ. Construction of parallel analysis of RNA ends (PARE) libraries for the study of cleaved miRNA targets and the RNA degradome. Nat Protoc. 2009;4:356–62.
51. Holt CE, Bullock SL. Subcellular mRNA localization in animal cells and why it matters. Science. 2009;326:1212–6.
52. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, Cerovina T, et al. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. Cell. 2007;131:174–87.
53. Wolke U, Weidinger G, Köprunner M, Raz E. Multiple levels of posttranscriptional control lead to germ line-specific gene expression in the zebrafish. Curr Biol. 2002;12:289–94.
54. Lipshitz HD, Smibert CA. Mechanisms of RNA localization and translational regulation. Curr Opin Genet Dev. 2000;10:476–88.
55. Chartrand P, Meng XH, Huttelmaier S, Donato D, Singer RH. Asymmetric sorting of Ash1p in yeast results from inhibition of translation by localization elements in the mRNA. Mol Cell. 2002;10:1319–30.

56. Lewis RA, Kress TL, Cote CA, Gautreau D, Rokop ME, Mowry KL. Conserved and clustered RNA recognition sequences are a critical feature of signals directing RNA localization in Xenopus oocytes. Mech Dev. 2004;121:101–9.
57. Macdonald PM, Struhl G. Cis-acting sequences responsible for anterior localization of bicoid mRNA in Drosophila embryos. Nature. 1988;336:595–8.
58. Cenik C, Chua HN, Zhang H, Tarnawsky SP, Akef A, Derti A, et al. Genome analysis reveals interplay between 5′UTR introns and nuclear mRNA export for secretory and mitochondrial genes. PLoS Genet. 2011;7:e1001366.
59. Palazzo AF, Springer M, Shibata Y, Lee C-S, Dias AP, Rapoport TA. The signal sequence coding region promotes nuclear export of mRNA. PLoS Biol. 2007;5:e322.
60. Arn EA, Cha BJ, Theurkauf WE, Macdonald PM. Recognition of a bicoid mRNA localization signal by a protein complex containing swallow, nod, and RNA binding proteins. Dev Cell. 2003;4:41–51.
61. Müller M, Heym RG, Mayer A, Kramer K, Schmid M, Cramer P, et al. A cytoplasmic complex mediates specific mRNA recognition and localization in yeast. PLoS Biol. 2011;9:e1000611.
62. Forrest KM, Gavis ER. Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for nanos mRNA localization in Drosophila. Curr Biol. 2003;13:1159–68.
63. Dever TE. Gene-specific regulation by general translation factors. Cell. 2002;108:545–56.
64. Muckenthaler M, Gray NK, Hentze MW. IRP-1 binding to ferritin mRNA prevents the recruitment of the small ribosomal subunit by the cap-binding complex eIF4F. Mol Cell. 1998;2:383–8.
65. Gebauer F, Grskovic M, Hentze MW. Drosophila sex-lethal inhibits the stable association of the 40S ribosomal subunit with msl-2 mRNA. Mol Cell. 2003;11:1397–404.
66. Grskovic M, Hentze MW, Gebauer F. A co-repressor assembly nucleated by Sex-lethal in the 3′UTR mediates translational control of Drosophila msl-2 mRNA. EMBO J. 2003;22:5571–81.
67. Beckmann K, Grskovic M, Gebauer F, Hentze MW. A dual inhibitory mechanism restricts msl-2 mRNA translation for dosage compensation in Drosophila. Cell. 2005;122:529–40.
68. Nelson MR, Leidal AM, Smibert CA. Drosophila cup is an eIF4E-binding protein that functions in Smaug-mediated translational repression. EMBO J. 2004;23:150–9.
69. Nakamura A, Sato K, Hanyu-Nakamura K. Drosophila cup is an eIF4E binding protein that associates with Bruno and regulates oskar mRNA translation in oogenesis. Dev Cell. 2004;6:69–78.
70. Stebbins-Boaz B, Cao Q, de Moor CH, Mendez R, Richter JD. Maskin is a CPEB-associated factor that transiently interacts with elF-4E. Mol Cell. 1999;4:1017–27.
71. Ostareck DH, Ostareck-Lederer A, Shatsky IN, Hentze MW. Lipoxygenase mRNA silencing in erythroid differentiation: the 3′UTR regulatory complex controls 60S ribosomal subunit joining. Cell. 2001;104:281–90.
72. Chaudhury A, Hussey GS, Ray PS, Jin G, Fox PL, Howe PH. TGF-beta-mediated phosphorylation of hnRNP E1 induces EMT via transcript-selective translational induction of Dab2 and ILEI. Nat Cell Biol. 2010;12:286–93.
73. Hussey GS, Chaudhury A, Dawson AE, Lindner DJ, Knudsen CR, Wilce MCJ, et al. Identification of an mRNP complex regulating tumorigenesis at the translational elongation step. Mol Cell. 2011;41:419–31.
74. Villalba A, Coll O, Gebauer F. Cytoplasmic polyadenylation and translational control. Curr Opin Genet Dev. 2011;21:452–7.
75. Reed R, Hurt E. A conserved mRNA export machinery coupled to pre-mRNA splicing. Cell. 2002;108:523–31.
76. LaCava J, Houseley J, Saveanu C, Petfalski E, Thompson E, Jacquier A, et al. RNA degradation by the exosome is promoted by a nuclear polyadenylation complex. Cell. 2005;121:713–24.

77. Hwang J, Sato H, Tang Y, Matsuda D, Maquat LE. UPF1 association with the cap-binding protein, CBP80, promotes nonsense-mediated mRNA decay at two distinct steps. Mol Cell. 2010;39:396–409.
78. Ruiz-Echevarría MJ, Peltz SW. The RNA binding protein Pub1 modulates the stability of transcripts containing upstream open reading frames. Cell. 2000;101:741–51.
79. Chester A, Somasekaram A, Tzimina M, Jarmuz A, Gisbourne J, O'Keefe R, et al. The apolipoprotein B mRNA editing complex performs a multifunctional cycle and suppresses nonsense-mediated decay. EMBO J. 2003;22:3971–82.
80. Inada T, Aiba H. Translation of aberrant mRNAs lacking a termination codon or with a shortened 3′-UTR is repressed after initiation in yeast. EMBO J. 2005;24:1584–95.
81. Kim YK, Furic L, Desgroseillers L, Maquat LE. Mammalian Staufen1 recruits Upf1 to specific mRNA 3′UTRs so as to elicit mRNA decay. Cell. 2005;120:195–208.
82. Kaygun H, Marzluff WF. Translation termination is involved in histone mRNA degradation when DNA replication is inhibited. Mol Cell Biol. 2005;25:6879–88.
83. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990;249:505–10.
84. Ray D, Kazan H, Chan ET, Castillo LP, Chaudhry S, Talukder S, et al. Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol. 2009;27:667–70.
85. Keene JD, Komisarow JM, Friedersdorf MB. RIP-Chip: the isolation and identification of mRNAs, microRNAs and protein components of ribonucleoprotein complexes from cell extracts. Nat Protoc. 2006;1:302–7.
86. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M. A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. Nat Methods. 2011;8:559–64.
87. Corcoran DL, Georgiev S, Mukherjee N, Gottwein E, Skalsky RL, Keene JD, et al. PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data. Genome Biol. 2011;12:R79.
88. Butter F, Scheibe M, Mörl M, Mann M. Unbiased RNA-protein interaction screen by quantitative proteomics. Proc Natl Acad Sci U S A. 2009;106:10626–31.
89. Khorshid M, Rodak C, Zavolan M. CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. Nucleic Acids Res. 2011;39: D245–52.
90. Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol. 2001;311:75–86.
91. Sigrist CJA, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, et al. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res. 2010;38:D161–6.
92. Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB. FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol. 2008;56:215–52.
93. Paradis E, Claude J, Strimmer K. APE: analyses of phylogenetics and evolution in R language. Bioinformatics. 2004;20:289–90.
94. Hiller M, Pudimat R, Busch A, Backofen R. Using RNA secondary structures to guide sequence motif finding towards single-stranded regions. Nucleic Acids Res. 2006;34:e117.
95. Pavesi G, Mauri G, Stefani M, Pesole G. RNAProfile: an algorithm for finding conserved secondary structure motifs in unaligned RNA sequences. Nucleic Acids Res. 2004;32:3258–69.
96. Simultaneous SD. Solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math. 1985;45:810–25.
97. Yao Z, Weinberg Z, Ruzzo WL. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics. 2006;22:445–52.

98. Rabani M, Kertesz M, Segal E. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. Proc Natl Acad Sci U S A. 2008;105:14885–90.
99. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ. Profiling condition-specific, genome-wide regulation of mRNA stability in yeast. Proc Natl Acad Sci U S A. 2005;102:17675–80.
100. Foat BC, Stormo GD. Discovering structural cis-regulatory elements by modeling the behaviors of mRNAs. Mol Syst Biol. 2009;5:268.
101. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q. RNAcontext: a new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput Biol. 2010;6:e1000832.

SNP Identification and Discovery

17

Christian Bharathi Antony Raj, Hemavathy Nagarajan, Mohamed Hameed Aslam, and Santhiya Panchalingam

Abstract

The global climate change has a negative influence on the quality of crop production and has been considered a threat in recent years. Henceforth, there is a need for advancement in technology to overcome the issue and improve both the quality and quantity of the crop plants by exploiting their genome. The various single nucleotide polymorphism (SNP) markers have been extensively used in crop breedings as molecular markers with advances in NGS (Next-generation sequencing) technology. These SNP markers are cost-effective for variety identification. SNPs have a deterministic role in protein expression. Thereby sequencing and genotyping have enabled crop improvement based on genomics with significant advances in NGS technologies which have also assisted in overcoming the drawbacks in detecting new functional SNPs associated with diverse traits. While SNP markers are found to be highly abundant and prevalent across the genome, functional SNPs are known to have a crucial impact on the phenotypes of plants. Besides, it was also known that SNP markers can be widely used and implemented in genotyping for the identification of structural variants due to their codominance, low cost, flexibility, speed, and ease of automation than other markers. Even in the genome-wide association study (GWAS), SNP markers were considered a significant tool in developing genome-wide haplotypes. Identified SNP patterns from GWAS are useful for understanding plant evolution. Genetic variations derived from the SNP patterns may execute desired phenotypes in the plant that benefits plant breeding and crop improvements. Even though SNPs widely use, technical advancements are

C. B. Antony Raj (✉) · H. Nagarajan · M. H. Aslam
Centre for Bioinformatics, Pondicherry University, Pondicherry, India

S. Panchalingam
Department of Biotechnology, Karpagam Academy of Higher Education, Coimbatore, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_17

needed to overcome the challenges in plant SNPs identification, to understand speciation and evolution through genomic divergence of plants, and to identify associate genomic variations of phenotypic traits.

Keywords

SNP discovery · Genetic variations · Polyploids · Haplotypes · Nanopore sequencing

Abbreviations

ASO	Allele-specific oligonucleotide
OLA	Oligonucleotide ligation assay
QTLs	Quantitative trait loci
RAPD	Random amplification of polymorphic DNA
RFLP	Restriction fragment length polymorphism
SNP	Single nucleotide polymorphism
SSCP	Single-strand conformation polymorphism
SSRs	Simple sequence repeats

17.1 Introduction

SNP (Single nucleotide polymorphism) occurs as a variation at a single position of DNA, which may occur due to the difference of two or more individual organisms based on a nucleotide [1]. SNP can either cause harmful effects or may not be lethal with no prominent impact even on the phenotype of the species. However, SNP genotyping is essential to validate and confirm the specificity of the SNPs occurring at the position of an interested gene [2]. The SNPs in the coding region are divided into two types; synonymous mutation and non-synonymous mutation, which may affect the protein structure. As SNPs are widely dominant across the entire genome, the incidence of SNPs is not only restricted to coding regions of a gene, as they are also known to be prevalent even in the non-coding regions. The mystery prevails over the specificity and functionality of the SNPs at the non-coding regions, and hence biologists are still investigating to unravel the role of SNPs at the non-coding region.

The history of SNP reminds us the efforts and contribution of many scientists to the current understanding of SNP and its application in healthcare and other fields. Mendel's laws of genetics (in 1865 and before) were supported by T. H. Morgan, and the rediscovery of Mendel's work in the late 1900's had an indispensable contribution to the understanding of SNP. The first linear map of genes completed by Alfred H. Sturtevant, rendered genetic mapping using gene crossing-over (in between 1913 and 1928). Further, an advancement in linkage map and genetic distance analysis was implemented by Walther Flemming, William Bateson, and Thomas H. Morgan. The outcome of their studies provided the basis for development of the chromosome theory of inheritance. Besides, Colin MacLeod, Oswald T. Avery, and Maclyn McCarty together established the concept that DNA is the genetic material (in 1944) [3]. Later, scientists focused much on the role and function of specific genes to understand their biological importance. Sequencing methods and automated sequencers boomed genomic research, especially the Human Genome Project in 1990, which contributed to the sharing of gene and genomic sequences across the world. Completion of the Human Genome Project led to a breakthrough in genomic studies in 2003 [4]. However, the increased computational power and growing bioinformatics approaches along with genomic data provided a much better understanding of the mysteries in biology. The results from genomic analysis also paved various ways for clinical studies and drug discovery. Their implementation improved health care system through the emerging new biological concepts.

SNPs are most common in the genome with many types of variations and they are considered essential tools for the discovery of markers and genetic mapping; however, the application of SNPs in the plant polyploid genomes has been quite challenging [5, 6]. In plants, the phenotyping-based selection was considered to increase the productivity of plants that are resistant to pathogens and can withstand unpredictable climate changes. Yet, identification and analysis of the minor genetic variations which affect the phenotypes are demanding in plants [7]. However, genomics-assisted breeding is an advanced technology and it needs prior knowledge of the markers/loci/genes associated with traits of our interest, which is growing trends in plants [8]. Hence, sequencing and genotyping have enabled crop improvement based on genomics with significant advances in next-generation sequencing technologies [9, 10], which have also assisted in overcoming the drawbacks in detecting new functional SNPs associated with diverse traits. Apart from the reference genome assembly, various re-sequencings methods are also initiated to understand genetic diversity among the species concerning SNPs. Small InDels (insertions/deletions) act as markers for enhancing the productivity of various varieties using the genomics approach. Among the widely used molecular markers, namely, RFLP, RAPD, AFLP, and SSRs (Simple sequence repeats), intergenic SSRs and SNPs markers are highly preferred [11, 12]. In general, SNP markers are abundant and prevalent across the genome, and functional SNPs are known to have a crucial impact on the phenotype of plants. Besides, it is known that SNP markers can be widely used and implemented in genotyping for the identification of structural variants due to their codominance, low cost, flexibility, speed, and ease of automation than other markers [13–16]. The other key role played by SNPs is genetic mapping, where the genetic map aids in identifying the location of a gene of interest, arrangement of chromosome based on the distance among genes [17], and identification of quantitative trait loci (QTLs) [18]. SNPs have been analyzed and collected from crop plants such as cotton [19], rice [20], maize [21], and soybean [22] from genetic maps and other approaches. Even in the genome-wide association study (GWAS), SNP markers were considered as a significant tool in developing genome-wide haplotypes [23]. Further, the evolution of plants can be inferred from the identified SNPs patterns and genetic variations and their corresponding effect on phenotypes could be an idea for breeding and crop improvements. In connection with the above information, this chapter attempts to describe SNP identification, SNP databases, methods for SNP discovery, assays to determine SNPs, SNP applications and challenges involved in plant SNP discovery. The experimental techniques are described based on the principle and usage of technology for the identification of SNPs and their interpretation; it was catagoried into non-sequencing, sequencing, and resequencing methods (Fig. 17.1). The application of plant SNPs in plant genomic research, breeding, and crop improvements, and the challenges in plant SNP identification and discoveries are mentioned at the end of the chapter. This chapter provides a framework and guidelines to analyze a plant genome and pinpoints the current difficulties faced by the plant genomic research community to consider careful evaluation during analysis and gives suggestions to overcome the important challenges.

17.2 Experimental Techniques to Identify SNPs

17.2.1 Non-sequencing Methods

17.2.1.1 RFLP (Restriction Fragment Length Polymorphism)

RFLP can identify a single nucleotide difference in a gene that provides the cutting sites for a restriction enzyme. This restriction enzyme cutting site may present in one allelic form and may not be present in the other form. This kind of polymorphisms are detected by amplifying the gene of interest using prior primers targeting the polymorphic site; subsequently, the amplified interested genetic material is digested using the restriction enzyme, and the undigested amplified DNA fragments are analyzed by gel electrophoresis [24]. A highly efficient PCR-based assay is designed to detect RFLPs of genes and the nature of RFLPs could be understood with the sequence information after sequencing it. Besides, due to their high genomic abundance, high reproducibility, co-dominant inheritance, RFLP markers are highly considered for developing genetic mapping [25] of a wide variety of crop plants, namely rice [26], sunflower [27], maize [28], wheat [29], soybean [30], tomato and potato [31], barley [32], and chickpea [33]. Moreover, RFLP is also known to have crucial application investigating the diversity and phylogenetic studies within the population and also to their closely associated species [34], which may aid in understanding the hybridization and also gene flow within the crops [35]. However, RFLP cannot be used for plant breeding applications since it entails DNA of both high quantity and quality [36], along with the need for expensive radioactive isotopes. Furthermore, the process involved is highly tedious, time-consuming with labour intensive procedures and lack of automated procedures [8, 37].

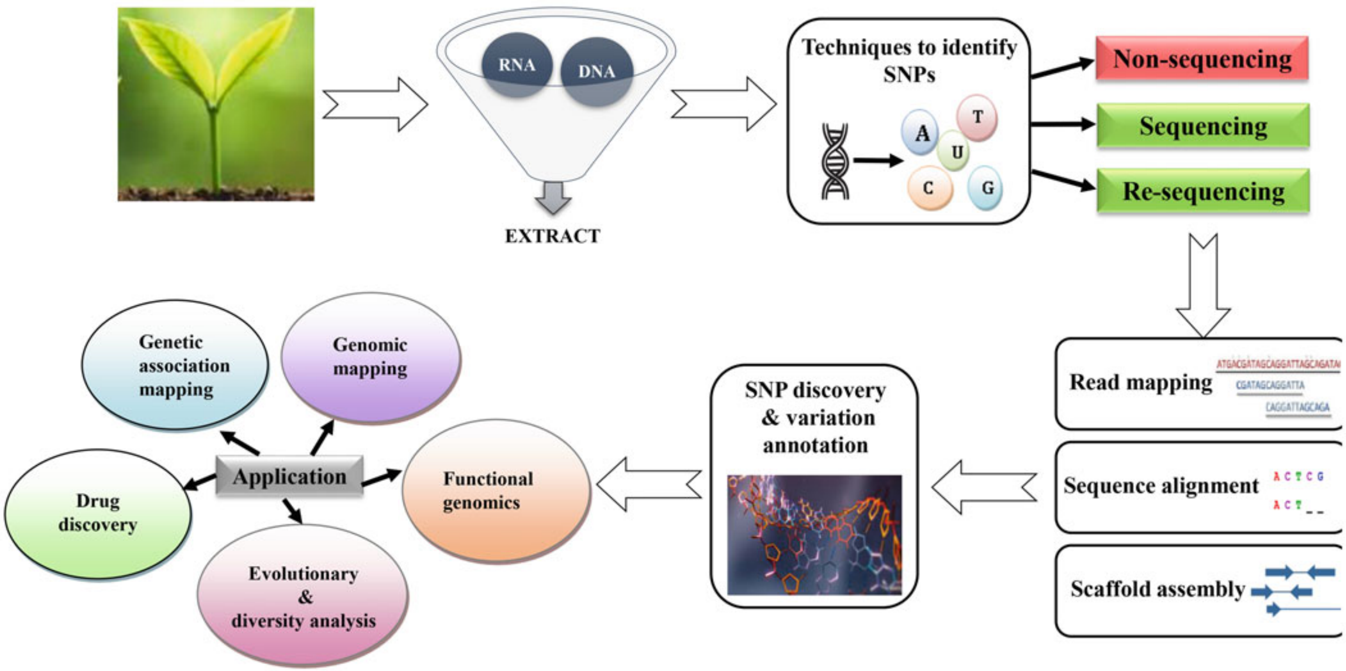


Fig. 17.1 Graphical representation and the overview of plant SNP identification and discovery

17.2.1.2 Oligonucleotide Ligation Assay (OLA)

Another method for detecting known alleles that differs by single nucleotide change was established by Hood and his team [38]. OLA requires the details of base pairing capability between the 3′-end of one strand and the 5′-end of another adjacent strand before ligase form a phosphodiester bond. OLA is also known as ligase-mediated gene detection. At first, the gene of interest is amplified using PCR and then the interesting complementary strands are hybridized through which the variant bases are identified. The variant base may or may not form complementarity at the 3′-end of the first ASO (Allele-specific oligonucleotide). The second strand tries to make perfect complementarity with the adjacent strand. If it exactly matches, then the ligase seals them by making a phosphodiester bond. The advantages of OLA are null-false positive and it is easy for automation [39].

17.2.1.3 Single-Strand Conformation Polymorphism (SSCP)

SSCP detects mutation based on the electrophoretic mobility of single-strand DNA attributed to secondary structure changes attained by the molecule due to the occurrence of variation [40]. A target single-strand DNA denatures at extreme temperature and tries to generate a DNA hybrid quickly after placing it on ice. It folds into the lowest free energy conformation and base stacking and here, hydrogen bonds play important role in forming complementary. If the strand has any change at a single base, the folding of DNA hybrid will not be the same as the native single-strand DNA and it may form a random coil or other conformation. The target and modified strands may show different mobility on an electrophoresis gel. Through this approach, both known or unknown allelic forms of a gene will be detected [41]. Since SSCP is a rapid and reliable technique for gene analysis, it is predominantly useful for the detection of point mutations and polymorphism typing. It identifies heterozygosity of DNA of with similar molecular weight fragments and detects bases responsible for the mutation through the mobility of single-stranded DNA changes while the GC content remains the same. SSCP is effective for DNA fragments of size varing about 200–800 bp that may be amplified using specific PCR primers of 20–25 bp. Gel electrophoresis of single-strand DNA detects sequence variation in those amplified fragments. The key factors for using SSCP are the pre-requisite for the low quantities of DNA and also the codominance of alleles. Even though, SSCP is highly considered as a potential tool for DNA polymorphism, yet in plants, it has to be well optimized and developed in discriminating progenies for designing crucial traits to improvise the crop plants.

17.2.1.4 Random Amplification of Polymorphic DNA

There is a need for sequence information to design primers to amplify the specific gene of interest. In 1990, single short random oligonucleotides of an interested gene sequence was used to prime the amplification of genomic sequences that can reproduce and detect polymorphisms [42, 43]. The short random oligonucleotides of the sequence will establish complementary to a locus or numerous loci within the genome. If the complementarities are in correct orientation with a distance optimum to amplify, it largely amplifies by PCR. The amplified DNA fragments are

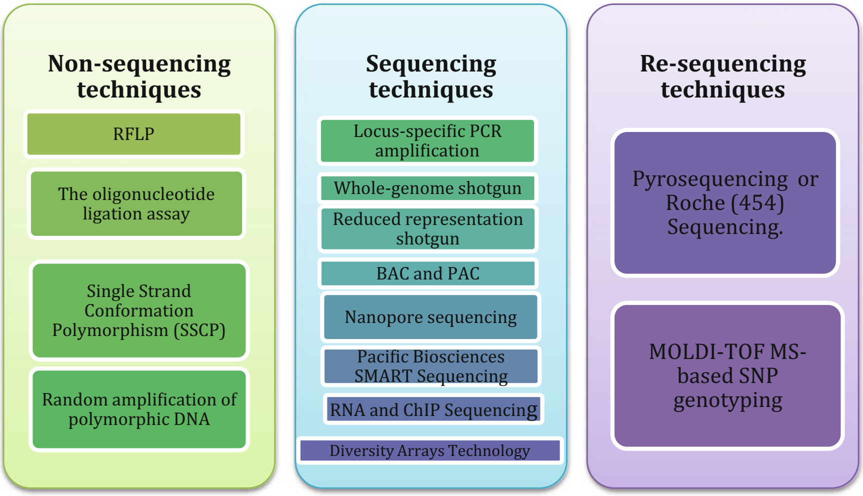

Non-sequencing techniques
RFLP
The oligonucleotide ligation assay
Single Strand Conformation Polymorphism (SSCP)
Random amplification of polymorphic DNA
Sequencing techniques
Locus-specific PCR amplification
Whole-genome shotgun
Reduced representation shotgun
BAC and PAC
Nanopore sequencing
Pacific Biosciences SMART Sequencing
RNA and ChIP Sequencing
Diversity Arrays Technology
Re-sequencing techniques
Pyrosequencing or Roche (454) Sequencing.
MOLDI-TOF MS-based SNP genotyping

Fig. 17.2 Experimental technique to identify and analyze plant SNPs. The standard experimental techniques and new methods are currently employed for plant SNP study and they are kept up-to-date to address the challenges. There are many techniques in non-sequencing, sequencing, and resequencing analysis of plant SNPs, and some of the widely used methods are mentioned

independent of one another and different in length; then, the amplified gene products are separated and identified by gel electrophoresis. If polymorphism presents in the DNA fragment, it can be detected after ethidium bromide staining by analyzing the presence or absence of a particular fragment band. The major factors such as least time consumption and need for low quantities of DNA with the feasibility of assay have led to the usage of RAPD in various applications. For instance, genetic mapping both at individual and between the closely related species, and also to overcome the barriers that are yet to be solved by other markers [43]. Nevertheless, it lacks reproducibility due to the need for higher quality DNA and a high standard experimental process that is highly sensitive to the reaction conditions. Although RAPD includes various techniques such as Arbitrarily Primed-Polymerase Chain Reaction (AP-PCR), DNA Amplification Fingerprinting (DAF) and Multiple Arbitrary Amplicon Profiling (MAAP), that are non-locus specific with a higher possibility of amplifying a variety of organisms and low reproducibility rate. Hence RAPD is known to be least considered for polyploidy-specific analysis (Fig. 17.2).

17.2.2 Sequencing Techniques

17.2.2.1 Locus-Specific PCR Amplification

Locus-specific PCR amplification (LSA) amplifies each locus of interest to produce several copies using locus-specific primers. The locus is amplified in a large number from a population sample and the amplified loci are compared among them to discover SNPs. The PCR product is applied on a nylon membrane that contains digoxigenin-ddUTP-labelled oligonucleotide probes for dot blot hybridization. And, Lumiphos 480 is treated with bound PCR products, and their signals are collected and analyzed. LSA is an expensive method for large-scale studies because it uses digoxigenin-labelled ddUTP, anti-dig Fab antibody, and Lumiphos 480 (a chemiluminescent) [44].

17.2.2.2 Whole-Genome Shotgun (WGS) Sequencing

The whole-genome shotgun, WGS, approach randomly fragments the entire genome and the size of the fragments ranges between 20 and 300 kb. Then, the fragments are cloned for producing a large number of copies. After sequencing all the fragments, they are reassembled to find the sequence of the whole-genome by aligning them. If a reference genome is available, they are easy to map. It is useful for rice genome analysis because the genome of *Oryza sativa* is available and few crop plant complete genomes are also available. If the reference genome is not present, de novo assembly has to be implemented to determine the genome, but the reliability is less compare to reference-based derived genome. With the improved techniques and computational powers, the accuracy of sequencing the whole genome has improved meantime. Nowadays, WGS is used to reduce error rates by overcoming the gaps in the genome and hence it is a more efficient sequencing method if a reference genome is available. WGS approach requires high computing power to align and assemble shotgun sequences against the reference genome and further analysis [45]. Using the WGS, the hexaploid wheat genome was determined through de novo assembly which employed both WGS-based assembly and sequence-based genetic map [46].

17.2.2.3 Reduced Representation Shotgun

In RRS, reduced representation shotgun, average DNA fragments size (D) is determined from the genome size (G) and several reads (N) and it is approximated for each run. The DNA fragments are digested by BglII and cloned into M13mp19/18 RFI DNA vectors. Then, the sequences of DNA fragments are obtained using standard methods through dye-primer chemistry, dye-terminator chemistry, or any capillary sequencer. It is a simple but powerful method for creating SNP maps. RRS can resample a subset of the genome from a population. Through comparative analysis, it determines accurate SNPs by efficient algorithms. The comparative analysis may extend to increase the yield of SNPs and mapping position against the genome by aligning to the available genome sequences. RRS facilitates faster and cost-effective construction of SNP maps based on a representative fraction of the genome and it could be implemented for various biomedical applications and agricultural importance [47–51]. This approach can be utilized for an organism that lacks a reference genome and the reads can be mapped to either with the reference or draft genome on their availability, which aids in deciphering additional functional insights [52–54]. For the species lacking reference genome, either by featuring non-bisulfite-converted samples or by inference of un-converted references from bisulfite-treated reads, accession-specific references can be generated only for the loci that are analyzed using RRBS. Apart from that, RRBS-loci were suggested to be markers for linkage mapping in plants based-on the characterization of variations in DNA methylation and higher heritability of cytosine methylation [55–57]. Although RRBS addresses various characteristics of DNA methylation, RRBS also aids in identifying the target gene loci in the non-model organisms by annotating and identifying the functional regions of a genome. However, limitation of the RRBS approach is the evolution of genes and their functions over time, thereby precluding the appropriate interpretation specifically in non-model species [58].

17.2.2.4 BAC and PAC

Bacterial artificial chromosome (BAC) intakes a DNA fragment, and it is widely used in molecular biology and genomic studies as an engineering DNA molecular tool. BACs have a huge role and impact in nucleotide sequencing as well as genomic sequencing. DNA fragments, ranging from 100 to about 300 kb, are inserted into BACs. Then, BACs with the DNA inserts are engulfed by bacterial cells. While bacterial cells grow, they multiply the genetic materials. Through this the BAC-DNA inserts also multiply in number. Finally, the DNA inserts are purified and sequenced using any sequencer [59, 60]. PAC, a P1-derived artificial chromosome, is a cloning system derived from P1 bacteriophage. P1 cloning vector system was developed by Nat Sternberg and colleagues [61]. This vector contains P1 packaging sites (PAC) where two P1 loxP recombination sites flanking the clone insert compactly using P1 Cre recombinase. It is capable of carrying large (in the range between 100 and 300 kbp) DNA fragments in *E. coli* cells for various bioengineering applications. PAC is used for sequencing the whole genome of few organisms within the size limit and is widely used for cloning larger DNA fragments. PAC method uptakes the DNA fragments via conjugation and there is no possibility of any additional DNA modifications and the DNA products from this method are largely useful for gene cluster analysis, synthetic biology, and so on. Both BAC and PAC are extensively involved in sequencing techniques to discover SNPs.

17.2.2.5 Nanopore Sequencing

Nanopore sequencing is an exclusive technology that detects real-time analysis of DNA or RNA through long-reads. It records electrical current changes by the molecules passing through, creates a characteristic disruption in current through a protein nanopore. The measure of the current can identify the molecule present in the nucleotide sequence. The resulting signal is decrypted to read the specific DNA or RNA sequence. Nanopore sequencing resolves complex structural variants and repetitive regions. It also simplifies de novo genome assembly and improves the quality of existing reference genomes. It helps to study both linkage and phasing and also enhances metagenomic analysis. It remarkably explores epigenetic modifications happening at each base by altering the current emitted from the respective bases. High-quality chromosome-level reference genomes of *Oryza* (*circum-basmati* and *cirum-aus*) were reported using the long-read nanopore sequencing method. There was a huge influence of structural variation and SNPs difference on chromosome 10 of the Oryza species [62]. DNA methylations are key players in the regulation of gene expression that controls many cellular processes like response to stimuli [63]. There is a dearth need for investigating CHG and CHH context-dependent methylation especially in plants [64], which is a major challenge facing by Nanopore sequencing of plant genomics. Besides, sequencing highly repetitive plant genome is a limitation of nanopore sequencing method [65, 66]. If the resolution of long-read sequence improved, it could revolutionize the plant genomic studies.

17.2.2.6 Pacific Biosciences SMRT Sequencing

PacBio sequencing is based on SMRT (Single-molecule real-time) technology, a mostly used third-generation sequencing technology, sequences the long-reads. Template DNAs (or SMRTbells) are carefully constructed with two hairpin-forming adaptors and ligated to form a stem-loop structure. The non-loop forming fragments are washed away and the loop forming region prefers to bind with the sequencing primers. The proper ligated template DNAs are amplified using a single polymerase which is immobilized at the bottom of ZMW (zero-mode waveguide). A phosphor-linked NTP (dNTP) is utilized for the synthesis by the polymerase and incoming dNTP makes fluorescence (each nucleotide is flagged with a different fluorophore). After forming the phosphodiester bond, it releases dye-linker-pyrophosphate into the ZMW and the cycles are continued for the next upcoming nucleotide temporarily. Until detecting non-regular nucleotides, it records the fluorescence pulse of dNTPs and it stops while it finds modified bases. PacBio sequencing reads real-time sequence and kinetic variations interpreted from the light-pulse movie which is also capable of detecting base modifications, such as methylation [67]. PacBio sequencing is applied in the biomedical application through studying the human genome extensively and it has to be improved for plant genome assembly to fulfil the difficulties in completing non-model plant genomes [68]. To improve the accuracy of detecting DNA modifications, and to reduce the required coverage of control data free of modifications, an empirical Bayesian hierarchical model was incorporated in the traditional PacBio sequencing [69].

17.2.2.7 RNA and ChIP Sequencing

RNA seq is a popular NGS technology for SNP identification, transcriptional profiling, RNA processing, differential gene expression, and other related analysis of a sample. First, the whole RNA of the sample is extracted and converted into cDNA fragments. With the cDNA fragments, appropriate adapter cum sequencing specific functional sites are added. Then, these fragments are amplified according to the need for the depth of sequence. The library of fragments is sequenced using available sequencers either in single-end or pair-end reads. The reads may be assembled against the reference genome or through de novo assembly methods. This method has been performed on various plant species including Arabidopsis [70], soybean [71], rice [72], and maize for transcript profile analysis and RNA splicing using a reference genome. The de novo assembly-based SNP discovery was performed in non-model plants like Eucalyptus, Rapeseed, and Alfalfa. Digital gene expression [73] and Illumina RNASeq [74] can perform both qualitative and quantitative genome analysis and the result permits the identification of rare transcripts and splicing variants through profiling alternative splicing events of a gene [75]. RNA-seq is one of the high-throughput based sequencing technologies and it provides high accuracy and cost-effectiveness. The advantage of RNA-seq analysis is that it simultaneously discovers thousands of SNPs along with expression levels of functional genes. The drawbacks of this method are prone to error through reverse transcriptase activity; the transcriptase may produce unauthentic complementary DNA and synthesis of artefactual cDNA as a result of template switching.

Chromatin immunoprecipitation and ChIP-seq together used to map DNA-binding protein and protein-DNA interactions [12] across the genome. ChIP-seq is also used to map histone modifications and high resolution nucleosomes, good coverage, and less noise. ChIP-Seq is not directly useful for SNP discovery, but the presence of SNP data together with ChIP-Seq provides allele-specific states of chromatin modifications and rearrangements. The sensitivity and specificity of mapping of TF (transcription factor) binding sites are increased and they facilitate TF binding motif discovery and target identification. The whole genome, RNA-seq, and ChIP-Seq data are analyzed using either a reference sequence or, de novo assembly by various software, algorithms, and pipelines, and they are kept on updating to improve the accuracy and efficiency. Together RNA-seq and ChIP-seq have the potential of elucidating the transcriptional regulatory network of important biological processes of plants along with gene expression analysis.

17.2.2.8 Diversity Arrays Technology (DArT) for SNP

DArT is a microarray hybridization-based technique that detects DNA variations across the polymorphic loci which are spread over the entire genome. For analysis of DArT data, we may use the softwares namely, DArTsoft and DArTdb. It is useful to identify SNP markers of non-model species; especially, it helps for crop plants whose genomes are polyploid. Initially, it needs to prepare the DNA fragments from genomic DNA using PstI adaptors. Libraries of arrays are prepared from the whole DNA fragments using diversity arrays technology which is special microarrays that is capable of detecting and analyzing DNA polymorphism in a genome. The cloned and amplified DNA fragments are transferred to microarray plates (Ex. Affymetrix). During purification of each fragment, fluorescence dye (Cy3 or Cy5) is added and then hybridization and washing are performed to remove non-DNA fragments. Later, an Affymetrix scanner is used for recording signal intensities from the sequences, and then the data are genotyped [76]. DArT was successfully applied for genotyping of barley plant and subsequent SNP markers analysis were performed to understand their functions [77]. DArT acts as an alternate strategy that preferentially targets low-copy genomic regions, specifically allows automation of data acquisition [76]. The features such as complexity reduction, cost competitiveness and the simultaneous typing of various polymorphic loci across the genome (i.e. based on DNA hybridization) have contributed to the application of DArT in mapping several ‘orphan’ crops. For example, the DArT technique is used to the analysis of many crop plants genomes; namely, rice, barley, eucalyptus, Arabidopsis, cassava, wheat, and pigeon-pea. The major limitations of DArT are the requirement of laboratory facilities, high investment, and skilled manpower.

17.2.3 Re-sequencing Techniques

17.2.3.1 Pyrosequencing or Roche (454) Sequencing

Pyrosequencing is one of the alternative methods of DNA sequencing using dideoxy chain termination technology which detects the released pyrophosphate (PPi). It is used for de novo sequencing and confirmation of sequence, and it has a limitation in determining the read length. It is also known as 454 sequencing which uses sequencing by synthesis method, and it is the first parallel sequencing technology [78]. The genomic DNA is fragmented and the fragments serve as a biotinylated pyrosequencing template. This biotinylated template is purified and added with sequencing primers. Then, the template is catalyzed by DNA polymerase, and each deoxyribonucleotide (dNTP) addition to the template complementary sequence releases pyrophosphate. Further, ATP sulfurylase acts on pyrophosphate and produces ATP from ASP (Adenosine 5′ phosphosulfate). The ATP molecule is used by the luciferase enzyme to convert luciferin to oxyluciferin thus generates visible light and it recorded by a charge-coupled device [78]. Apyrase degrades unincorporated dNTPs and ATP before moving on to the addition of the next nucleotide, and subsequent nucleotides are added through repeating the protocols. As the process continues, the signals are recorded in the program which generates the sequence data. Using appropriate bioinformatics tools, the data are analyzed.

17.2.3.2 MALDI-TOF MS-Based SNP Typing

MALDI is one of the revolutionary high-throughput methods that altered the conventional method of sequencing, which involves the separation of DNA fragments resulting from enzymatic sequencing using gel electrophoresis [79]. MALDI technique is also used for ionization relative to mass-analysis of large biomolecules [80]. Hence, in general, there is predominant detection of single charged molecular ions (both negative and positive) by MALDI–TOF. Initially, polymorphisms containing genes of interest need to be amplified using PCR from genomic DNA. The PCR product will be purified and used for MALDI-TOF analysis; before that allele-specific primer is annealed with the PCR product for SNP analysis. Then, the PCR products are again amplified using dNTP or/and ddNTPs with a DNA polymerase, and the products are subjected to MALDI-TOF MS analysis (Fig. 17.3). The

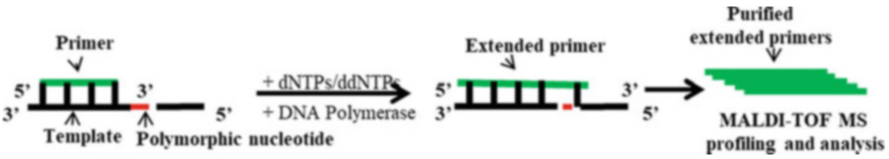


Fig. 17.3 A nucleotide-specific primer is synthesized to anneal with a PCR template immediate downstream of polymorphic position. Four deoxy or dideoxy-nucleotide triphosphate (ddNTP or dNTP) and DNA polymerase are mixed and the primer gets extended by a nucleotide (N), where N is complementary to the nucleotide at the polymorphic position. The extended primer is purified and subjected to MS analysis (*m*/*z* value). Accurate mass measurement of the extended primer identifies the polymorphic nucleotide exactly

extended primers, allele-specific primers, are purified through the MALDI matrix and detected by the mass-charge ratio (*m*/*z* value) specific to the added nucleotide to the extension. This approach is more effective in detecting and discovering novel SNPs [81]. The problems with this approach are the loss of signal intensity and mass resolution with the increasing length of DNA. Enzymatic DNA sequencing coupled with MALDI–TOF MS analysis has been proposed in recent studies for effective identification of unknown single-nucleotide substitutional mutations/SNPs.

17.3 Plant SNP Databases

Various databases are available related to plant SNPs, and databases are implemented based on various algorithms for processing the user queries, the imporant databases are mentioned in Table 17.1. **CropSNPdb** [87] is a simple user interface database that provides genotyping arrays of crop plants and is also based on comparing genotyping algorithms. It enables comparative analysis of the SNP alleles among the crop plants as queried by the user. **CropSNPdb** includes genotyping arrays of 535 Brassica lines from reported datasets and 309 Wheat lines from T3-Wheat. In **PLANET-SNP** [82], the users are provided with different algorithm options such as Bayes Net Naïve, Bayes SVM, J48, and Random Forest for SNP detection and evaluation. **PLANET-SNP** has an interactive GUI (Graphical User Interface) for the detection and analysis of SNPs and also provides a different representation of SNPs. The **Rice Stress-Resistant SNP** (RSRS) database [91] contains more than 9.5 million SNPs of stress-resistant strains and 797 stress-resistant genes of rice from 400 plus rice varieties. This database comprises of SNP function, phenotype information, and genome annotation. **SNP-Seek II** [83] implemented BWA (Burrows-Wheeler Aligner) and GATK (Genome Analysis Toolkit) identified nearly 40 M SNP variants from the 3000 Rice Genomes Project dataset. The database covers genotype, phenotype, and variety information of *Oryza sativa* L. (rice plant). SNP genotyping datasets are obtained from the five reference genomes namely; Nipponbare (*Temperate japonica*), IR 64 (*Indica*), 93–11 (*Indica*), DJ 123 (*Aus*), and Kasalath (*Aus*). The database provides access to rice research information and facilitates rice variety improvement via discovering new gene–trait associations and accelerated breedings. Although **Information Commons for Rice** (**IC4R**) [92] provides ultra-high-density maps of rice variations, the raw SNPs could not be valid for public use. Raw genotypic data of 18 million SNPs could be used by different bioinformatics pipelines for different utilizations like population genetics, evolutionary study, and genomic breeding of new rice varieties. The outcomes of rice genome SNPs is to develop an integrative SNP resource of rice genome database; that is **SnpReady for Rice** (**SR4R**) **database** [93]. The SR4R database web interface utilizes the ridge regression BLUP algorithm for retrieving SNPs and utilization of online toolkits for analyzing population genetics. Besides, the SNPs have enabled the users to access four different reference SNP panels comprising of HapMap SNPs, tagSNPs, fixed SNPs, and barcodeSNPs, thereby providing extensive analysis information of the genetic diversity among the *Oryza sativa* population

Table 17.1 Databases for plant SNPs

S. no.	Tools	Working algorithm					References
1.	**PLANET-SNP**	Bayes Net	Naive Bayes	SVM	J48	Random Forest	[82]
2.	**SNP-Seek II**	Shuffle-LAGAN glocal chaining algorithm					[83]
3.	**Rice RelativeGD**	Markov cluster algorithm					[84]
4.	**SR4R**	Ridge regression BLUP					[85]
5.	**HapRice**	Minimum distance method	Neighbour-joining method				[86]
6.	**CropSNPdb**	Comparing the genotyping algorithm					[87]
7.	**FreeBayes**	Bayesian					https://github.com/freebayes/freebayes
8.	**GABi**	Biclustering algorithm					[88]
9.	**SNiPlay**	Median joining algorithm					[89]
10.	**QualitySNPng**	SNP calling algorithm					[90]

with the least SNP redundancy. **HapRice** [86], is an SNP haplotype database and a web tool for rice. The SNP haplotypes are well-defined by the allele frequency in cultivar groups such as Aus, Indica, tropical japonica, and temperate japonica for the world, and non-irrigated and three irrigated groups for Japanese. HapRice is also finding polymorphic SNPs in between two rice varieties to identify polymorphic sequence markers by implementing a minimum distance method and neighbour-joining method. **Gramene** [94] database has various plant genomes that are suitable for genomic colinearity and comparative genomics studies. The data are acquired either from EST (expressed sequence tag) sequences or genetic maps/publications. It contains well-curated data of rice mutants including molecular markers. The Rice Annotation Project Database (RAP-DB) [95] contains comprehensive data sets of gene annotations of rice, especially *Oryza sativa* (japonica group) cv. Nipponbare. RAP-DB and satellite databases are together to offer a better platform for plant and genome researchers. The ESTree DB [96] represents a collection of *Prunus persica* ESTs and is intended as a resource for peach functional genomics. The ESTree DB encompasses nearly 18,630 sequences, where the contig assembly is performed with CAP3, while the AutoSNP program is used for identifying the putative single nucleotide polymorphism (SNP). This database provides external access to NiceZyme (Expasy) and the KEGG metabolic pathways. An annotated SNP database for crop plants, known as **AutoSNPdb** [39], facilitates the identification of polymorphic sequences by BLAST or keyword searches and annotation will be derived from UniRef90 and GenBank which compares with the reference genomes. Besides, SNPs between any two varieties are worthwhile for targeting genomic mapping and association studies.

17.3.1 Methods for SNP Discovery

SNP-RFLP is the earliest method used to detect SNPs. The advancement of massively parallel high-throughput sequencing techniques like next-generation sequencing has reduced the cost and time in SNPs identification. Few SNP discovery NGS tools are as follows: Among the SNP calling tools predominantly used tools are Samtools/mpileup, SNVer, SOAPsnp, FreeBayes, GATK, Platypus, VarScan, and VarDict. One of the popular stand-alone NGS file conversion (SAM to BAM and vice-versa) package is Samtools. It has various modules for mapping statistics, variant calling, and assembly visualization. SOAPsnp (Short Oligonucleotide Analysis Package) shows great accuracy and consistency (Table 17.2).

SOAPsnp aligns the raw sequencing reads against available references to find the consensus sequence and then predicts SNPs. Another stand-alone statistical variant calling tool SNVer runs very fast and has excellent scalability for whole-genome sequencing. A haplotype-based variant detector tool, **Freebayes**, [100] detects variants based on read sequence match to an appropriate position in the genome while aligning to a particular target, but it does not align by the precision. This model is a direct approach for generalization than the previous approaches and it discovers and detects based on the alignments. The advantage is that it avoids one of the core

Table 17.2 Methods for SNP discovery

Method	Tools	References
The earliest method to detect SNP	SNP-RFLPing 2	http://bio.kuas.edu.tw/snp-rflping2
Advanced method for SNP discovery	Samtools/ mpileup	http://www.htslib.org/
	SNVer	[97]
	GATK	https://gatk.broadinstitute.org/hc/en-us
	FreeBayes	https://github.com/freebayes/freebayes
	SOAPsnp	http://soap.genomics.org.cn/soapsnp.html
	Platypus	https://github.com/andyrimmer/Platypus
	VarScan	[98]
	VarDict	[99]

problems, that identical sequences may align to multiple possible positions of the genome.

17.3.2 Assays to Determine SNP

The simplest and cost-effective genotyping systems are KASPar and SNPline which are capable of discovering thousands of genotyping SNPs using a laboratory protocol. The SNPline is updated into SNPlite or SNPline XL versions with flexibility in sample numbers and SNP assays methods. Primer extension chemistry and MALDI-TOF are joined together for genotyping in the iPLEX Gold technology which is based on the MassARRAY system. The iPLEX Gold system has been implemented and has become a popular technique due to its better precision and cost-effectiveness (Table 17.3). GeneChip arrays (Affymetrix) and BeadChips (Illumina) are high-throughput chip-based genotyping assays that can validate millions of SNPs per reaction in a genome.

17.3.3 SNP Visualization Tools

The interactive SNP visualization tools that are publically available are Tablet, SNP-VISTA or Savant, ViewGene, SNPVersity, Quality SNPng (SNP detection and visualization), and CircosVCF (Fig. 17.4).

Table 17.3 Assays to determine SNPs

S. no.	Assay	Number of SNPs genotyped per reaction	Sample required	Advantage	Disadvantage	References
1.	KASPar	Few to 1000 of SNPs	≥5 ng	The flexible and low genotyping error rate	Cost-effective and time-consuming	[101]
2.	SNPline (SNPlite or SNPline XL)	Few to thousands of SNPs	5 ng	Scalable, cost-effective and flexible	Time-consuming than NGS technologies	https://www.biosearchtech.com/products/instruments-and-consumables/genotyping-instruments/snpline-genotyping-automation
3.	iPLEX Gold	>1000 SNPs	5 ng	Multiplex, sensitive, fast, and accurate	Well-trained personnel and computational knowledge are required	https://www.cd-genomics.com/Sequenom-MassARRAY-iPLEX-Gold.html
4.	AffymetrixGeneChip arrays	A million SNPs per reaction	0.13–3μg	The rapid result, monitor every gene expression in the genome, and widely used	Well-trained personnel and computational knowledge are required	https://www.affymetrix.com/about_affymetrix/outreach/educator/microarray_curricula.affx
5.	IlluminaBeadChips	A million SNPs per reaction	200–400 ng	Millions of SNPs assayed at once and accurate	Well-trained personnel and computational knowledge are required	[102]

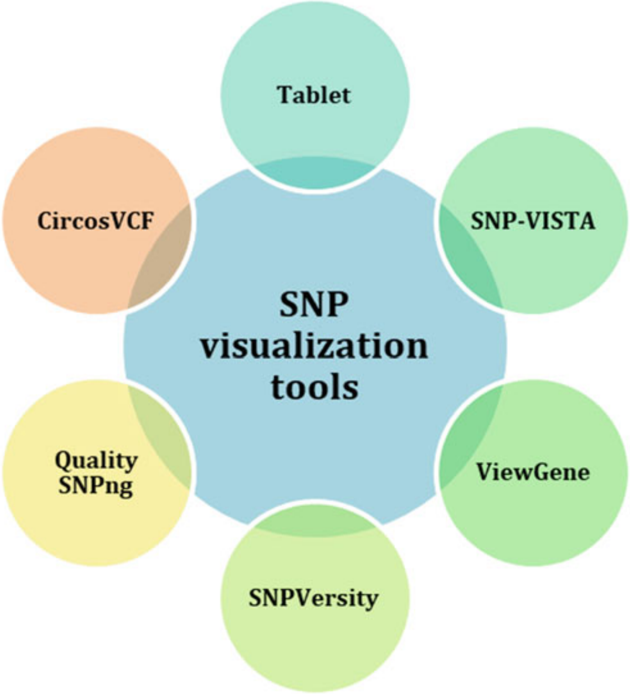


Fig. 17.4 SNP visualization tools are given for SNP analysis

17.4 Applications

SNPs aid in constructing high-resolution genetic maps. Genetic linkage maps of some economically important plants like rice [20], cotton [19], and Brassica [103] were assembled using SNPs. Genetic maps of crops are widely useful in molecular breeding to improve yield, stress tolerance, drought tolerance, and so on. By considering SNP-based genetic maps, flowering-specific genes of Brassica and maize [21] plants were identified and molecular marker genes at particular plant growth stage will be able to determine by SNPs. And, these kinds of gene-specific SNP markers are mostly discovered by RNA-Seq and/or ESTs [104].

The special advantage of SNPs over other polymorphisms is that they are least pretentious by homoplasy hence such kind of inter-species evolution could be explained by SNPs. The SNPs of samples can be successfully implemented for phylogenetic analysis with reduced homoplasy and similar analysis can be applied to understand the phylogeny among crops [105]. SNPs of important genes involved in drug metabolism are most effective for pharmacogenetics investigation and it helps personalized medicine, and Cytochrome P450 (CYP2D6) gene is one of the best examples [106]. A GWAS study of rice identified SNPs and constructed a high-density haplotype map for phenotypic variances that are useful for rice genetics and breeding [107]. Further, the GWAS investigation unravelled the salt tolerance-specific SNPs of rice at the flowering stage thus aid in breeding salt tolerance varieties and agronomic traits [108]. The SNPs of plant growth-regulating genes will have a huge impact on crop yielding and growth, hence understanding and underpinning functional SNPs could be effectively implemented to the development of desired plant breedings [109].

17.5 Challenges

SNPs have been considered as the significant genetic marker in plant genomics to enhance both crop improvement and breeding by predicting the phenotypic changes based on the changes in genotype. The variations within the genome are highly very crucial in plant genomes due to its genetic diversity among the population. With polyploid species and their subgenomes, it is even more challenging to distinguish the homologous SNPs that are specific to the species from the allelic SNPs [18]. Hence several advancements were developed in the genome assembly-based analysis to overcome the barriers and accomplish species-focused plant genome reference assembly to decipher the relative interaction of the subgenomes within the polyploidy and heterozygous species. The applications of NGS technologies are SNP identification and discovery along with finding linkage map construction, understanding genetic diversity, constructing association maps, and marker-assisted selection breeding. These approaches are applied for several plant species and remaining plants and crops need to be explored by these analysis. Polyploids, repeated genetic elements, paralog sequences, and partial or incorrect reference genomes make uncertainties in SNP calling, identification, and discovery. The read mapping in NGS may cause misalign that leads to inaccurate SNP identification if the sequencing contains an error (due to erroneous base calling). Hence, there is a huge need for programs to discover SNPs with more accuracy and minimize erroneous SNP calling. Sequencing and read mapping errors may affect the SNP validation rate so these two major factors have to be improved. However, there are different approaches to reduce the cost and simplify the SNP marker identification and discovery. Henceforth, the techniques of sequencing such as haplotype phasing, structural variant analysis, and de novo pan-genomics are the emerging frontiers in plant genome assembly [110]. But, there is a lack of consensus among the various tools and algorithms involved in SNP identification. Several techniques have been reported for the detection of SNPs in crop plants yet most of them have their respective pitfalls. In the case of genotyping by sequencing (GBS), the major disadvantage is the improper digestion of restriction enzymes that may lead to absence of important regions in the genome from the genomic libraries [111]. Moreover, this may also lead to erroneous data from sequencing [112]. The restriction-site associated DNA sequencing (RAD-seq) technique does not need the prerequisite of reference genome and this method is highly feasible; however, loss of sheared restriction sites may occur due to sequence polymorphisms [113]. The specific locus amplified fragment sequencing (SLAF-seq) is a major cost-effective method implemented for SNP-based genotyping on a large scale, yet it lacks whole-genome coverage [114]. On considering gene expression analysis tools, application of Chromatin immunoprecipitation with ChIP-seq approach in plants and crops is challenging due to vigorous disruption of the cell wall, presence of phenolic compounds, and polysaccharides, and selecting good quality antibodies for plant is the most difficult task. Even though NGS and SNP genotyping technologies made SNPs the most widely used marker for genetic studies, still there is a need for technical advancements to overcome the challenges involved in plant SNPs identification, genomic divergences of plants and in determination of the associate genomic variations of phenotypic traits.

17.6 Future Perspective

SNP identification and discovery certainly made a quantum leap with growing sequencing features and technologies, and there are plenty of SNPs available for several genomes including large and complex plant and animal genomes. Unlike model organisms such as humans and Arabidopsis, SNPs in plants and crops remain inadequate for the time being, but with the help of advanced sequencing and computational tools, the reference genomes of most of the crops and other plants will be sequenced in the near future. Their understanding and SNPs will make ground-breaking results for improved plant breeds and understanding of the functions of plant genes and genomes.

Acknowledgements The authors thank the Centre for Bioinformatics for providing the necessary facility for the work.

Conflict of Interest There is no conflict of interest to declare.

References

1. Consortium IH. A second generation human haplotype map of over 3.1 million SNPs. Nature. 2007;449:851.
2. Syvänen A-C. Toward genome-wide SNP genotyping. Nat Genet. 2005;37:S5–S10.
3. MacLeod C. Obituary notice. Oswald Theodore Avery, 1877–1955. Microbiology. 1957;17:539–49.
4. Mayor S. First human chromosome is sequenced. Br Med J. 1999;319:1453.
5. Khan AW, Garg V, Roorkiwal M, Golicz AA, Edwards D, Varshney RK. Super-pangenome by integrating the wild side of a species for accelerated crop improvement. Trends Plant Sci. 2020;25:148–58.
6. Powell W, Machray GC, Provan J. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1996;1:215–22.
7. Jang S-J, Sato M, Sato K, Jitsuyama Y, Fujino K, Mori H, Takahashi R, Benitez ER, Liu B, Yamada T. A single-nucleotide polymorphism in an endo-1,4-β-glucanase gene controls seed coat permeability in soybean. PLoS One. 2015;10:e0128527.

8. Varshney R, Graner A, Sorrells M. Genomics-assisted breeding for crop improvement. Trends Plant Sci. 2005;10:621–30.
9. Bevan MW, Uauy C, Wulff BB, Zhou J, Krasileva K, Clark MD. Genomic innovation for crop improvement. Nature. 2017;543:346–54.
10. Yuan Y, Bayer PE, Batley J, Edwards D. Improvements in genomic technologies: application to crop genomics. Trends Biotechnol. 2017;35:547–58.
11. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF. Complementary DNA sequencing: expressed sequence tags and human genome project. Science. 1991;252:1651–6.
12. Solomon MJ, Larsen PL, Varshavsky A. Mapping proteinDNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene. Cell. 1988;53:937–47.
13. Alkan C, Coe BP, Eichler EE. Genome structural variation discovery and genotyping. Nat Rev Genet. 2011;12:363–76.
14. Chung W-H, Jeong N, Kim J, et al. Population structure and domestication revealed by high-depth resequencing of Korean cultivated and wild soybean genomes. DNA Res. 2014;21:153–67.
15. Rafalski A. Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol. 2002;5:94–100.
16. Shi Z, Liu S, Noe J, Arelli P, Meksem K, Li Z. SNP identification and marker assay development for high-throughput selection of soybean cyst nematode resistance. BMC Genomics. 2015;16:314.
17. Semagn K, Bjørnstad Å, Ndjiondjop MN. Principles, requirements and prospects of genetic mapping in plants. Afr J Biotechnol. 2007;525(25):2569–87. https://doi.org/10.4314/ajb.v5i25.56082.
18. Ganal MW, Altmann T, Röder MS. SNP identification in crop plants. Curr Opin Plant Biol. 2009;12:211–7.
19. Byers RL, Harker DB, Yourstone SM, Maughan PJ, Udall JA. Development and mapping of SNP assays in allotetraploid cotton. Theor Appl Genet. 2012;124:1201–14.
20. Xie W, Feng Q, Yu H, Huang X, Zhao Q, Xing Y, Yu S, Han B, Zhang Q. Parent-independent genotyping for constructing an ultrahigh-density linkage map based on population sequencing. PNAS. 2010;107:10578–83.
21. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, Ersoz E, Flint-Garcia S, Garcia A, Glaubitz JC. The genetic architecture of maize flowering time. Science. 2009;325:714–8.
22. Akond M, Liu S, Schoener L, Anderson JA, Kantartzi SK, Meksem K, Song Q, Wang D, Wen Z, Lightfoot DA. A SNP-based genetic linkage map of soybean using the SoySNP6K Illumina Infinium BeadChip genotyping array. Plant Genet Genomics Biotechnol. 2013;1:80–9.
23. Yano K, Yamamoto E, Aya K, et al. Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. Nat Genet. 2016;48:927–34.
24. Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985;230:1350–4.
25. Tanksley SD, Young ND, Paterson AH, Bonierbale MW. RFLP mapping in plant breeding: new tools for an old science. Bio/Technology. 1989;7:257–64.
26. McCouch SR, Kochert G, Yu ZH, Wang ZY, Khush GS, Coffman WR, Tanksley SD. Molecular mapping of rice chromosomes. Theor Appl Genet. 1988;76:815–29.
27. Gentzbittel L, Vear F, Zhang Y-X, Berville A, Nicolas P. Development of a consensus linkage RFLP map of cultivated sunflower (Helianthus annuus L.). Theor Appl Genet. 1995;90:1079–86.

28. Helentjaris T, Slocum M, Wright S, Schaefer A, Nienhuis J. Construction of genetic linkage maps in maize and tomato using restriction fragment length polymorphisms. Theor Appl Genet. 1986;72:761–9.
29. Chao S, Sharp PJ, Worland AJ, Warham EJ, Koebner RMD, Gale MD. RFLP-based genetic maps of wheat homoeologous group 7 chromosomes. Theor Appl Genet. 1989;78:495–504.
30. Keim P, Diers BW, Olson TC, Shoemaker RC. RFLP mapping in soybean: association between marker loci and variation in quantitative traits. Genetics. 1990;126:735–42.
31. Tanksley SD, Ganal MW, Prince JP, de Vicente MC, Bonierbale MW, Broun P, Fulton TM, Giovannoni JJ, Grandillo S, Martin GB. High density molecular linkage maps of the tomato and potato genomes. Genetics. 1992;132:1141–60.
32. Varshney RK, Marcel TC, Ramsay L, Russell J, Röder MS, Stein N, Waugh R, Langridge P, Niks RE, Graner A. A high density barley microsatellite consensus map with 775 SSR loci. Theor Appl Genet. 2007;114:1091–103.
33. Radhika P, Gowda SJM, Kadoo NY, Mhase LB, Jamadagni BM, Sainani MN, Chandra S, Gupta VS. Development of an integrated intraspecific map of chickpea (Cicer arietinum L.) using two recombinant inbred line populations. Theor Appl Genet. 2007;115:209–16.
34. Miller JC, Tanksley SD. RFLP analysis of phylogenetic relationships and genetic variation in the genus Lycopersicon. Theor Appl Genet. 1990;80:437–48.
35. Desplanque B, Boudry P, Broomberg K, Saumitou-Laprade P, Cuguen J, Van Dijk H. Genetic diversity and gene flow between wild, cultivated and weedy forms of Beta vulgaris L. (Chenopodiaceae), assessed by RFLP and microsatellite markers. Theor Appl Genet. 1999;98:1194–201.
36. Kesawat MS, Kumar BD. Molecular markers: it's application in crop improvement. J Crop Sci Biotechnol. 2009;12:169–81.
37. Gupta M, Chyi Y-S, Romero-Severson J, Owen JL. Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theor Appl Genet. 1994;89:998–1006.
38. Landegren U, Kaiser R, Sanders J, Hood L. A ligase-mediated gene detection technique. Science. 1988;241:1077–80.
39. Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D. AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res. 2009;37:D951–3.
40. Orita M, Suzuki Y, Sekiya T, Hayashi K. Rapid and sensitive detection of point mutations and DNA polymorphisms using the polymerase chain reaction. Genomics. 1989;5:874–9.
41. Barany F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proc Natl Acad Sci. 1991;88:189–93.
42. Welsh J, McClelland M. Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res. 1990;18:7213–8.
43. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV. DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res. 1990;18:6531–5.
44. Cereb N, Maye P, Lee S, Kong Y, Yang SY. Locus-specific amplification of HLA class I genes from genomic DNA: locus-specific sequences in the first and third introns of HLA-A,-B, and-C alleles. Tissue Antigens. 1995;45:1–11.
45. Weber JL, Myers EW. Human whole-genome shotgun sequencing. Genome Res. 1997;7:401–9.
46. Chapman JA, Mascher M, Buluç A, Barry K, Georganas E, Session A, Strnadova V, Jenkins J, Sehgal S, Oliker L. A whole-genome shotgun approach for assembling and anchoring the hexaploid bread wheat genome. Genome Biol. 2015;16:1–17.
47. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, Lander ES. An SNP map of the human genome generated by reduced representation shotgun sequencing. Nature. 2000;407:513–6.

48. Gu H, Smith ZD, Bock C, Boyle P, Gnirke A, Meissner A. Preparation of reduced representation bisulfite sequencing libraries for genome-scale DNA methylation profiling. Nat Protoc. 2011;6:468–81.
49. Wang S, Lv J, Zhang L, et al. MethylRAD: a simple and scalable method for genome-wide DNA methylation profiling using methylation-dependent restriction enzymes. Open Biol. 2015;5:150130.
50. van Gurp TP, Wagemaker NCAM, Wouters B, Vergeer P, Ouborg JNJ, Verhoeven KJF. epiGBS: reference-free reduced representation bisulfite sequencing. Nat Methods. 2016;13:322–4.
51. Trucchi E, Mazzarella AB, Gilfillan GD, Lorenzo MT, Schönswetter P, Paun O. BsRADseq: screening DNA methylation in natural populations of non-model species. Mol Ecol. 2016;25:1697–713.
52. Gugger PF, Fitz-Gibbon S, PellEgrini M, Sork VL. Species-wide patterns of DNA methylation variation in Quercus lobata and their association with climate gradients. Mol Ecol. 2016;25:1665–80.
53. Lea AJ, Altmann J, Alberts SC, Tung J. Resource base influences genome-wide DNA methylation levels in wild baboons (Papio cynocephalus). Mol Ecol. 2016;25:1681–96.
54. Weyrich A, Lenz D, Jeschek M, Chung TH, Rübensam K, Göritz F, Jewgenow K, Fickel J. Paternal intergenerational epigenetic response to heat exposure in male wild guinea pigs. Mol Ecol. 2016;25:1729–40.
55. Johannes F, Porcher E, Teixeira FK, Saliba-Colombani V, Simon M, Agier N, Bulski A, Albuisson J, Heredia F, Audigier P. Assessing the impact of transgenerational epigenetic variation on complex traits. PLoS Genet. 2009;5:e1000530.
56. Cortijo S, Wardenaar R, Colomé-Tatché M, Gilly A, Etcheverry M, Labadie K, Caillieux E, Aury J-M, Wincker P, Roudier F. Mapping the epigenetic basis of complex traits. Science. 2014;343:1145–8.
57. Hofmeister BT, Lee K, Rohr NA, Hall DW, Schmitz RJ. Stable inheritance of DNA methylation allows creation of epigenotype maps and the study of epiallele inheritance patterns in the absence of genetic variation. Genome Biol. 2017;18:1–16.
58. Pavey SA, Bernatchez L, Aubin-Horth N, Landry CR. What is needed for next-generation ecological and evolutionary genomics? Trends Ecol Evol. 2012;27:673–8.
59. O'Connor M, Peifer M, Bender W. Construction of large DNA segments in Escherichia coli. Science. 1989;244:1307–12.
60. Ariyadasa R, Stein N. Advances in BAC-based physical mapping and map integration strategies in plants. J Biomed Biotechnol. 2012;2012:184854.
61. Sternberg N. Bacteriophage P1 cloning system for the isolation, amplification, and recovery of DNA fragments as large as 100 kilobase pairs. PNAS. 1990;87:103–7.
62. Choi JY, Lye ZN, Groen SC, Dai X, Rughani P, Zaaijer S, Harrington ED, Juul S, Purugganan MD. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. Genome Biol. 2020;21:21.
63. Simpson JT, Workman RE, Zuzarte PC, David M, Dursi LJ, Timp W. Detecting DNA cytosine methylation using nanopore sequencing. Nat Methods. 2017;14:407–10.
64. Law JA, Jacobsen SE. Establishing, maintaining and modifying DNA methylation patterns in plants and animals. Nat Rev Genet. 2010;11:204–20.
65. Jiao W-B, Schneeberger K. The impact of third generation genomic technologies on plant genome assembly. Curr Opin Plant Biol. 2017;36:64–70.
66. Schmidt MH-W, Vogel A, Denton AK, et al. De novo assembly of a new Solanum pennellii accession using nanopore sequencing. Plant Cell. 2017;29:2336–48.
67. Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B. Real-time DNA sequencing from single polymerase molecules. Science. 2009;323:133–8.
68. VanBuren R, Bryant D, Edger PP, et al. Single-molecule sequencing of the desiccation-tolerant grass Oropetium thomaeum. Nature. 2015;527:508–11.

69. Feng Z, Fang G, Korlach J, Clark T, Luong K, Zhang X, Wong W, Schadt E. Detecting DNA modifications from SMRT sequencing data by modeling sequence context dependence of polymerase kinetic. PLoS Comput Biol. 2013;9:e1002935.
70. Weber APM, Weber KL, Carr K, Wilkerson C, Ohlrogge JB. Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. Plant Physiol. 2007;144:32–42.
71. Libault M, Farmer A, Joshi T, Takahashi K, Langley RJ, Franklin LD, He J, Xu D, May G, Stacey G. An integrated transcriptome atlas of the crop model Glycine max, and its use in comparative analyses in plants. Plant J. 2010;63:86–99.
72. Lu T, Lu G, Fan D, Zhu C, Li W, Zhao Q, Feng Q, Zhao Y, Guo Y, Li W. Function annotation of the rice transcriptome at single-nucleotide resolution by RNA-seq. Genome Res. 2010;20:1238–49.
73. Ozsolak F, Ting DT, Wittner BS, Brannigan BW, Paul S, Bardeesy N, Ramaswamy S, Milos PM, Haber DA. Amplification-free digital gene expression profiling from minute cell quantities. Nat Methods. 2010;7:619–21.
74. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet. 2009;10:57–63.
75. Zhao H, Sun L, Xiong T, et al. Genetic characterization of the chromosome single-segment substitution lines of O. glumaepatula and O. barthii and identification of QTLs for yield-related traits. Mol Breed. 2019;39:51.
76. Jaccoud D, Peng K, Feinstein D, Kilian A. Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res. 2001;29:e25.
77. Wenzl P, Carling J, Kudrna D, Jaccoud D, Huttner E, Kleinhofs A, Kilian A. Diversity arrays technology (DArT) for whole-genome profiling of barley. PNAS. 2004;101:9915–20.
78. Ronaghi M, Uhlén M, Nyrén P. A sequencing method based on real-time pyrophosphate. Science. 1998;281:363–5.
79. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, Connell CR, Heiner C, Kent SBH, Hood LE. Fluorescence detection in automated DNA sequence analysis. Nature. 1986;321:674–9.
80. Smith LM. The future of DNA sequencing. Science. 1993;262:530–2.
81. Storm N, Darnhofer-Patel B. MALDI-TOF mass spectrometry-based SNP genotyping. In: Kwok P-Y, editor. Single nucleotide polymorphisms: methods and protocols. New York: Springer; 2003. p. 241–62.
82. Bhardwaj A, Bag SK. PLANET-SNP pipeline: PLants based ANnotation and Establishment of True SNP pipeline. Genomics. 2019;111:1066–77.
83. Mansueto L, Fuentes RR, Chebotarov D, et al. SNP-seek II: a resource for allele mining and analysis of big genomic data in Oryza sativa. Curr Plant Biol. 2016;7–8:16–25.
84. Mao L, Chen M, Chu Q, et al. RiceRelativesGD: a genomic database of rice relatives for rice research. Database (Oxford). 2019;2019:baz110. https://doi.org/10.1093/database/baz110.
85. Yan J, Zou D, Li C, Zhang Z, Song S, Wang X. SR4R: an integrative SNP resource for genomic breeding and population research in rice. Genomics Proteomics Bioinform. 2020;18:173–85.
86. Yonemaru J, Ebana K, Yano M. HapRice, an SNP haplotype database and a web tool for rice. Plant Cell Physiol. 2014;55:e9.
87. Scheben A, Verpaalen B, Lawley CT, Chan C-KK, Bayer PE, Batley J, Edwards D. CropSNPdb: a database of SNP array data for Brassica crops and hexaploid bread wheat. Plant J. 2019;98:142–52.
88. Curry EW. A framework for generalized subspace pattern mining in high-dimensional datasets. BMC Bioinform. 2014;15:355.
89. Dereeper A, Nicolas S, Le Cunff L, Bacilieri R, Doligez A, Peros J-P, Ruiz M, This P. SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC Bioinform. 2011;12:134.

90. Nijveen H, van Kaauwen M, Esselink DG, Hoegen B, Vosman B. QualitySNPng: a user-friendly SNP detection and visualization tool. Nucleic Acids Res. 2013;41:W587–90.
91. Tareke Woldegiorgis S, Wang S, He Y, et al. Rice stress-resistant SNP database. Rice. 2019;12:97.
92. Consortium IP. Information commons for rice (IC4R). Nucleic Acids Res. 2016;44: D1172–80.
93. Yan R, Zhang Y, Li Y, Xia L, Guo Y, Zhou Q. Structural basis for the recognition of SARS-CoV-2 by full-length human ACE2. Science. 2020;367:1444–8.
94. Ware D, Jaiswal P, Ni J, et al. Gramene: a resource for comparative grass genomics. Nucleic Acids Res. 2002;30:103–5.
95. Sakai H, Lee SS, Tanaka T, et al. Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol. 2013;54:e6.
96. Lazzari B, Caprera A, Vecchietti A, Stella A, Milanesi L, Pozzi C. ESTree db: a tool for peach functional genomics. BMC Bioinform. 2005;6:1–6.
97. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H. SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res. 2011;39: e132.
98. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L. VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics. 2009;25:2283–5.
99. Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, Johnson J, Dougherty B, Barrett JC, Dry JR. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic Acids Res. 2016;44:e108.
100. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907; 2012.
101. Smith SM, Maughan PJ. SNP genotyping using KASPar assays. Methods Mol Biol. 2015;1245:243–56.
102. Wu MC, Kuan P-F. A guide to illumina BeadChip data analysis. In: Tost J, editor. DNA methylation protocols. New York, NY: Springer; 2018. p. 303–30.
103. Li F, Kitashiba H, Inaba K, Nishio T. A Brassica rapa linkage map of EST-based SNP markers for identification of candidate genes controlling flowering time and leaf morphological traits. DNA Res. 2009;16:311–23.
104. Kruglyak L. The use of a genetic map of biallelic markers in linkage studies. Nat Genet. 1997;17(1):21–4. https://doi.org/10.1038/ng0997-21.
105. Nakanishi N, Wada T, Arikawa K, Millet J, Rastogi N, Iwamoto T. Evolutionary robust SNPs reveal the misclassification of Mycobacterium tuberculosis Beijing family strains into sublineages. Infect Genet Evol. 2013;16:174–7.
106. Rai AJ, Yee J, Fleisher M. Biomarkers in the era of personalized medicine—a multiplexed SNP assay using capillary electrophoresis for assessing drug metabolism capacity. Scand J Clin Lab Invest. 2010;70:15–8.
107. Huang X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z, Li M. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42:961.
108. Lekklar C, Pongpanich M, Suriya-arunroj D, Chinpongpanich A, Tsai H, Comai L, Chadchawan S, Buaboocha T. Genome-wide association study for salinity tolerance at the flowering stage in a panel of rice accessions from Thailand. BMC Genomics. 2019;20:76.
109. Huq MA, Akter S, Nou IS, Kim HT, Jung YJ, Kang KK. Identification of functional SNPs in genes and their effects on plant phenotypes. J Plant Biotechnol. 2016;43:1–11.
110. Wendel JF. Genome evolution in polyploids. In: Doyle JJ, Gaut BS, editors. Plant molecular evolution. Dordrecht: Springer Netherlands; 2000. p. 225–49.

111. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML. Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet. 2011;12:499–510.
112. Kim K-T, Lee HW, Lee H-O, Song HJ, Shin S, Kim H, Shin Y, Nam D-H, Jeong BC, Kirsch DG. Application of single-cell RNA sequencing in optimizing a combinatorial therapeutic strategy in metastatic renal cell carcinoma. Genome Biol. 2016;17:1–17.
113. Andrews KR, Good JM, Miller MR, Luikart G, Hohenlohe PA. Harnessing the power of RADseq for ecological and evolutionary genomics. Nat Rev Genet. 2016;17:81.
114. Ma J-Q, Huang L, Ma C-L, Jin J-Q, Li C-F, Wang R-K, Zheng H-K, Yao M-Z, Chen L. Large-scale SNP discovery and genotyping for constructing a high-density genetic map of tea plant using specific-locus amplified fragment sequencing (SLAF-seq). PLoS One. 2015;10: e0128798.

Microsatellite Markers from Whole Genome and Transcriptomic Sequences 18

Manoj Kumar Gupta, Ravindra Donde, S. Sabarinathan, Gayatri Gouda, Goutam Kumar Dash, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, Priyadarsini Sanghamitra, C. Parameswaran, and Lambodar Behera

Abstract

Microsatellites (MS) or simple sequence repeats (SSRs) is a DNA sequence set comprising of tandemly repeated motifs. SSRs with codominant inheritance, higher amounts, moderately conservative flanking sequences, and rich polymorphism are commonly distributed throughout the plants and animals' genome. MS has already been employed in several crop plants for determining their seed lots' genetic integrity and to evaluate the capacity of plant varieties to defend their intellectual property. Thus, the key objective of this chapter is to include a revised and comprehensive overview of the SSR marker and its applications in various biological domains. Additionally, we have discussed genomic occurrence and the advantage/disadvantages of employing microsatellites as genetic markers in agricultural research.

M. K. Gupta · R. Donde · G. Gouda · G. K. Dash · P. Sanghamitra · C. Parameswaran ·
L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_18

Keywords

Microsatellites · Simple sequence repeats · Biological functions · Plants

Abbreviations

ESTs	Expressed sequence tags
GNMS	Genic non-coding microsatellite
InDel	Insertion-deletion
MAB	Marker-assisted backcrossing
MAS	Marker-assisted selection
MFRs	Microsatellite flanking regions
MS	Microsatellites
NGS	Next-generation sequencing
QTL	Quantitative trait loci
SNP	Single nucleotide polymorphism
SSRs	Simple sequence repeats (SSRs)

18.1 Introduction

DNA sequences set comprising of tandemly repeated motifs is known as microsatellites (MS) or simple sequence repeats (SSRs) [1, 2]. SSRs with codominant inheritance, higher amounts, moderately conservative flanking sequences, and rich polymorphism are commonly distributed throughout the plants and animals' genome. The SSRs polymorphisms, generated from motif repeat times, can be quickly identified by amplification of PCR [3]. SSR markers are thus known to be a class of molecular markers that are cost-effective, accurate, and strongly polymorphic and are commonly used in genetic and breeding research [4]. Even today, while molecular markers like single nucleotide polymorphism (SNP) or insertion-deletion (InDel) are rapidly emerging through third-generation sequencing, SSR markers are still important in molecular breeding, crop genetic analysis, and quantitative trait loci (QTL) mapping. From sequences generated from transcriptome, genome, and expressed sequence tags (EST), corresponding g-SSR, EST-SSR, and transcriptome-SSR markers are developed, respectively [5]. The g-SSRs displayed a higher polymorphism rate relative to the EST-SSRs and transcriptome-SSRs [6, 7]. In addition, the majority of the g-SSRs derive from the chromosome's known physical position and provide more precise information, particularly in gene/QTL mapping. SSR has already been employed in several crop plants for determining their seed lots' genetic integrity [8], and to evaluate the capacity of plant varieties to defend their intellectual property [9]. These markers are often used to a large degree to determine genetic variation and associations between species and lines, as well as for distinguishing crop varieties. The benefits of SSRs over SNPs are that SSRs support relative ease of transition amongst nearly related species [10, 11] and strong allelic diversity [12, 13].

However, SSRs have certain limitations in comparison to SNPs: the production process for multi-locus assays is very long and costly, and the throughput is comparatively poor. Recently, to overcome these issues, progress has been made in the synthesis of multi-locus assays in many ways, indicating that SSR markers remain, at least for basic applications as well as genetic studies, as important molecular tools [14]. In reality, SSR genotyping based on PCR has rapidly evolved within plants. Techniques have been established for the concurrent multiple marker loci amplification augmented with semi-automated identification systems [15]. Due to the advent of next-generation sequencing (NGS) technologies, the detection and discovery of SSR markers have also become cheaper and quicker. In addition, it has become much simpler to multiplex complex combinations of microsatellite markers. The existence of capillary electrophoresis device dependent on computerized "laser-induced fluorescence DNA technology" has also encouraged the implementation as well as usage of this approach within relevant breeding programs [16–18]. For classifying as well as identifying a cultivated variety and for establishing its uniformity, stability and differentiation, and genotypic features, SSR analysis is an appropriate molecular method accessible to all organisms. They are capable of assisting phenotypic reflection (DUS testing) [18–20]. Thus, the key objective of this chapter is to include a revised and comprehensive overview of the SSR marker and its applications in various biological domains.

18.2 Microsatellites: Definition and Genomic Occurrence

Tandem repeats with very small nucleotide motifs, such as TCCTCCTCCTCC, are known as SSRs or microsatellite. The microsatellite repeats may vary from two (AG) 2 to a few dozen (ACAT)11 nucleotides, while several dozens of repeated motifs are known as minisatellites [21]. In eukaryotes and prokaryotes, microsatellites are largely distributed across the genome, specifically within eukaryotes' nuclear euchromatin and organellar DNA coding as well as a non-coding region [22, 23]. To date, numerous studies have been conducted to support that SSRs are not spread around the genome at random. In 2006, Lawson and Zhang showed that SSR distribution is extremely non-random employing comparative analysis, and it varies distinctly in various regions of the rice and *Arabidopsis thaliana* genome [24]. In the major cereals, researchers have also tended to categorize microsatellites based on various parameters. SSRs were divided into two groups within the barley and Avena species: those with distinct sequences on either flank and those deeply entwined with retrotransposons and other scattered repetitive components. In oat cultivars, the second form shows less polymorphism [25, 26]. In 2001, Temnykh and the team classified microsatellites dependent on length employing publicly accessible DNA sequence data of the rice genome and found that larger perfect repeats ($\geq$20 nucleotides) were strongly polymorphic [27]. It has been discovered that microsatellites with SSRs $<$12 bp have a mutation capacity no diverse from most of the distinct sequences. In addition, researchers recorded that approximately 80% of GC-rich trinucleotides appeared within exons, while AT-rich trinucleotides were nearly equally spread among all genomic components (untranslated regions, coding sequences, intergenic and introns spaces). The tetranucleotide SSRs are largely found within the rice genome's non-coding regions, primarily intergenic regions. Later it was detected that the SSR distributions were non-random in the maize genome's various regions, and untranslated regions (UTR) have the highest density, eventually dropping off within the regions of the promoter, intergenic, intron, and coding region [28].

Correlations of microsatellite distributions in the chromosomes of *Silene latifolia* and *Rumex acetosa*, on the other hand, revealed that certain motifs (e.g., TAA or CAA) in the sex chromosome's (Y) non-recombinant regions are highly accumulated in both plant organisms [29]. Likewise, in a group of fish species (*Leporinus* spp.) that share the ZW sex system, a rather broad accumulation comprising primarily of microsatellites on the heterochromatic W chromosome was recorded, suggesting an interconnectedness amongst heterochromatinization and the repetitive sequences accumulation that has been suggested as the basis for the sex chromosome evolution [30]. Commonly SSRs experience a large mutation rate that is associated with gene expression, and thus it may be claimed that the incidence of SSRs is lowest within gene regions. Studies suggest that there is an SSRs preponderance with tri- and hexanucleotide gene motifs within coding areas, the product of selection pressure against mutations that change the reading frame [31, 32]. In humans, the consensus is that SSRs can often originate within coding areas, thereby contributing to the presence within protein sequences having repeated patterns. Tandem repeats in several proteins have been documented in protein sequence database analyses, and the underlying mechanisms within their genesis may lead to the rapid evolution of proteins [33, 34].

Repeat polymorphisms in the SSRs are typically the product of the complete repeat units or motifs being inserted or omitted. Therefore, various individuals show changes in repeat numbers as discrepancies. In other words, the polymorphisms found within SSRs are the variations in the template number repeats induced by DNA replication or recombination errors owing to polymerase strand-slippage. Strand-slippage replication is a defect of DNA replication in which there is a mismatch between the "template and nascent" strands. That implies that the strand of the template will loop out, thereby causing contraction. The "nascent" strand may also loop out, contributing to the extension being repeated. In addition, recombination events, like uneven crossing over as well as gene transfer, can lead to expansions and contractions of the SSR sequence. The purer and longer the repeat, the higher the risk of mutation, while shorter repeats of lower purity experience lower mutation frequency. New alleles at SSR loci are produced by mutations which have evaded correction through the DNA mismatch repair mechanism. For this cause, at a given SSR locus, there might be multiple alleles, which suggests that SSRs are more descriptive in comparison to other molecular markers, like SNPs. As for their arrangement, SSRs may be classified by motif as: "(a) perfect if composed entirely of repeats of a single motif; (b) imperfect if a base pair not belonging to the motif occurs between repeats; (c) interrupted if a sequence of a few base pairs is inserted into the motif; or (d) composite if formed by multiple, adjacent, repetitive motifs" [35, 36].

18.3 Advantage and Disadvantages of Microsatellites as Genetic Markers

The key benefits of microsatellite markers are codominant (heterozygotes may be differentiated from homozygotes) transmission, locus-specific in nature, strongly polymorphic and hypervariable, strong pattern and knowledge quality, the relative abundance of standardized genome distribution, higher-than-average mutation rate, and simple preparation for sampling. Potential benefits of microsatellites like short size range, continuous stretches of similar repeat units, strong polymorphism proportion, insights gained in recognizing the mutational mechanism that helps to improve statistical interpopulation comparison procedures, their prevalence in fish and other organisms' genomes, and the existence of microsatellite cloning methodologies have all contributed to the creation of statistical interpopulation comparison procedures. Microsatellites of tetranucleotides are often very effective for paternity and human forensic examinations. New technologies such as digital recording and automatic identification and scoring systems like automated DNA sequences and fluorescent imaging instruments have benefited from the beneficial properties of microsatellites [37].

Microsatellite drawbacks include the existence of shadow or stutter bands, the occurrence of null alleles (existing alleles not detected using normal assays), homoplastic, and too many alleles at some positions that would need very large sample sizes for analysis [38]. "Microsatellite flanking regions" (MFRs) often include lengthy mutants that can generate similar lengthy variants that could undermine studies of microsatellite population level (and comparisons of variance levels for homologous loci across species) and phylogenetic inferences as these lengthy variants in the flanking regions may potentially reduce the variation of allele duration within the repeat region [39].

18.4 Biological Functions of SSRs

The introduction of repetitive sequences into genomes of eukaryotes can confer an evolutionary adaptability advantage to novel environments [40, 41]. There have been well-documented discussions on the SSRs' functional role(s) within species adaptation as well as survival [42, 43]. The effects of contraction and expansion of the SSR motifs inside genes, however, have stimulated the assignment to SSRs' biological function. To date, the loci associated with fragile-X and Huntington's disease [44] are the best-reported examples of SSRs with phenotypic effects. SSRs can also be active in controlling the expression of neighbouring genes in the UTR areas, as shown by the GT repeat within the "Tilapia prolactin 1" gene in fish in regard to the salt-related conditions [45]. By affecting mRNA splicing and its translocation to cytoplasm, intronic SSRs can control gene expression, as seen by the CCTG repeat within the first intron of "zinc finger protein 9" (ZNF9) of human, in which repeat extension induces failure of one intron splicing, which in turn results in myotonic dystrophy [46]. While biological roles have not yet been reported for SSRs in plants, related roles are predicted in plant genes for these biomolecular markers.

There is always a debate on whether certain SSRs functionally have an ecological value and are they appropriate for studying biodiversity and environmental protection of endangered species. Few researchers have suggested that many of the molecular markers used in population genetics did not undergo filtering and were thus largely neutral. The frequency of alleles in neutral theory is calculated by strictly stochastic processes [47]. Neutral molecular markers can be useful in conservation biology to provide fundamental knowledge regarding community mating forms, gene distribution, and the population background of a species [48]. However, as assessed by neutral RAPD markers, there was a substantial difference between genetic divergence and that evaluated through the monkey puzzle tree's quantitative genetic traits (*Araucaria araucaucana*) (a fragile tree that is endemic to the southern region of South America) [49]. Tienderen and the team [50] suggested that gene-targeted functional markers can lead to ecological diversity studies, endangered species protection, and ex situ genetic resource management. Holderegger and the team [51] included an "adaptive versus neutral diversity" hypothesis for studying landscape genetics in which the diversity evaluated through neutral markers is better adapted to the gene flow processes analysis within ecosystems, while the diversity tested through quantitative genetic tests utilizing functional markers is ideally suited to the assessment of the evolutionary or adaptive ability of a species. They suggested that these variations amongst adaptive and neutral genetic variance must be considered via ecologists when elucidating the landscape genetic studies' outcome. It is pertinent to note that, nevertheless, that variance within functional genes may depict the selection's previous effect, which may vary in every gene and may influence the past, migration, and drift variation profiles [51]. Since genomic SSR markers are largely neutral, certain adaptive functions may be maintained by genetic SSRs from ESTs or cDNAs. This duality of selection and adaptation gives the usage of SSRs another benefit in characterizing the genetic variability of the capital held in various institutes of germplasm.

18.5 Development of SSR Markers

Fully sequenced genomes aid in the identification as well as the development of the huge amount of microsatellite gene-based markers. Rice (*Oryza sativa* L.), for example, is the first cereal whose genome has been completely sequenced, which in turn led to the development of a significant number of microsatellite markers [52]. Polymorphic 52,485 microsatellite markers amongst japonica and *indica* were recently developed by Zhang et al. [53]. The main problem, however, lies in detecting the most appropriate as well as insightful microsatellite markers from that much large dataset for genotyping rice. By building smaller as well as detailed microsatellite marker databases comprising of markers situated in potentially usable gene sequences having higher polymorphic potential, this problem can be overcome. Parida and the team [54] reported 19,555 perfect "genic non-coding microsatellite (GNMS)" repeats on chromosomes 12 and 1 within the rice, taking into account the excellent genetic qualities as well as higher expected informativity of GNMS markers. With the entire genome of rice sequenced now, within each genes' few thousand base pairs, microsatellite markers can be created. Research by Goff et al. [55], for example, indicates that there is a mean of one microsatellite repetition (demarcated as at ~8 repetitions of a 2–4 bp motif) per 8 kb, thereby generating nearly 48,351 markers within the whole genome [56].

Saarinen and Austin [57] have designed PCR primers to amplify microsatellite markers employing the online program, namely, Primer3 through subjectively selecting primers "flanking the repeat regions". The primers were subsequently retrieved from numerous industrial manufacturers and checked for their capacity for amplifying the microsatellites as well as distinguish polymorphisms amongst the parental lines employed in the mapping studies of Pi-z. The RM6836 and RM527 markers, which were earlier situated nearby the Pi-z locus [58], were retrieved from the gramene (http://www.gramene.org/) database [59] and employed in the primary tests [60]. Different NGS technologies, like GS FLX, Roche 454, and HighSSR, are already being employed for microsatellite exploration, with substantial cost as well as less time [57, 61–64]. Rapid developments of NGS technologies have lowered costs significantly while enhancing throughput and precision exponentially. At present, the IlluminaHiSeq2000 [65] is the most cost-effective NGS tool, which can minimize costs by 3400 times compared to conventional sequencing methods; it is fair to believe that continuing advances would result in still lower costs [66].

Earlier DNA markers have been designed by Fjellstrom and the team [67] employing four distinct approaches. Five of these markers (RM155, RM138, RM101, RM166 and RM144) are focused on a previous microsatellites group found at the University of Texas A&M through scanning repetitive sequences from the publicly available DNA sequence database of the NCBI, as defined in Temnykh et al. [68]. Through mapping these genes in several populations identified by Conaway et al., the genomic positions of Pi-k, Pi-ta2, and Pi-b and their actual linkages have been detected [69]. Three additional closely related markers have been subsequently defined through mapping these initial markers comparative to microsatellite markers produced employing conventional approaches at Cornell University [68]: RM266, RM224, and RM208. The authors mapped these markers nearby the "blast resistance genes", namely, *Pi-k, Pi-ta2,* and *Pi-b*, on rice chromosomes 11, 12, and 2, respectively, after discovering candidate microsatellite markers employing various publicly available database tools. Subsequent to the public availability of the "Monsanto rice microsatellite database", two additional associated microsatellite markers, namely RM7102 and RM1233, were released [70]. Though several DNA markers have been produced for rice blasting resistance, the majority are not appropriate for regular usage within the MAS program comprising of large progeny number. Pibdom, a dominant *Pi-b* gene marker, has also been established dependent on the cloned *Pi-b* gene sequences [71] (GenBank accession AB013448). The introgression, as well as the pyramidization of these three blast tolerance genes into new rice cultivars as well as elite lines [67], should be encouraged by these markers. In addition, the "International Rice Microsatellite Initiative" (IRMI) has established "a high-density microsatellite map with a genome coverage of approximately one microsatellite per 0.5 cM" [70], which can be used to establish closely connected markers for a number of agronomic traits, like blast resistance. The freely accessible full genomic sequences accessibility of the rice subspecies japonica and indica (http://www.genomics.org.cn; http://rgp.dna.affrc.go.jp;) has allowed rice researchers for producing additional markers for high-quality mapping of targeted genes. Novel SSR, CAPS, and InDel markers [72] have been developed for acquiring a "high-density linkage map" to generate "fine-scale mapping" within their target area, utilizing the publicly accessible genome sequence of the rice (http://rgp.dna.affrc.go.jp).

18.6 Applications of Microsatellites

Because of its hypervariable existence and broad genome coverage, microsatellites have been a marker of choice for large number of plant species. In a germplasm set, these are employed for testing genetic variance at the molecular level to allow sufficient selection of parents for gene mapping, crosses (i.e., hybrid breeding), and QTLs for agronomic as well as disease resistance characteristics, genome mapping, MAS and "marker assisted backcrossing" (MAB) during breeding projects, and gender identification. Diversity study, genetic similarity measurements or variations between plant species are valuable knowledge for the conservation of crops and varietal growth [73]. In addition, the knowledge is often helpful in characterizing plant germplasm collections and for taxonomic studies. Because of codominant existence, the high degree of polymorphism, high quality, as well as reproducibility, microsatellite markers have proven to be a powerful instrument for estimating genetic diversity and phylogenetic relationships of organisms dependent on sequence conservation in recent years. In cultivar detection, microsatellites are beneficial and are often advantageous in pedigree research since they display a single locus. These multiallelic markers enables accurate comparative detection of allelic heterogeneity through a broad variety of germplasm [74]. The determination of hybridity is another significant application of microsatellites, in which the codominant structure of microsatellites plays an important role and permit for the allelic contribution of each parent to be established in sexual as well as somatic hybrids [75].

The female plants are economically prized for the development of fruits (papaya, seabuckthorn, kiwi fruit, dates, etc.) and seeds. A select group of flowering plants are sexually dimorphic (pistachio, nutmeg, black pepper, etc.). However, the most dioecious plants' sex is not morphologically revealed, and at the seedling stage, the female and male plants cannot be differentiated. In species where an organism's sex is only disclosed subsequent to flowering, may take many months (papaya, Coccinia) to many years, this issue can even be more complex in few plants (nutmeg, date palm, and jojoba). In many species, sex-related microsatellite markers have been identified, like hemp [76], *Actinidia chinensis* [77], hop [78], wild strawberry [79], *Carica papaya* [80]. Parasnis and the team (1999) employed a microsatellite probe (GATA)4 in papaya as a diagnostic biomarker and revealed the sex-specific DNA variation at every point of plant development. Recently, utilizing 644 microsatellite markers, Fraser et al. [77] have developed gene-rich male, female as well as consensus linkage maps of the diploid species namely *A. chinensis*. They established genetic linkage maps identifying the haploid genome's 29 linkage classes, unveiled the location and scale of the locus that decides sex, and also defined putative X and Y chromosomes by sex-linked markers.

The creation of unique organelle markers (i.e., mtSSR and cpSSR) had a significant influence on the identification of structure and diversity, as well as phylogenetic relationships, within a natural population. The uniparental inheritance mode retained gene order as well as absence of heteroplasmia and recombination of organelle genomes that render them an attractive instrument for evolutionary studies, especially levels of differentiation, migration trends, as well as population histories [81]. ESTs, though, are often employed for such research when one specifically aims at the development of functioning genes in such experiments [74]. Genetic complexity assessment and phylogenetic associations have culminated in the discovery of certain reclassified misclassified accessions. The examination of genetic diversity and the creation of phylogenetic relationships would provide valuable knowledge for the collection of parental lines for carrying out breeding studies, accessions classification to plant germplasm, and further curation as well as the acquisition of new accessions to plant germplasm [82].

In identifying particular genomic regions associated with the development of essential agronomic and physiological features, microsatellite markers have been efficiently employed. In addition, microsatellite markers may also be employed for evaluating QTLs associated with the detection of candidate interest trait genes that are especially important for a breeding program such as yield, disease resistance, consistency of seed and fruit, and resilience to stress [73, 83]. Nevertheless, in comparison to genomic library microsatellite markers, EST-SSRs may lead to the direct selection of alleles since they have putative or established roles and may be correlated with targeted phenotype functions [84]. Association mapping relating to a substantial molecular marker associated with a phenotypic feature is particularly helpful for the application of marker-assisted quantitative characteristics selection in plant breeding programs [85]. QTL mapping typically employs a bi-parental cross-community, while association mapping employs a set of people of varying ethnicity. Genetic maps of many plant species have been prepared in recent years, including corn, barley, potato, wheat, sorghum, cotton, white clover, ryegrass, and raspberry. Microsatellite markers, once mapped, may be used to tag some individual traits that are especially important for a breeding program. In several significant crop species such as potatoes, wheat, corn, and soybeans, association mapping using SSR markers has been successfully carried out [82]. In potato cultivars, the interaction between the microsatellite marker and QTL for resistance to *Verticillium dahliae* has been established, which in turn contributes to QTL cloning for resistance to *V. dahliaee* [86]. The link amongst SSR markers and the wheat kernel size was established using elite germplasm interaction mapping [85].

In addition, for MAS, a wide number of monogenic as well as polygenic loci for different characteristics may be defined and used [74]. MAS will enable breeders to circumvent conventional choices focused on phenotypes in the region, thus speeding up breeding programs. In three backcross generations, the rice variety Swarna could be effectively transformed into a submergence-resistant variety, requiring a duration of 2–3 years utilizing marker-aided backcrossing [83]. Employing "~ 9,892 subtracted drought stress, ESTs of sorghum" accessible in the NCBI dbEST database, Srinivas et al. [87] investigated microsatellite loci and proposed that it could be relevant for drought stress in QTL research. In another study, ~20,162 salinity-responsive as well as drought-ESTs from 10 separate root tissue cDNA libraries of chickpea were created through Varshney and the team [88]. The created collection of ESTs functions as a resource for high-quality gene discovery transcripts as well as the production of functional markers related with tolerance of abiotic stress. It is also possible to merge transgenic methods with MAS for the production of insect- as well as disease-resistant cultivars.

Another area where microsatellites are being employed widely is genome mapping. The mapping of genomes includes physical mapping, genetic mapping, association mapping, and comparative mapping. Microsatellite marker genetic mapping in plants was first recorded in tropical trees and subsequently reported in rice, soybeans, etc. Over 80 genetic maps have been built thus far, through employing SSR markers from several plant organisms. In several plant types, including the legumes, Solanaceae family, crucifers, and grasses, comparative mapping has been successfully carried out [82]. There is a strong opportunity for comparative genomics of relatives of Arabidopsis to enhance our knowledge of the molecular structure and evolutionary processes. A significant context for comparative genomics analysis is provided by recent analyses of phylogenetic relationships within Brassicaceae. An ancestral karyotype of these species has been concluded through comparative linkage mapping as well as chromosome painting in the near *Arabidopsis* relatives. Furthermore, Brassica's comparative mapping established genomic blocks that have been preserved since the Arabidopsis and Brassica lineages diverged [89]. For comparative visualization, microsatellite markers amongst *Castanea sativa* (Mill.) and *Quercus robur* (L.) have been employed [90]. In rice, wheat, barley, and rye, EST-SSR markers have been employed for comparative visualization. The conservative chromosome areas amongst wheat and rice and the occurrence of barley EST-SSR orthologues in various organisms have been reported as well as recognized [84, 91, 92]. For the creation of whole-genome physical maps of model crop organisms, SSR markers have also been employed. SSR markers have been employed for anchoring and evaluating the physical and genetic soybean map frames [93, 94]. Employing BAC end sequences and SSR markers, a ~ 2 Mb BAC contig's physical map was built in the region of ~80 cM of chromosome 2 of *Arabidopsis thaliana* [95].

Because of their cost-efficiency and their application in broad-scale genotyping, the emergence of novel technology has not influenced the utility of microsatellites. Microsatellite-based markers, including varietal and cultivar discrimination, map-based cloning, marker-assisted breeding, and gene flow studies, are accurate and simple-to-use instruments for fingerprinting applications. In view of the presence of multiple main and secondary gene pools in several plants' types, the future of microsatellite-based markers is encouraging. These gene pools' categorization is an enormous challenge, and the only inexpensive, reproducible, and effective approach that will be capable for these characterization studies is microsatellite-dependent markers.

18.7 Software for Microsatellite Development

To date several softwares have been developed for microsatellite development (Table 18.1) and its analysis (Table 18.2). Few have been described below.

18.7.1 Geneious

Geneious is a desktop applications package for sequence knowledge organization and interpretation in molecular biology [96]. In order to satisfy the particular needs of users, microsatellite architecture includes many plugins (e.g., MISA, Phobos, and Primer3). It is a commercial software that involves the procurement of an activation certificate, boosting the budget for testing. Phobos, which can be operated free of charge separately from Geneious, is the part that searches for microsatellite loci. Phobos has both command-line and GUI interfaces, and it easily processes massive data. In less than an hour on a regular laptop, every data set checked completed the quest ("2.5-GHz Intel Core i5, 8 GB RAM"). Phobos may not communicate specifically with Primer3. However, the effects of the position scan in Phobos will quickly be piped to Primer3 if Phobos is used by Geneious. Phobos is quick and user-friendly for the production of microsatellite locations.

18.7.2 GMATo

GMATo comes with a graphical Java GUI and is available for use shortly after execution [97]. The findings of GMATo are provided as a table of statistics for SSR loci. It works fast; it finished the job within 52 min on a Windows desktop computer for the "HiSeq2 data set (a 5.7-GB file) (Eight Core 3.4-GHz Intel Core i7-2600 CPU, 16 GB RAM)". However, the distribution of repeat number motifs cannot be regulated by the individual, and any repeat duration must be set to the same amount. This software is incapable of generating marker generation, primer configuration, or electronic mapping markers.

Table 18.1 Software's/tools used in microsatellite development

Sl. no.	Software's/tools	Description	Software interface	Language	Link	Operating system	Cost
1	Staden Package	A set of DNA sequence assembly, editing and analysis tools (automated microsatellite marker development)	Web server	Perl	http://staden.sourceforge.net/	Linux, Mac OS X, Microsoft Windows	Free
2	RepeatMasker	For interspersed repeats and low complexity DNA sequences	Command line interface	Perl, Python	http://www.repeatmasker.org/	Linux, Mac OS X, Microsoft Windows	Free
3	Troll	Tandem Repeat Occurrence Locator (SSR finder based on Aho-Corasick algorithm)	Command line interface	Matlab	http://finder.sourceforge.net/	Linux	Free
4	WebSat	For microsatellite marker development (visualization of microsatellites and the design of primers)	Web server	PHP and JavaScript	https://bioinfo.inf.ufg.br/websat/	Linux, Mac OS X, Microsoft Windows	Free
5	msatcommander	Detection of microsatellite repeat arrays and automated, locus-specific primer design	Graphical interface	Python	https://github.com/brantfaircloth/msatcommander	Linux, Microsoft Windows	Free
6	Primer3	Design PCR primers from DNA sequence (widely used)	Web-based, command-line	Perl, C	http://primer3.org/	Linux, Mac OS X, Microsoft Windows	Free
7	RISA	Rapid Identification of SSRs and Analysis of primers	Web server	Perl	http://sol.kribb.re.kr/RISA/	Linux	Free
8	SSR Locator	Discovery of Integrated with Primer Design and PCR Simulation	Web server, graphic interface	Perl	http://microsatellite.org/ssr.php	Linux, Mac OS X, Microsoft Windows	Free

9	ESAP plus	For EST-SSR marker development (Analysis Pipeline Plus) clustering, pre-processing, assembly, SSR mining and SSR primer design	Web-based server	Perl and bash, MySQL server, HTML, CSS, PHP, Java	http://gbp.kku.ac.th/esap_plus/index.php	Linux	Free
10	FullSSR	SSR detection and primer designing software	Command-line interface	Perl	https://sourceforge.net/projects/fullssr/	Linux	Free
11	BatchPrimer3	For PCR and sequencing primer design	Web server	HTML	https://wheat.pw.usda.gov/demos/BatchPrimer3/	Linux, Mac OS X, Microsoft Windows	Free
12	PolyMorphPredict	For Rapid Polymorphic Microsatellite Marker Discovery from Transcriptome Data and Whole Genome	Web server	Perl, R, Java	http://webtom.cabgrid.res.in/polypred/	Linux, Mac OS X, Microsoft Windows	Free
13	SAT	For SSR marker development (integration, analysis and display of sequence data)	Web server	Standalone command-line interface	http://sat.cirad.fr/sat	Linux, Mac OS X, Microsoft Windows	Free
14	SSRIT	Simple Sequence Repeat Identification Tool (to finds all perfect simple sequence repeats in a given sequence)	Web server	Stand-alone	https://archive.gramene.org/db/markers/ssrtool	Linux, Mac OS X, Microsoft Windows	Free
15	Microsatellite repeats finder	Finds microsatellites in DNA sequences (to generate the microtatellites database)	Web server	–	http://insilico.ehu.es/mini_tools/microsatellites/	Linux, Mac OS X, Microsoft Windows	Free
16	SciRoKo	For whole genome microsatellite investigation and search	Web server	Perl	https://www.kofler.or.at/digitalbase/index.html	Linux, Mac OS X, Microsoft Windows	Free

(continued)

Table 18.1 (continued)

Sl. no.	Software's/tools	Description	Software interface	Language	Link	Operating system	Cost
17	PolySSR	To identify Candidate Polymorphic SSRs based on Multiple Assembled Sequences	User interface	MySQL	http://www.plantkingdomgdb.com/CandiSSR/index.html	Linux, Mac OS X, Microsoft Windows	Free
18	MISA	Microsatellite prediction tool (detects, interrupted and compound SSRs)	Web server	PHP, Perl	https://webblast.ipk-gatersleben.de/misa/	Linux, Mac OS X, Microsoft Windows	Free
19	Tandem Repeats Finder	Database of tandem repeats to run sequences (to locate and display tandem repeats in DNA sequences)	Web server	–	https://tandem.bu.edu/trf/trf.html	Linux	Free
20	Imperfect SSR Finder	To help geneticists find SSR, aka microsatellites/Short Tandem Repeats	Web server	Perl	https://ssr.nwisrl.ars.usda.gov/	Linux, Mac OS X, Microsoft Windows	Free

DNA deoxyribonucleic acid, *PCR* polymerase chain reaction

Table 18.2 Software and tools used for analyzing microsatellite

Sl. no.	Software's/ tools	Description	Software interface	Language	Link	Operating system	Cost
1	Darwin	Dissimilarity analysis and representation (diversity and phylogenetic analysis)	User interface	Microsoft™ Visual Basic Studio.Net 2010	https://darwin.cirad.fr/	Windows	Free
2	Structure	Multi-locus genotype data to investigate population structure	User interface	–	https://web.stanford.edu/group/pritchardlab/structure.html	Linux, Mac OS X, Microsoft Windows	Free
3	Power marker	For genetic marker data analysis (summary statistics, population structure, phylogenetic analysis, association study and design)	Graphical user interface	–	https://brcwebportal.cos.ncsu.edu/powermarker/	Windows	Free
4	GenAlex	Population genetic analysis tool (AMOVA, PCoA, diversity and Frequency-Based Population Genetic Analysis)	User interface (MS Excel add-in software)	–	https://biology-assets.anu.edu.au/GenAlEx/Welcome.html	Windows	Free
5	Tassel	To evaluate traits associations, evolutionary patterns, and linkage disequilibrium	User interface	–	https://www.maizegenetics.net/tassel	Linux, Mac OS X, Microsoft Windows	Free
6	NTSYSpc	Numerical Taxonomy and Multivariate Analysis System (to discover pattern and structure in multivariate data) cluster analysis, ordination analyses and biplots, principal components analysis, principal coordinates analysis, PCOORDA, nonmetric	User interface	–	http://www.appliedbiostat.com/ntsyspc/ntsyspc.html	Windows	Free and paid

(continued)

Table 18.2 (continued)

Sl. no.	Software's/ tools	Description	Software interface	Language	Link	Operating system	Cost
		multidimensional scaling, Burnaby's method for size adjustment, analysis of shape using landmark coordinates and comparison of dis/similarity matrices					
7	GGT	For Visualization and Analysis of Genetic Data (marker trait associations, calculation of genetic distances, marker linkage disequilibrium)	User interface, graphical user interface	–	https://www.plantbreeding.wur.nl/#home	Windows	Free
8	POPGENE	For Population Genetic Analysis (computes both comprehensive and complex genetic statistics)	User interface	–	https://sites.ualberta.ca/~fyeh/index.html	Windows	Free
9	Qtl cartographer	A Program Package for finding Quantitative Trait Loci (single-marker analysis, interval mapping, composite interval mapping, Bayesian interval mapping, multiple interval mapping, multiple trait analysis, multiple trait MIM analysis)	User interface, Graphical user interface	LaTeX2e, Perl	https://brcwebportal.cos.ncsu.edu/qtlcart/index.php	Windows, Linux	Free
10	QTL IciMapping	Building Genetic Linkage Maps and Mapping Quantitative Trait Genes	User interface	C#, Fortran 90/95	http://www.isbreeding.net/software/?type=detail&id=18	Paid	Free
11	Arlequin	For Population Genetics Data Analysis (intra-population and inter-population methods)	User interface	R	http://cmpg.unibe.ch/software/arlequin35/	Linux, Mac OS X,	Free

						Microsoft Windows	
12	GeneMapper	Provides DNA sizing and quality allele. Size, genotype, and process quality reports and genotype plots (autoanalysis)	Command line operation	–	https://www.thermofisher.com/order/catalog/product/4370784#/4370784	Windows	Free
13	GMATA	Genome-wide Microsatellite Analyzing Toward Application (for genomic SSR marker) SSR mining, statistical analysis and plotting, SSR loci graphic viewing, marker designing, electronic mapping and marker transferability investigation	User-friendly graphical and command line interfaces	Java, Perl, R	https://sourceforge.net/projects/gmata/	Linux, Mac OS X, Microsoft Windows	Free
14	KINGROUP	For pedigree relationship reconstruction and kin group assignments using genetic markers	User interface	Java	https://sourceforge.net/projects/kingroup/	Linux, Mac OS X, Microsoft Windows	Free
15	GENEPOP	Population genetics (null alleles, exact tests, algorithms for exact tests, accuracy of *P* values estimated by the Markov chain algorithms, test statistics, estimating *F*-statistics and related quantities, Bootstraps, Mantel test)	–	R, C++	https://genepop.curtin.edu.au/		
16	GENECLASS2	For Genetic Assignment and First-Generation Migrant Detection	User interface	–	http://www1.montpellier.inra.fr/CBGP/software/GeneClass/	Linux, Microsoft Windows	Free
17	FSTAT	To estimate and test population genetics parameters (gene diversities and *F*-statistics)	User interface	–	https://www2.unil.ch/popgen/softwares/fstat.htm	Microsoft Windows	Free

18.7.3 HighSSR

In the PCR primers for retrieved loci, HighSSR recognizes microsatellites and reduces redundancy [64]. Employing "Tandem Repeats Finder" (TRF; [98]), it detects as well as grades SSRs within raw sequencing reads and stores them in a PostgreSQL database, reporting summary information, for instance, the alleles number of each SSR locus that can be evaluated with other applications. HighSSR demultiplexes pooled libraries, evaluates polymorphism of the locus and applies Primer3 for the configuration of the primer. Finally, for refining crude clusters and extract loci from them, Muscle [99] is used. A "Java virtual machine", however, and database access on a PostgreSQL server are necessary. In addition, it allows it impossible to use non-universal parameter settings, separate Java codes as well as shell scripts. We could open just our smallest test data file for the TRF executable file (PacBio; 445 MB).

18.7.4 MISA

MISA is an abbreviation for the "MIcroSAtellite identification tool", which was originally developed to produce SSR loci from EST results [100]. When Perl is enabled and runs fast, it operates immediately; the 5.7-GB HiSeq2 data collection was completed in 1.8 h ("one node, one processor, and 4 GB of memory"). By modifying a configuration file ("misa.ini"), users are able to adjust the default configurations, and MISA is capable of building primers. The reports are in tabular form, providing a description of the numerous figures, like the occurrence of a particular sort of microsatellite. Some reports, however, show that MISA might have been redundantly mined in overlapping microsatellites [97, 101].

18.7.5 MSATCOMMANDER

MSATCOMMANDER facilitates quick and automatic identification of microsatellites, the locus-specific configuration of the primer and labelling [102]. It needs Python and provides output files in the format of a "comma-separated value" (CSV). The findings, however, are hard to see and do not contain general overview statistics on the types of microsatellite loci detected. In order to classify simple data, the user must expend significant time filtering the output file (e.g., the dinucleotide repeats number detected). It utilizes Primer3 as its engine for primer modelling and primer-tagging.

18.7.6 PAL_FINDER

Directly from raw NGS sequencing reads, PAL FINDER discovers microsatellite repeat elements and subsequently designs PCR primers for amplifying these repeat loci ("potentially amplifiable loci [PAL]") through association with Primer3 [103]. This is a command-line program that can be changed openly through the user via the configuration file required. Its efficiency, however, is highly sensitive to data coverage (quality and quantity of PALs; [103]). We were unable to get the FASTQ mode to function despite approximately 24 h of effort by modifying FASTQ input files. In "454" mode, we could use any form of FASTA file, even paired-end Illumina records, as long as all the readings were in a single file. Compared to the other software packages checked, this application has a slow run time (">24 h for data sets >4 GB on a regular [2.5-GHz Intel Core i5 with 8 GB RAM] laptop").

18.7.7 QDD3

QDD3 is made up of four individual operating modules with quality trimming, microsatellite identification, redundancy elimination, contamination management, primer architecture, and established transposable components comparative functions [104]. It can be employed both on the command line as well as through Galaxy [105] and operates via RepeatMasker and a number of other NGS software [106]. The runtime is reasonably long (for a 5.7 GB data collection, 9.5 h on a high-performance computer), and users are unable to adjust the default SSR search settings (e.g., specifying various repeats number for distinct length motifs).

18.7.8 SSR Locator

The SSR Locator combines SSR search features, motif frequency, primer architecture, and PCR simulation with other databases. This simulation enables the execution of global alignments as well as identity and homology searches between several amplified sequences [107]. Employing a GUI with an optimized menu system set, it performs all the module calls. It needs some file reformatting, though, which raises the computational time. For the HiSeq2 data collection, it took 10 min for the Windows framework to reformat and 69 min for the SSR quest ("8 Core 3.4-GHz Intel Core i7-2600 CPU, 16 GB RAM").

18.7.9 SSR_pipeline

The SSR pipeline is a command-line software for detecting high-throughput sequencing data microsatellites employing a Python environment [108]. With components for consistency filtering as well as alignment of Illumina raw data, it detects SSRs within paired-end reads of Illumina. Through utilizing the SSR detector module separately, SR-pipeline can also evaluate knowledge from other sequencing platforms, like 454 and Ion Torrent. Nevertheless, after 24 h of work by a biologist competent in bioinformatics, some researchers did not effectively run test data via the SSR pipeline.

18.7.10 STAMP

STAMP is a revised STADEN kit for microsatellite detection as well as primer design [109], with extensive Phobos incorporation [110] for tandem repeat recognition and evaluation. TROLL [111] is used by STAMP to trackback primer pairs to series trace directories, Primer3 to interactively build and visualize primers, and SQLite as a database to store the effects of research. Inclusive, STAMP is a highly versatile, high-throughput, interactive instrument for the design of traditional and multiplex microsatellite markers, preventing redundant markers from being produced. It is complex, however, and involves several tool command language modules and the STADEN kit pre-installation, and it is not appropriate for low-coverage NGS data [104].

18.8 Conclusion and Future Perspective

Since the identification of polymorphisms is a limiting factor in numerous breeding strategies, microsatellite markers serve as an invaluable tool for plant breeders as well as geneticists. Over the long term, in the science of rice breeding, the development of the allele-specific marker for genes controlling both abiotic and biotic resistance traits will become gradually important. The selection of the best suitable marker systems for a given program depends on numerous issues, including the technology platforms available, the cost of developing markers, the transferability of species, the content of information, and the ease of documentation. Moreover, additional resources for genomic analysis as well as breeding will be provided by a higher degree of genetic variability and the localization of more markers on the linkage map. There is therefore still scope for the development of more efficient breeding programs, which, in the future, will help us to develop novel biotic and abiotic resistant crop varieties.

Conflict of Interest None

References

1. ul Haq S, Jain R, Sharma M, Kachhwaha S, Kothari SL. Identification and characterization of microsatellites in expressed sequence tags and their cross transferability in different plants [Internet]. Int J Genomics. Hindawi; 2014 [cited 2021 Jan 27]. p. e863948. Available from: https://www.hindawi.com/journals/ijg/2014/863948/.
2. Gupta PK, Balyan HS, Sharma PC, Ramesh B. Microsatellites in plants: a new class of molecular markers. Curr Sci. 1996;70:45–54.
3. Varshney RK, Thiel T, Stein N, Langridge P, Graner A. In silico analysis on frequency and distribution of microsatellites in ESTs of some cereal species. Cell Mol Biol Lett. 2002;7:537–46.
4. Kalia RK, Rai MK, Kalia S, Singh R, Dhawan AK. Microsatellite markers: an overview of the recent progress in plants. Euphytica. 2011;177:309–34.

5. Zhao C, Qiu J, Agarwal G, Wang J, Ren X, Xia H, et al. Genome-wide discovery of microsatellite markers from diploid progenitor species, Arachis duranensis and A. ipaensis, and their application in cultivated peanut (A. hypogaea). Front Plant Sci. 2017;8:1209.
6. Balfourier F, Roussel V, Strelchenko P, Exbrayat-Vinson F, Sourdille P, Boutet G, et al. A worldwide bread wheat core collection arrayed in a 384-well plate. Theor Appl Genet. 2007;114:1265–75.
7. Han B, Wang C, Tang Z, Ren Y, Li Y, Zhang D, et al. Genome-wide analysis of microsatellite markers based on sequenced database in Chinese spring wheat (Triticum aestivum L.). PLoS One. 2015;10:e0141540.
8. Kumar MC, Vishwanath K, Shivakumar N, Prasad R, Radha S, Ramegowda BN. Utilization of SSR markers for seed purity testing in popular rice hybrids (Oryza sativa L.). Ann Plant Sci. 2012;1:1–5.
9. Ibaňez J, Van Eeuwijk FA, Spain H. Microsatellite profiles as a basis for intellectual property protection in grape. Acta Hortic. 2003;603:41–7.
10. Fan L, Zhang M-Y, Liu Q-Z, Li L-T, Song Y, Wang L-F, et al. Transferability of newly developed pear SSR markers to other Rosaceae species. Plant Mol Biol Rep. 2013;31:1271–82.
11. Satya P, Paswan PK, Ghosh S, Majumdar S, Ali N. Confamiliar transferability of simple sequence repeat (SSR) markers from cotton (Gossypium hirsutum L.) and jute (Corchorus olitorius L.) to twenty two Malvaceous species. 3 Biotech. 2016;6:65.
12. Emanuelli F, Lorenzi S, Grzeskowiak L, Catalano V, Stefanini M, Troggio M, et al. Genetic diversity and population structure assessed by SSR and SNP markers in a large germplasm collection of grape. BMC Plant Biol. 2013;13:1–17.
13. Filippi CV, Aguirre N, Rivas JG, Zubrzycki J, Puebla A, Cordes D, et al. Population structure and genetic diversity characterization of a sunflower association mapping population using SSR and SNP markers. BMC Plant Biol. 2015;15:1–12.
14. Guichoux E, Lagache L, Wagner S, Chaumeil P, Léger P, Lepais O, et al. Current trends in microsatellite genotyping. Mol Ecol Resour. 2011;11:591–611.
15. Masi P, Zeuli PS, Donini P. Development and analysis of multiplex microsatellite markers sets in common bean (Phaseolus vulgaris L.). Mol Breed. 2003;11:303–13.
16. Ganal MW, Röder MS. Microsatellite and SNP markers in wheat breeding. In: Varshney RK, Tuberosa R, editors. Genomics-assisted crop improvement. New York: Springer; 2007. p. 1–24.
17. Gonzaga ZJ, Aslam K, Septiningsih EM, Collard BC. Evaluation of SSR and SNP markers for molecular breeding in rice. Korean Soc Breeding Sci. 2015;3(2):139–52.
18. Sardaro MLS, Marmiroli M, Maestri E, Marmiroli N. Genetic characterization of Italian tomato varieties and their traceability in tomato food products. Food Sci Nutr. 2013;1:54–62.
19. Baleiras-Couto MM, Eiras-Dias JE. Detection and identification of grape varieties in must and wine using nuclear and chloroplast microsatellite markers. Anal Chim Acta. 2006;563:283–91.
20. Pasqualone A, Montemurro C, Caponio F, Blanco A. Identification of virgin olive oil from different cultivars by analysis of DNA microsatellites. J Agric Food Chem. 2004;52:1068–71.
21. Nevo E. Genetic diversity. In: Levin SA, editor. Encyclopedia of biodiversity [Internet]. 2nd ed. Waltham: Academic Press; 2001 [cited 2021 Jan 31]. p. 662–77. Available from: http://www.sciencedirect.com/science/article/pii/B9780123847195000654.
22. Pérez-Jiménez M, Besnard G, Dorado G, Hernandez P. Varietal tracing of virgin olive oils based on plastid DNA variation profiling. PLoS One. 2013;8:e70507.
23. Phumichai C, Phumichai T, Wongkaew A. Novel chloroplast microsatellite (cpSSR) markers for genetic diversity assessment of cultivated and wild Hevea rubber. Plant Mol Biol Rep. 2015;33:1486–98.
24. Lawson MJ, Zhang L. Distinct patterns of SSR distribution in the Arabidopsis thaliana and rice genomes. Genome Biol. 2006;7:R14.

25. Ramsay L, Macaulay M, Cardle L, Morgante M, Ivanissevich SD, Maestri E, et al. Intimate association of microsatellite repeats with retrotransposons and other dispersed repetitive elements in barley. Plant J. 1999;17:415–25.
26. Li CD, Rossnagel BG, Scoles GJ. The development of oat microsatellite markers and their use in identifying relationships among Avena species and oat cultivars. Theor Appl Genet. 2000;101:1259–68.
27. Temnykh S, DeClerck G, Lukashova A, Lipovich L, Cartinhour S, McCouch S. Computational and experimental analysis of microsatellites in rice (Oryza sativa L.): frequency, length variation, transposon associations, and genetic marker potential. Genome Res. 2001;11:1441–52.
28. Qu J, Liu J. A genome-wide analysis of simple sequence repeats in maize and the development of polymorphism markers from next-generation sequence data. BMC Res. 2013;6:1–10.
29. Kejnovsky E, Hobza R, Cermak T, Kubat Z, Vyskot B. The role of repetitive DNA in structure and evolution of sex chromosomes in plants. Heredity. 2009;102:533–41.
30. Poltronieri J, Marquioni V, Bertollo LAC, Kejnovsky E, Molina WF, Liehr T, et al. Comparative chromosomal mapping of microsatellites in Leporinus species (Characiformes, Anostomidae): unequal accumulation on the W chromosomes. Cytogenet Genome Res. 2014;142:40–5.
31. Zhang L, Yuan D, Yu S, Li Z, Cao Y, Miao Z, et al. Preference of simple sequence repeats in coding and non-coding regions of Arabidopsis thaliana. Bioinformatics. 2004;20:1081–6.
32. Xu JIE, Liu L, Xu Y, Chen C, Rong T, Ali F, et al. Development and characterization of simple sequence repeat markers providing genome-wide coverage and high resolution in maize. DNA Res. 2013;20:497–509.
33. Katti MV, Sami-Subbu R, Ranjekar PK, Gupta VS. Amino acid repeat patterns in protein sequences: their diversity and structural-functional implications. Protein Sci. 2000;9:1203–9.
34. Huntley M, Golding GB. Evolution of simple sequence in proteins. J Mol Evol. 2000;51:131–40.
35. Oliveira EJ, Pádua JG, Zucchi MI, Vencovsky R, Vieira MLC. Origin, evolution and genome distribution of microsatellites. Genet Mol Biol. 2006;29:294–307.
36. Mason: plant genotyping—Google Scholar [Internet]. [cited 2021 Jan 31]. Available from: https://scholar.google.com/scholar_lookup?title=Plant+Genotyping&author=AS+Mason&publication_year=2015&.
37. O'Connell M, Wright JM. Microsatellite DNA in fishes. Rev Fish Biol Fish. 1997;7:331–63.
38. Webster MS, Reichart L. Use of microsatellites for parentage and kinship analyses in animals. Methods in enzymology [Internet]. New York: Academic Press; 2005 [cited 2021 Jan 31]. p. 222–38. Available from: http://www.sciencedirect.com/science/article/pii/S0076687905950143.
39. Hansen MM, Kenchington E, Nielsen EE. Assigning individual fish to populations using microsatellite DNA markers. Fish Fish. 2001;2:93–112.
40. Marcotte EM, Pellegrini M, Yeates TO, Eisenberg D. A census of protein repeats11Edited by J. M. Thornton. J Mol Biol. 1999;293:151–60.
41. Wren JD, Forgacs E, Fondon JW, Pertsemlidis A, Cheng SY, Gallardo T, et al. Repeat polymorphisms within gene regions: phenotypic and evolutionary implications. Am J Hum Genet. 2000;67:345–56.
42. Li Y-C, Korol AB, Fahima T, Beiles A, Nevo E. Microsatellites: genomic distribution, putative functions and mutational mechanisms: a review. Mol Ecol. 2002;11:2453–65.
43. Li Y-C, Korol AB, Fahima T, Nevo E. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol. 2004;21:991–1007.
44. Cummings CJ, Zoghbi HY. Fourteen and counting: unraveling trinucleotide repeat diseases. Hum Mol Genet. 2000;9:909–16.
45. Streelman JT, Kocher TD. Microsatellite variation associated with prolactin expression and growth of salt-challenged tilapia. Physiol Genomics. 2002;9:1–4.

46. Liquori CL, Ricker K, Moseley ML, Jacobsen JF, Kress W, Naylor SL, et al. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science. 2001;293:864–7.
47. Holmes E. In: Li W-H, Graur D, editors. Fundamentals of molecular evolution. Sunderland, MA: Sinauer; 1991. + ± 284 pp. $22.95 (paper). American Journal of Physical Anthropology. 1991;85:363–5.
48. Ennos RA. Utilising genetic information in plant conservatio...—Google Scholar [Internet]. 1996 [cited 2021 Jan 31]. Available from: https://scholar.google.com/scholar_lookup?title=Utilising+Genetic+Information+in+Plant+Conservation+Programmes&author=Ennos,+R.A.&publication_year=1996&pages=278%E2%80%93291.
49. Bekessy SA, Ennos RA, Burgman MA, Newton AC, Ades PK. Neutral DNA markers fail to detect genetic divergence in an ecologically important trait. Biol Conserv. 2003;110:267–75.
50. van Tienderen PH, de Haan AA, van der Linden CG, Vosman B. Biodiversity assessment using markers for ecologically important traits. Trends Ecol Evol. 2002;17:577–82.
51. Holderegger R, Kamm U, Gugerli F. Adaptive vs. neutral genetic diversity: implications for landscape genetics. Landsc Ecol. 2006;21:797–807.
52. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature. 2005;436:793–800.
53. Zhang Z, Deng Y, Tan J, Hu S, Yu J, Xue Q. A genome-wide microsatellite polymorphism database for the Indica and Japonica rice. DNA Res. 2007;14:37–45.
54. Parida SK, Dalal V, Singh AK, Singh NK, Mohapatra T. Genic non-coding microsatellites in the rice genome: characterization, marker design and use in assessing genetic and evolutionary relationships among domesticated groups. BMC Genomics. 2009;10:140.
55. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science. 2002;296:92–100.
56. Mackill DJ. Applications of genomics to rice breeding. Int Rice Res Notes. 2003;28:9–15.
57. Saarinen EV, Austin JD. When technology meets conservation: increased microsatellite marker production using 454 genome sequencing on the endangered Okaloosa darter (Etheostoma okaloosae). J Heredity. 2010;101:784–8.
58. Conaway-Bormans CA, Marchetti MA, Johnson CW, McClung AM, Park WD. Molecular markers linked to the blast resistance gene Pi-z in rice for use in marker-assisted selection. Theor Appl Genet. 2003;107:1014–20.
59. Ware D, Jaiswal P, Ni J, Pan X, Chang K, Clark K, et al. Gramene: a resource for comparative grass genomics. Nucleic Acids Res. 2002;30:103–5.
60. Fjellstrom R, McClung AM, Shank AR. SSR markers closely linked to the Pi-z locus are useful for selection of blast resistance in a broad array of rice germplasm. Mol Breed. 2006;17:149–57.
61. Abdelkrim J, Robertson BC, Stanton J-AL, Gemmell NJ. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. BioTechniques. 2009;46:185–92.
62. Dutta S, Kumawat G, Singh BP, Gupta DK, Singh S, Dogra V, et al. Development of genic-SSR markers by deep transcriptome sequencing in pigeonpea [Cajanus cajan (L.) Millspaugh]. BMC Plant Biol. 2011;11:1–13.
63. Santana QC, Coetzee MP, Steenkamp ET, Mlonyeni OX, Hammond GN, Wingfield MJ, et al. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques. 2009;46:217–23.
64. Churbanov A, Ryan R, Hasan N, Bailey D, Chen H, Milligan B, et al. HighSSR: high-throughput SSR characterization and locus development from next-gen sequencing data. Bioinformatics. 2012;28:2797–803.
65. Liu L, Li Y, Li S, Hu N, He Y, Pong R, et al. Comparison of next-generation sequencing systems. J Biomed Biotechnol. 2012;2012:251364.
66. Moran C. Looking back to move forward—a personal perspective on pig molecular genetics from RFLPs to nextgen sequencing. Database; 2012.

67. Fjellstrom R, Conaway-Bormans CA, McClung AM, Marchetti MA, Shank AR, Park WD. Development of DNA markers suitable for marker assisted selection of three Pi genes conferring resistance to multiple Pyricularia grisea pathotypes. Crop Sci. 2004;44:1790–8.
68. Temnykh S, Park WD, Ayres N, Cartinhour S, Hauck N, Lipovich L, et al. Mapping and genome organization of microsatellite sequences in rice (Oryza sativa L.). Theor Appl Genet. 2000;100:697–712.
69. Conaway C, Cartinhour S, Ayres N, McClung AM, Lai XH, Marchetti MA, et al. PCR based markers linked to blast resistance genes in rice. In: Proceedings of the 27th rice technical working group meeting, Reno-Sparks, NV, USA. 1998. p. 1–4.
70. McCouch SR. Development and mapping of 2240 new SSR markers for rice (Oryza sativa L.). DNA Res. 2002;9:199–207.
71. Wang Z-X, Yano M, Yamanouchi U, Iwamoto M, Monna L, Hayasaka H, et al. The Pib gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. Plant J. 1999;19:55–64.
72. Zhang Y-X, Wang Q, Jiang L, Liu L-L, Wang B-X, Shen Y-Y, et al. Fine mapping of qSTV11 KAS, a major QTL for rice stripe disease resistance. Theor Appl Genet. 2011;122:1591–604.
73. Romero G, Adeva C, Battad Z. Genetic fingerprinting: advancing the frontiers of crop biology research. Philipp Sci Lett. 2009;2:8–13.
74. Joshi SP, Ranjekar PK, Gupta VS. Molecular markers in plant genome analysis. Curr Sci. 1999;77:230–40.
75. Powell W, Machray GC, Provan J. Polymorphism revealed by simple sequence repeats. Trends Plant Sci. 1996;1:215–22.
76. Rode J, In-Chol K, Saal B, Flachowsky H, Kriese U, Weber WE. Sex-linked SSR markers in hemp. Plant Breed. 2005;124:167–70.
77. Fraser LG, Tsang GK, Datson PM, De Silva HN, Harvey CF, Gill GP, et al. A gene-rich linkage map in the dioecious species Actinidia chinensis (kiwifruit) reveals putative X/Y sex-determining chromosomes. BMC Genomics. 2009;10:1–15.
78. Jakse J, Stajner N, Kozjak P, Cerenak A, Javornik B. Trinucleotide microsatellite repeat is tightly linked to male sex in hop (Humulus lupulus L.). Mol Breed. 2008;21:139–48.
79. Spigler RB, Lewers KS, Main DS, Ashman TL. Genetic mapping of sex determination in a wild strawberry, Fragaria virginiana, reveals earliest form of sex chromosome. Heredity. 2008;101:507–17.
80. Microsatellite (GATA)n reveals sex-specific differences in Papaya—ProQuest [Internet]. [cited 2021 Feb 6]. Available from: https://search.proquest.com/openview/e2f039b242ec448f15ead2abb665e958/1?pq-origsite=gscholar&cbl=54040.
81. Provan J, Powell W, Hollingsworth PM. Chloroplast microsatellites: new tools for studies in plant ecology and evolution. Trends Ecol Evol. 2001;16:142–7.
82. Wang ML, Barkley NA, Jenkins TM. Microsatellite markers in plants and insects. Part I: applications of biotechnology. 2009 [cited 2021 Feb 6]. Available from: https://pubag.nal.usda.gov/catalog/44058.
83. Neeraja CN, Maghirang-Rodriguez R, Pamplona A, Heuer S, Collard BCY, Septiningsih EM, et al. A marker-assisted backcross approach for developing submergence-tolerant rice cultivars. Theor Appl Genet. 2007;115:767–76.
84. Varshney RK, Graner A, Sorrells ME. Genic microsatellite markers in plants: features and applications. Trends Biotechnol. 2005;23:48–55.
85. Breseghello F, Sorrells ME. Association mapping of kernel size and milling quality in wheat (Triticum aestivum L.) cultivars. Genetics. 2006;172:1165–77.
86. Simko I, Costanzo S, Haynes KG, Christ BJ, Jones RW. Linkage disequilibrium mapping of a *Verticillium dahliae* resistance quantitative trait locus in tetraploid potato (*Solanum tuberosum*) through a candidate gene approach. TAG Theor Appl Genet. 2004;108:217–24.
87. Srinivas G, Satish K, Madhusudhana R, Seetharama N. Exploration and mapping of microsatellite markers from subtracted drought stress ESTs in Sorghum bicolor (L.) Moench. Theor Appl Genet. 2009;118:703–17.

88. Varshney RK, Hiremath PJ, Lekha P, Kashiwagi J, Balaji J, Deokar AA, et al. A comprehensive resource of drought- and salinity-responsive ESTs for gene discovery and marker development in chickpea (Cicer arietinum L.). BMC Genomics. 2009;10:523.
89. Schranz ME, Song B-H, Windsor AJ, Mitchell-Olds T. Comparative genomics in the Brassicaceae: a family-wide perspective. Curr Opin Plant Biol. 2007;10:168–75.
90. Barreneche T, Casasoli M, Russell K, Akkak A, Meddour H, Plomion C, et al. Comparative mapping between Quercus and Castanea using simple-sequence repeats (SSRs). Theor Appl Genet. 2004;108:558–66.
91. Yu J-K, Dake TM, Singh S, Benscher D, Li W, Gill B, et al. Development and mapping of EST-derived simple sequence repeat markers for hexaploid wheat. Genome. 2004;47:805–18.
92. Stein N, Prasad M, Scholz U, Thiel T, Zhang H, Wolf M, et al. A 1,000-loci transcript map of the barley genome: new anchoring points for integrative grass genomics. Theor Appl Genet. 2007;114:823–39.
93. Shultz JL, Kazi S, Bashir R, Afzal JA, Lightfoot DA. The development of BAC-end sequence-based microsatellite markers and placement in the physical and genetic maps of soybean. Theor Appl Genet. 2007;114:1081–90.
94. Shoemaker RC, Grant D, Olson T, Warren WC, Wing R, Yu Y, et al. Microsatellite discovery from BAC end sequences and genetic mapping to anchor the soybean physical and genetic maps. Genome. 2008;51:294–302.
95. Wang ML, Huang L, Bongard-Pierce DK, Belmonte S, Zachgo EA, Morris JW, et al. Construction of an ~2 Mb contig in the region around 80 cM of Arabidopsis thaliana chromosome 2. Plant J. 1997;12:711–30.
96. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28:1647–9.
97. Wang X, Lu P, Luo Z. GMATo: a novel tool for the identification and analysis of microsatellites in large genomes. Bioinformation. 2013;9:541–4.
98. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27:573–80.
99. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
100. Thiel T, Michalek W, Varshney R, Graner A. Exploiting EST databases for the development and characterization of gene-derived SSR-markers in barley (Hordeum vulgare L.). Theor Appl Genet. 2003;106:411–22.
101. Hodel RGJ, Segovia-Salcedo MC, Landis JB, Crowl AA, Sun M, Liu X, et al. The report of my death was an exaggeration: a review for researchers using microsatellites in the 21st century. Appl Plant Sci [Internet]. 2016 [cited 2021 Feb 6];4. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4915923/.
102. Faircloth BC. Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour. 2008;8:92–4.
103. Castoe TA, Poole AW, de Koning APJ, Jones KL, Tomback DF, SJ O-MC, et al. Rapid microsatellite identification from illumina paired-end genomic sequencing in two birds and a snake. PLoS One. 2012;7:e30953.
104. Meglécz E, Pech N, Gilles A, Dubut V, Hingamp P, Trilles A, et al. QDD version 3.1: a user-friendly computer program for microsatellite selection and primer design revisited: experimental validation of variables determining genotyping success rate. Mol Ecol Resour. 2014;14:1302–13.
105. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Cech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Res. 2018;46:W537–44.
106. Tarailo-Graovac M, Chen N. Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinform. 2009;Chapter 4:Unit 4.10.

107. da Maia LC, Palmieri DA, de Souza VQ, Kopp MM, de Carvalho FIF, Costa de Oliveira A. SSR locator: tool for simple sequence repeat discovery integrated with primer design and PCR simulation. Int J Plant Genomics. 2008;2008:412696.
108. Miller MP, Knaus BJ, Mullins TD, Haig SM. SSR_pipeline: a bioinformatic infrastructure for identifying microsatellites from paired-end Illumina high-throughput DNA sequencing data. J Heredity. 2013;104(6):881–5.
109. Kraemer L, Beszteri B, Gäbler-Schwarz S, Held C, Leese F, Mayer C, et al. STAMP: extensions to the STADEN sequence analysis package for high throughput interactive microsatellite marker design. BMC Bioinformatics. 2009;10:41.
110. Mayer C. Phobos: highly accurate search for perfect and imperfect tandem repeats in complete genomes by Christoph Mayer [Internet]. Christoph Mayer; 2007 [cited 2021 Feb 6]. Available from: https://www.ruhr-uni-bochum.de/spezzoo/cm/cm_phobos.htm.
111. Castelo AT, Martins W, Gao GR. TROLL—tandem repeat occurrence locator. Bioinformatics. 2002;18:634–6.

Status and Prospectives of Genome-Wide Association Studies in Plants

19

Goutam Kumar Dash, S. Sabarinathan, Ravindra Donde, Gayatri Gouda, Manoj Kumar Gupta, Lambodar Behera, and Padmini Swain

Abstract

The genome-wide association study (GWAS) is one of the potential approaches for identifying QTLs/genes and complex traits associated with target traits quickly using natural variations. With the advancement in genome sequencing technology, it is possible to examine genome-wide genetic variants of agro-morphological, physiological, biochemical, and molecular traits across diverse genetic materials. In nature, natural variants of crops are generated due to spontaneous mutations and manual breeding in the wild progenitors. Traditional landraces are adapted to various environmental conditions, rich sources of alleles, and genes linked with various traits valuable for variety improvement through molecular breeding because of the availability of high-throughput sequencing technologies and a reference genome sequence, accurate re-sequencing of a significant number of crop genomes as possible. It aided in understanding the genetic basis of phenotypic variance and allows for functional studies of evolutionary changes in crops. This rapid development will significantly improve crop design research using genomics-assisted breeding, allowing it to be used in gene recognition, cloning, QTL identification, and crop breeding using marker-assisted selection or genetic engineering. This book chapter presents an overview of the entire process of a typical GWAS, various software applications and, its limitations, and future perspectives.

G. K. Dash (✉) · R. Donde · G. Gouda · M. K. Gupta · L. Behera · P. Swain
ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_19

Keywords

Genome-wide association study · Linkage disequilibrium · Rice

Abbreviations

BC	Bonferroni correction
BILs	Backcross inbred lines
CMLM	Compress Mixed Linear Model
EMMA	Effective Mixed-Model Association
EMMAX	Efficient Mixed-Model Association eXpedited
FDR	False discovery rate
GAPIT	Genome Association and Prediction Integrated Tool
GLM	General linear model
GWAS	Genome-wide association study
LD	Linkage disequilibrium
MAGIC	Multiparent Advanced Generation Inter-Cross
MAS	Marker assisted selection
MLM	Mixed linear model
NAM	Nested association mapping
NILs	Near Isogenic lines
QTL	Quantitative trait locus
RILs	Recombinant Inbred Lines
SNP	Single nucleotide polymorphism
TILLING	Targeting induced local lesions in genome

19.1 Introduction

Since the beginning of cultivation, a huge number of crops have been adapted to various environmental conditions because of spontaneous mutations that existed in their wild progenitors, producing natural variants. In this aspect, crop domestication influenced the genetic diversity of crops. Improvement in crop productivity to satisfy the need for food demand of the growing population requires a critical understanding of the genetic basis of phenotypic variation that can be employed in an advanced breeding program [1–3]. With advanced genomic technologies such as genome sequencing, "genome-wide association studies," haplotype map, and genetic-transformation technique [4, 5], researchers are able to mine natural variants and their associated phenotypic variations [6, 7]. In recent years genome-wide association study has got a lot of attention. Advancement in genomic technology initiated a wave of association mapping in model and crop plants that enabled the study of trait variation across diverse genetic backgrounds [8]. Rapid and continued progress in sequencing technologies, along with the availability of reference genome, high-density genotyping array, and accurate phenotypic trait measurement using high-throughput phenotyping, has made the GWAS a method of choice [9]. In the recent decade, GWAS has transformed from a new promising tool to a powerful, ubiquitous technique for dissecting complex traits in plants [8, 10–12].

If a phenotype exists within a subpopulation, it must be linked to the neighboring genetic variations in their recent ancestors. This phenomenon is known as linkage disequilibrium (LD) based on which GWAS operates. Recent advanced genotyping technology and progress in information technology enabled genotype to phenotype association study on a genome-wide scale in a very large population, thus accelerating quantitative traits mapping. In contrast to linkage mapping, which is time-consuming and has relatively low resolution, GWAS can discover genomic regions associated with a particular phenotype in a relatively high resolution and in an unbiased manner. Also, the experimental materials used in GWAS are more diverse than those used in traditional linkage mapping (derived from biparental crossing) due to many historical recombination events [13]. Most of the traits of agronomic and evolutionary importance are complex traits that are influenced by multi genes, environment, and their interactions [14]. To date, in many crops, genome-wide association studies have been conducted [10, 15–17] for different traits such as plant height, flowering time, grain yield, kernel number, and stress tolerance [10, 15, 16, 18]. It is also used to identify genes with geographical divergence and adaptation during domestication, [19] genes associated with biochemical and molecular phenotypes, including fatty acid, flavonoid, amino acid, and nucleic acid metabolites [20]. The high-throughput automated phenotyping system has facilitated the measurement of the complex traits that speed up in observing the natural variants in a very large population with much accuracy [21, 22]. It is also used to validate the loci identified through other approaches, improve transgenic research by recognizing genes [23], and have huge applications to identify target genes for editing [24]. The global landscape or the genetic architecture of a trait can also be revealed by GWAS, which includes the number of alleles of that particular gene, their distribution pattern and interactions, and their effects [25]. Despite multiple advantages, there are issues regarding population structure and occurence of false-negative results due to low frequency of causal alleles [26]. In the recent years, using several efficient statistical algorithms feasible for plant populations (the mixed model) [27–34], hundreds of associated loci are identified in rice. The experiments and critical factors required for a good GWAS are described in this chapter, along with the generation of a GWAS population, softwares used, its limitations, and future perspectives.

Association mapping is a powerful tool for recognizing particular genetic regions/markers in crops that are associated with agronomically important traits [35]. This method has great potential for evaluating and characterizing a wide variety of alleles. These polymorphic alleles are very tightly linked with a locus that influences the phenotypic effect significantly and associated with traits in a randomly mating population, thus allowing for a much finer resolution than genetic mapping. As a result, association mapping based on linkage disequilibrium (LD) could lead to discovering the genes responsible for QTLs. The MAS is more likely to succeed in different backgrounds since the association method works for a wide variety of germplasm. In GWAS, there is also the possibility of identifying QTLs that are linked to several traits. Thus, applying this approach may take advantage of historical recombination events in natural populations, minimizing the expense and time required for research. This book chapter has enlisted the application of association mappings, for identifying QTLs linked with different agronomic traits.

19.2 History and Development of GWAS Study

In a typical GWAS, genotypes and phenotype data are collected for a large, diverse population. After that, the significant associations of genetic markers with the phenotype data are established using statistical methods. However, the linked genetic markers may not always present within the causative gene of the studied phenotype; that is why GWAS relies on linkage disequilibrium (LD) between markers under testing and functional polymorphisms of the gene of interest. Generally, the loci nearer to each other on a chromosome have fewer chances of getting separated through recombination than those that are distant from each other. This nonrandom association of alleles at two loci is called linkage disequilibrium. The SNPs nearer to the causative locus can be in high LD with the functional polymorphism thus associated with the phenotype of interest. These genomic regions are identified and marked through genome-wide association studies. If the period after the last common ancestor in which functional polymorphisms were produced by mutation is considered in the unrelated populations desired in GWAS, the genomic regions in LD can be narrow, making them well suited for high-resolution mapping of the gene responsible [12, 36–38].

Quantitative trait locus (QTL) mapping through linkage analysis is the older version of GWAS, which studies individuals with known relationships instead of taking diverse individuals. For example, in linkage mapping, the individuals or populations used are the progeny of biparental crosses (either F_2 or Recombinant Inbred Lines (RIL), Backcross Inbred Lines (BIL), and Near Isogenic Lines (NIL)). Here, the QTL-linked genetic markers co-segregate with the phenotype of interest. Since the individuals came from a biparental cross, the number of recombination events from their most common ancestors is low, creates large linkage blocks that can be detected with genetic markers less dense than GWAS. After the QTL detection and validation, the targeted genomic area is used for fine mapping and QTL cloning [10, 14, 26, 36, 39]. This technique was adopted to identify QTLs before the sequencing technologies came into the picture. Through this technique, the first genome-wide QTL identification was made in tomato in 1988 [40], 14 years before the first GWAS technique was used [41]. Both linkage analysis and GWAS are used to understand complex traits in different species in the present day.

The primary advantage of GWAS over linkage analysis is that GWAS do not need any experimental crosses [42] and can detect genes with smaller effects [43] with improved resolution having smaller blocks of LD [42]. However, previous association studies carried out before next-generation sequencing could address only a small part of the genome or the region of interest already identified by other methods [44]. Later, researchers made the association study genome-wide with dense genetic markers covering the whole-genome [36, 45–47]. But the major drawback with this technique [36, 45] was that it used to give high rates of false-positive because of population structure [42] and multiple testing [48]. That is why they have to wait till the publication of the draft of the human genome in 2001 (International Human Genome Sequencing Consortium, 2001) and subsequent availability of early SNP datasets [49] and HapMap (The International HapMap Consortium, 2003). After that, the first paper of GWAS was published in 2002, taking 65,000 SNPs and 94 individuals [41].

The first GWAS paper in plants was published in 2005 on *Arabidopsis* [50]. Afterward, many GWAS papers were published in which statistical methods were used to associate genetic markers with phenotype to find the causative SNPs associated with variations in phenotypes [51], which was by conducting simple ANOVA on each SNPs considering the assumption that the difference between the trait means for any genotype group (i.e., AA, Aa, and aa) is nil and can then be tested for every SNP. However, this approach gave high rates of false-positives (the association declared significant even though they were not) because of many statistical tests. To deal with this, a significance threshold of 0.05 was used, which means the false-positive rate is accepted only up to 5%, which was an acceptable risk. The common methods to multiple testing correction are limiting the false discovery rate (FDR) [52, 53] or using the Bonferroni correction, which is the proportion of desired significance threshold to the total number of tests conducted to determine the corrected significance threshold. However, setting an appropriate significance threshold presents additional challenges in GWAS. Besides this, the close relatedness among the individuals forming subpopulation with the diverse population used for GWAS imposing other major factors contributes to the false-positives in GWAS [27, 54]. As a result, some spurious association was displayed by the SNPs common in the subpopulations and the phenotype of interest, if the phenotype was present at high frequency in that group.

19.3 How GWAS Works

The size of the population is one of the important factors that determine the success of the experiment. It is generally recommended to increase the size of the population to at least above hundred individuals to avoid Beavis effects that greatly overrated the phenotypic variance when the population size is small [37]. After selecting enough number of individuals, the genotypes need to be phenotyped accurately for a particular trait or a group of traits. Accurate phenotyping is essential and should be repeated over replications, locations, and years. The broad-sense heritability must be calculated since higher heritability shows that the trait of interest is mostly under genetic regulation, which is essential for detecting genotype–phenotype associations. Then the genotyping is done taking the same set of individuals of which phenotyping has been done using DNA markers. Usually, GBS-based genotyping is done that covers the whole genome and identifies many SNPs (e.g., wheat, barley). Before initiating analysis, the GBS-generated SNPs must be free from missing data, heterozygosity, and minor allele frequency. Also, the population structure must be tested before running GWAS to select the appropriate model.

There are two statistical models, general linear model (GLM) and mixed linear model (MLM), used for performing GWAS. The GLM does not consider population structure into account [55, 56], whereas MLM considers population structure in its model (Kinship or kinship + Q matrix + PCAs). Using appropriate software (e.g., TASSEL), phenotype and genotype data are combined through which the causative alleles for the particular trait are detected. Phenotyping should be done before genotyping, especially for those with no prior information, to save time and money since the individuals of the population collected from different regions may not adapt to the phenotyping environment where the genotypes are going to be tested and may be lost due to poor adaptation.

After the analysis, the significance of the marker–trait association is examined through the false discovery rate (FDR) or Bonferroni correction (BC). Bonferroni correction method is applied by dividing the level of significance by the number of markers at each locus that gave the information about the threshold of significant markers for several traits at once [57, 58]. As a result, a fixed BC *p*-value is generated. However, in false discovery rate returns, all the associations' actual significant associations are found [59]. In this test, the *p*-value of each marker is given a rank after arranging them in ascending order. The FDR is calculated by dividing the product of the rank number of marker *p*-value and a factor (usually 0.05) with the total number of markers. Deriving FDR values for each trait makes it a powerful tool in analyzing agronomic and developmental traits in crops. FDR *p*-values are more flexible in detecting a highly significant association between marker and trait than the fixed *p*-values of BC. In both cases, if the *p*-value for FDR or BC at a particular locus is less or equal to the *p*-value obtained from GWAS, then the association is considered to be true. Generally, the significance for the marker–trait association is tested at 0.05 and 0.01 levels [60, 61]. However, the significance of some markers–trait association measured at 20% using FDR since the marker can detect minor effects [38]. The level of significance opted for the marker–trait association is based on the requirement of the study in which low FDR is used to identify candidate genes or loci. In contrast, high FDR is used to predict the whole picture of the genetic architecture of the trait [46].

19.4 Software for Performing GWAS Study

For performing GWAS analysis, many statistical softwares are used. Frequently used softwares for GWAS analysis is discussed here. TASSEL (Trait Analysis by Association, Evolution, and Linkage) is the most frequently used for GWAS in plants. It includes many statistical methods like GLM, MLM, and FaST for performing GWAS [47]. Analyze population structure using kinship, and PCA also includes LD. The software is usually used for association studies in barley [62]. However, the recent version of TASSEL can perform SNP calling using GBS data and analyze genetic diversity. It includes many visualizing tools such as scatter plots for PCA, LD, Manhattan plot, heatmap for genetic distance, and phylogenetic tree construction using archaeopteryx. The newer version also enables the researcher to quickly view genotypes, markers, missing data, heterozygous, and the number of markers on each chromosome. The older version of the TASSEL, like TASSEL v.2.1 can analyze taking any type of markers (SNP, SSR, AFLP, RAPD, etc.); however, the newer version accepts only SNP data. It is a completely free software.

GenStat is another Windows-based statistical software used for marker–trait association analysis in a genetically diverse population using bi-allelic and multi-allelic markers. It also uses GLM and MLM models along with population structure correction to analyze GWAS. In this software, the threshold of the significance level can be selected, and LD decay can also be estimated. Also, the effect of each SNP can be estimated using its visualization of the location of the significant markers and Q-Q plot. Therefore, it is widely used to detect the causative alleles. It is not a free software.

PLINK facilitates the study of a large dataset of phenotypes and genotypes [63]. It is a free program that includes features like population stratification identification, simple interaction checks, meta-analyses, and other tests like gene-based association tests and epistasis screening. This app can display graphical images for Manhattan plots, Q-Q plots, and multidimensional scaling (for population structure). Tables produced by PLINK can also be used to present the results of GWAS and LD among SNP markers.

R statistical environment (https://www.r-project.org/) also provides a useful package Genome Association and Prediction Integrated Tool (GAPIT) for performing GWAS, which can deal with large numbers of SNPs and genotypes with less computational time without conceding the statistical power [64]. Many statistical approaches such as MLM, previously defined population parameters (P3D), and effective mixed-model association (EMMA) are included in this package. Manhattan maps, quantile-quantile (Q-Q) plots, and a table with the *p*-value, minor allele frequency, sample size, phenotypic variation explained by markers R2, and corrected *P*-value after a false discovery rate can be used to demonstrate GWAS findings [59]. Kinship analyses are shown in a heat map and a table, as well. Furthermore, through graphs with various compression levels, the heritability estimates and probability functions are presented. Because of the features mentioned above, GAPIT is considered the most effective and valuable method for association analysis [52–54]. Since GenSta was one of the first tools to perform the tests and has several features not found in other software, there is a strong pattern of using it for QTL and candidate gene recognition. For example, the evaluation of phenotypic and genotypic data, the calculation for BLUE values, LD, and population structure using PCA and kinship, and finally GWAS using either GLM or MLM can be performed GenStat. In addition to the G X E relationship, the output contains all the relevant plots and statistics about the marker–trait correlations, such as the impact size of the marker on the trait. Finally, Bonferroni correction is used to confirm the significant correlations. For GWAS in other software/packages, each move must also be measured separately. For GWAS, each step needs to be calculated separately in other software/packages. The detailed list of softwares used for GWAS is provided in the Table 19.1.

19.5 Linkage Disequilibrium Mapping in Crops

A massive sampling size typically requires thousands of individuals to decipher the genetic foundation of complex QTLs, such as grain yield and stress tolerance. It is now possible to genotype thousands of genomes using high-throughput sequencing techniques. As compared with genotyping techniques, not much development has taken place to date in phenotyping. Most phenotyping and field studies were time-consuming and stressful; it necessitated the evaluation of several traits at multiple time points in a large-scale experiment through a variety of ecosystems. In this context, some sensor-based platforms for measuring biomass have been created with near-infrared spectroscopy, and spectral reflectance on agriculture harvesters plant canopies, respectively [165]. The advancement of phenotyping technology in the future will hasten genetic mapping and gene exploration in crops.

For association analyses, the LD decay interval offers critical knowledge about marker densities [166, 167]. The distribution of markers has a major impact on the LD decay resolution for association analysis [168]. The big red blocks revealed a high degree of LD decay between the loci, which resulted from less recombination of LD blocks along with the triangle plot's diagonal [35, 169, 170]. In 2003, Garris and team have discovered LD decay at 100–200 kb intervals through a single area of chromosome 5 [171]. Similarly, Olsen and the team have investigated a 500-kb area on chromosome 6 and found a 250-kb selective for a waxy locus, which resulted in an elevated LD region [172].

For genetic mapping in crops, segregating mapping populations such as F_2 groups, "Recombinant Inbred Lines" (RILs), and "Backcross Inbred Lines" (BILs) are commonly used. Fine mapping and gene cloning are often carried out, using specialized backcross-derived populations. A mapping population has been produced from a cross between *Oryza sativa* ssp. *indica* Kasalath and *Oryza sativa* ssp. *japonica*. According to a study reported by Harushima and the team in 1998, the "Nipponbare varieties" have allowed for the discovery and cloning of tens of QTLs underlying a diverse variety of traits [173]. This method has been widely used in functional genomics studies on crops, but there are two significant drawbacks to QTL mapping in typical recombinant populations. There are just a few recombination events in the mapping population; for example, in rice segregating populations, one or two recombination events occur in each chromosome. Thus, fewer mapping samples result in a poor resolution; so, very massive populations are used to get significant result. Second, since the sequence variation between the preferred parents accounts for just a small fraction of all genetic variance within a group of species,

Table 19.1 List of different softwares/applications used in GWAS analysis

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
1	AlphaDrop_beta	QTL effects, population, and pedigree structure (stimulate sequence and SNP data (genomic and phenotype data))	Script	Fortran 95	https://sites.google.com/site/hickeyjohn/alphadrop	Linux, Microsoft Windows	–	[65]
2	Altools	Indels and SNP data (population and haplotype structure)	Desktop graphical user interface	R, Java, C++	https://sourceforge.net/projects/altools/	Linux	Free	[55]
3	aSPU	SNP data (gene- and pathway-based tests)	Command-line user interface, Library	R	https://cran.r-project.org/web/packages/aSPU/	Linux, Mac OS X, Microsoft Windows	Free	[56]
4	Bayenv	Bayesian method and standardized allele frequencies (SNPs)	Command-line interface	C	https://gcbias.org/	Linux, Mac OS X	Free	[57]
5	BigTop	Manhattan plot in 3D (a visualization framework in virtual reality)	Web user interface	JavaScript	https://github.com/dnanexus/bigtop	Linux, Mac OS X, Microsoft Windows	Free	[58]
6	BioBin	Automate the binning of low frequency variants	Command-line user interface	C++	https://ritchielab.org/software/biobin-download	Microsoft Windows	Free	[60]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
7	BLINK	"Bayesian-information and Linkage-disequilibrium" Iteratively Nested Keyway (MLM)	Command-line user interface	R, C	http://zzlab.net/blink/	Linux, Mac OS X, Microsoft Windows	Free	[61]
8	BWMR	Bayesian weighted Mendelian randomization (variational expectation-maximization)	Command-line user interface	R	https://github.com/jiazhao97/BWMR	Linux, Mac OS X, Microsoft Windows	Free	[66]
9	CERENKOV	Noncoding SNPs in loci identified by GWAS	Command-line interface, R package	R	https://github.com/ramseylab/cerenkov	Linux	Free	[67]
10	cit	Likelihood-based hypothesis testing approach	Command-line user interface	R	https://cran.r-project.org/web/packages/cit/index.html	Mac OS X, Microsoft Windows	Free	[68]
11	CPBayes	Bayesian meta-analysis method (cross-phenotype genetic associations)	Command-line user interface, Library	R	https://cran.r-project.org/web/packages/CPBayes/index.html	Linux, Mac OS X, Microsoft Windows	Free	[69]
12	EMMAX	Testing efficient mixed model association	Command-line user interface	C++, C	http://genetics.cs.ucla.edu/emmax/	Linux	Free	[30]
13	EPIQ	SNP Epistasis detection for Quantitative GWAS	Command-line interface	C++	https://github.com/yaarasegre/EPIQ	Linux, Mac OS X	Free	[70]

14	Evoker	Graphical tool for visualizing genotype intensity data	Desktop graphical user interface	Java, Perl	https://www.sanger.ac.uk/tool/evoker/	Linux, Microsoft Windows, Mac OS X	Free	[71]
15	famFLM	Region-based association analysis (functional linear models)	Command-line user interface	R	https://mga.bionet.nsc.ru/soft/famFLM/	Linux	Free	[72]
16	FarmCPU	"Fixed and random model Circulating Probability Unification" (GLM)	Command-line user interface	R	http://zzlab.net/FarmCPU/	Linux, Microsoft Windows, Mac OS X	Free	[73]
17	FaST-LMM	Factored Spectrally Transformed Linear Mixed Models	Command-line user interface, Library	Python	https://www.microsoft.com/en-us/research/project/fastlmm/	Linux, Microsoft Windows, Mac OS X	Free	[31, 33]
18	FORGE	Functional element Overlap analysis (identify tissue specific signals)	Graphical user interface, Web user interface	Perl	https://github.com/iandunham/Forge	Linux, Mac OS X	Free	[74]
19	FunciSNP	Integrate functional noncoding data sets (Identify Candidate Regulatory SNPs)	Command-line interface, library	R	https://www.bioconductor.org/packages/release/bioc/html/FunciSNP.html	Linux, Microsoft Windows, Mac OS X	–	[75]
20	G2P	Simulation tool for Genotype and phenotype data	Command-line user interface, graphical user interface	Java	https://github.com/XiaoleiLiuBio/G2P	Linux, Microsoft Windows, Mac OS X	Free	[76]
21	GAPIT	MLM, CMLM, GLM, P3D/EMMAX, SUPER, MLMM,	Command-line user interface	R	http://zzlab.net/GAPIT/	Linux, Microsoft	Free	[64]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
		FARMCPU, BLINK, GENOMIC BLUP, COMPRESSED GBLUP, SUPER GBLUP				Windows, Mac OS X		
22	garfield	Nonparametric functional enrichment analysis	Command-line user interface, Library	R	http://bioconductor.org/packages/release/bioc/html/garfield.html	Linux, Microsoft Windows, Mac OS X	Free	[77]
23	GBOOST	Gene–gene interaction analysis (Boolean operation-based screening and testing)	Graphical user interface	–	http://bioinformatics.ust.hk/BOOST.html#GBOOST	Linux, Microsoft Windows	Free	[78]
24	gboosting	Automated variable selection in survival data analysis	Library	R	https://sites.google.com/site/bestumich/issues	Linux, Microsoft Windows, Mac OS X	Free	[79]
25	gcatest	Association test for GWAS that controls for population structure under a general class of trait.	Command-line user interface, Library	R	http://bioconductor.org/packages/release/bioc/html/gcatest.html	Linux, Microsoft Windows, Mac OS X	Free	[80]
26	GCORE-sib	Fast gene–gene interaction test interaction test for discordant sib pairs	Command-line interface	C++	https://sourceforge.net/projects/gcore-sib/	Linux	Free	[81]

27	GDT	Generalized disequilibrium test (family-based)	Command-line user interface	–	http://people.virginia.edu/~wc9c/GDT/	Linux, Microsoft Windows, Mac OS X	Free	[82]
28	GEMMA	Genome-wide efficient mixed model association (ULMM, MLMM, BSLMM)	Command-line user interface	C++	http://www.xzlab.org/software.html	Linux, Mac OS X	Free	[32]
29	GenoWAP	Signal prioritization method (integrates genomic functional annotation)	Graphical user interface	Python	https://github.com/rlpowles/GenoWAP-V1.2	Microsoft Windows, Mac OS X	Free	[83]
30	GIGSEA	Genotype Imputed Gene Set Enrichment Analysis (differential gene expression)	Command-line user interface, Library	R	https://github.com/zhushijia/GIGSEA	Linux, Microsoft Windows, Mac OS X	Free	[84]
31	GISPA	"Gene Integrated Set Profile Analysis" (integrated genomic analysis)	Command-line user interface, Library	R	http://bioconductor.org/packages/release/bioc/html/GISPA.html	Linux, Microsoft Windows, Mac OS X	Free	[85]
32	GLOGS	Genome-wide LOGistic mixed model/Score test (risk covariate effects increase)	Command-line user interface	R, C	http://www.bioinformatics.org/~stanhope/GLOGS/	Linux	Free	[86]
33	GMDR	"Gene–Gene and Gene–Environment Interactions with Generalized multifactor dimensionality methods"	Graphical user interface	Java, Perl	http://ibi.zju.edu.cn/software/GMDR/download.html	Linux, Microsoft Windows, Mac OS X	Free	[87]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
34	GPA	Genetic analysis incorporating Pleiotropy and Annotation	Library	R, C++	http://dongjunchung.github.io/GPA/	Linux, Microsoft Windows, Mac OS X	Free	[88]
35	gPLINK	To perform large data sets and novel approaches	Graphical user interface	Java	http://zzz.bwh.harvard.edu/plink/gplink.shtml	Linux	Free	[89]
36	graph-GPA	A graphical model for GWAS prioritization (using a secret Markov random field method to multiple phenotypes)	Library	R, C++	https://dongjunchung.github.io/GGPA/	Linux	Free	[90]
37	GraphAT	A graph-theoretic methodology is used to evaluate the relationship between different sources (functional genomics data)	Command-line user interface, library	R	http://bioconductor.org/packages/release/bioc/html/GraphAT.html	Microsoft Windows, Mac OS X	Free	[91]
38	GWAS catalog	Publicly available, manually curated collection of published GWAS assaying at least 100k SNPs	Database portal	–	https://www.ebi.ac.uk/gwas/	Linux, Microsoft Windows, Mac OS X	Free	[92]
39	GWAS Pipeline	The pipeline method can sort, generate a kinship matrix, and generate covariate files (which can be used to run LMM, Manhattan, and QQ plots)	Command-line user interface	Python	http://www.ricediversity.org/tools/	Linux	Free	[93]

40	GWAS4D	Tissue/cell type-specific epigenetic details, transcriptional regulatory motifs, Hi-C data processing, noncoding variant functional annotations, interactive SNP aim visualization interaction	Web user interface	JavaScript, Perl	http://mulinlab.tmu.edu.cn/gwas4d	Linux, Microsoft Windows, Mac OS X	Free	[94]
41	gwascat	To represent and EMBL-EBI model	Command-line user interface, Library	R	http://bioconductor.org/packages/release/bioc/html/gwascat.html	Microsoft Windows, Mac OS X	Free	[95]
42	GWASTools	Quality control analysis and SNP annotation	Command-line user interface, Library	R	http://bioconductor.org/packages/release/bioc/html/GWASTools.html	Microsoft Windows, Mac OS X	Free	[96]
43	GWAtoolbox	Quality control and handling of multiple data files	Command-line user interface	R	http://www.eurac.edu/en/research/health/biomed/services/Pages/GWAtoolbox.aspx	Linux, Microsoft Windows, Mac OS X	Free	[97]
44	GWIZ-Rscript	Using summary-level GWAS info, produce ROC curves and calculate the AUROC.	Command-line user interface	R	https://github.com/jonaspatronjp/GWIZ-Rscript/	Linux, Microsoft Windows, Mac OS X	Free	[98]
45	HaploView	To analyze and visualize linkage disequilibrium haplotype maps	Graphical user interface	Java	https://www.broadinstitute.org/haploview/haploview	Linux, Microsoft Windows, Mac OS X	Free	[99]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
46	IGES	Integrating genotype details at the population level with summary statistics (identification of risk variants and risk prediction)	Command-line user interface	R, C++	https://github.com/daviddaigithub/IGESS	Microsoft Windows, Mac OS X	Free	[100]
47	iPat	Intelligent Prediction and Association Tool (GAPIT, PLINK, FarmCPU, BLINK, BLUP, and BGLR models)	Graphical user interface	Java	http://zzlab.net/iPat/	Linux, Microsoft Windows, Mac OS X	Free	[101]
48	IPGWAS	Quality control analysis (Manhattan plot, quantile–quantile plot, and PLINK)	Graphical user interface	Perl	https://sourceforge.net/projects/ipgwas/	Linux, Microsoft Windows, Mac OS X	Free	[102]
49	JASS	The Omnibus method and multiple weighted total of *Z*-score tests (Joint Analysis of Summary Statistics)	Command-line user interface, Web-user interface	Python	https://statistical-genetics.pages.pasteur.fr/jass/	Linux	Free	[103]
50	LAMPLINK	Cutting-edge method to detect statistically significant SNP combinations	Command-line interface	C++	http://a-terada.github.io/lamplink/	Linux	Free	[104]

51	LDSC	Estimating heritability and genetic correlation (linkage disequilibrium score regression test statistics that quantifies examining the relationship between test statistics and linkage disequilibrium)	Command-line user interface	Python	https://github.com/bulik/ldsc	Linux, Microsoft Windows, Mac OS X	Free	[105, 106]
52	lme4	Linear and generalized linear mixed-effects models can be equipped	Command-line user interface	R, Eigen C ++ library	https://cran.r-project.org/web/packages/lme4/index.html	Linux, Microsoft Windows, Mac OS X	Free	[107]
53	LocusZoom	Plotting tool (utilizes LD information from HapMap and gene information from the UCSC browser)	Web user interface	Python, R	http://locuszoom.sph.umich.edu/	Linux	Free	[108]
54	lodGWAS	Analysis of biomarkers for the limit of detection (parametric survival analysis method)	Command-line user interface	R	https://cran.r-project.org/web/packages/lodGWAS/	Linux, Microsoft Windows, Mac OS X	Free	[109]
55	MACLEAPS	To predict disease risk (based on top-validated SNPs)	Command-line user interface	Java	http://www.ra.cs.uni-tuebingen.de/software/MACLEAPS/index.htm	Linux, Microsoft Windows, Mac OS X	Free	[110]
56	Manhattan	To transpose Manhattan plot, annotate plots, and display annotations	Command-line user interface	R	https://genome.sph.umich.edu/wiki/Code_Sample:_Generating_Manhattan_Plots_in_R	Linux, Microsoft Windows, Mac OS X	Free	[111]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
57	Manhattan-Harvester	Automatically detecting and characterizing peaks (Manhattan Plots)	Command-line user interface	C++	https://genomics.ut.ee/en/tools/manhattan-harvester	Linux, Microsoft Windows, Mac OS X	Free	[112]
58	martini	Low power inherent (the vertices of the network SNPs)	Command-line user interface, library	R	http://bioconductor.org/packages/release/bioc/html/martini.html	Linux, Microsoft Windows, Mac OS X	Free	[113]
59	Matapax	Using a genome browser to display candidate regions and genes and include appropriate annotation details (GAPIT and EMMA)	Web user interface	JavaScript, R	https://matapax.mpimp-golm.mpg.de/	Linux, Microsoft Windows, Mac OS X	Free	[114]
60	metaCCA	Performs multivariate regression on a single or several studies, with genotype and phenotype representation (canonical correlation analysis)	Command-line user interface, library	R, MATLAB	http://bioconductor.org/packages/release/bioc/html/metaCCA.html	Linux, Microsoft Windows, Mac OS X	Free	[115]
61	Metal	Meta-analysis of genome-wide association scans (improving power complex traits gene mapping studies)	Command-line user interface	C++	https://genome.sph.umich.edu/wiki/METAL	Linux, Microsoft Windows, Mac OS X	Free	[116]

62	MetaSKAT	Meta-analysis methods for gene- or region-based rare variants tests (burden tests and variance component tests)	Library	R	https://www.hsph.harvard.edu/skat/metaskat/	Linux, Microsoft Windows, Mac OS X	Free	[116]
63	MSS	Maximal Segmental Score (regional empirical *P*-values to classify genomic segments and score systems-based Fisher's *P*-value combining approach to turn locus-specific importance levels into region-specific scores)	Command-line user interface	R	http://www.csjfann.ibms.sinica.edu.tw/eag/programlist/MSS/MSS.html	Microsoft Windows	Free	[117]
64	MULTIPOW	Joint and replication-based research (general multi-stage genetic association studies)	Command-line user interface	R	https://www.hsph.harvard.edu/peter-kraft/software/	Linux, Microsoft Windows, Mac OS X	Free	[118]
65	mvBIMBAM	Multivariate association analysis (Bayesian statistic approach for genetic association analysis of multiple related phenotypes)	Command-line user interface	C++	https://github.com/heejungshim/mvBIMBAM	Linux, Mac OS X	Free	[119]
66	NAM	Nested association mapping (MLM and Manhattan plot)	Command-line user interface	R	https://cran.r-project.org/web/packages/NAM/index.html	Linux, Microsoft Windows, Mac OS X	Free	[120]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
67	OmicABEL	Mixed model-based tests (involving single and multiple phenotypes)	Command-line user interface	C	https://github.com/GenABEL-Project/OmicABEL	Linux	Free	[121]
68	OrdinalGWAS	Ordered categorical phenotypes (multinomial model)	Command-line user interface	Julia	https://github.com/OpenMendel/OrdinalGWAS.jl	Linux	Free	[122]
69	PAPA	A set of pleiotropic pathways analysis	Command-line user interface	R, C	https://sourceforge.net/projects/papav1/	Linux, Mac OS X	Free	[123]
70	PARIS	Pathway Analysis by Randomization Incorporating Structure (unique methodology and Multiple testing)	Command-line user interface	C++	https://ritchielab.org/software/paris-download	Linux	Free	[124]
71	Pascal	Pathway scoring algorithm (new gene-based analysis method MAGMA or VEGAS)	Command-line user interface	Pascal	https://www2.unil.ch/cbg/index.php?title=Pascal	Linux, Mac OS X	Free	[125]
72	PC-select	FaST-LMM and principal components analysis	Command-line user interface	MATLAB	https://github.com/gjtucker/PC-Select	Linux	Free	[126]
73	PEPIS	Pipeline for calculating EPIStatic hereditary impact (new linear mixed model and used to predict the performance of hybrid rice)	Web user interface	R, C++, C, Fortran	http://bioinfo.noble.org/PolyGenic_QTL/	Linux, Microsoft Windows, Mac OS X	Free	[127]

74	PExFInS	Post-GWAS Explorer for Functional Indels and SNPs (to analyze functional insertions, indels, LD, and SNPs)	Command-line user interface	Perl, SAS, Java	https://sourceforge.net/projects/pexfins/	Linux, Microsoft Windows, Mac OS X	Free	[128]
75	PheGWAS	To analyze many variants and one phenotype (visualization of Manhattan plots in 3D landscape)	Command-line user interface	R	https://github.com/georgeg0/PheGWAS	Linux, Microsoft Windows, Mac OS X	Free	[129]
76	PheWAS	Phenome-wide association studies (to discover many genotype–phenotype relationships)	Library	R	https://github.com/PheWAS/PheWAS	Linux, Microsoft Windows, Mac OS X	Free	[130]
77	PLATO	Data Analysis, Translation, and Organization Platform on a Large Scale (detection of gene-gene interactions)	Command-line user interface	C++	https://ritchielab.org/software/plato-download	Linux	Free	[131]
78	PLINK	Popular and well-documented method (for data management, summary statistics, population stratification, association analysis, and estimating identification by descent).	Command-line user interface	C++, C	http://zzz.bwh.harvard.edu/plink/	Linux, Mac OS X	Free	[89, 132]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
79	powerGWASinteraction	To identify gene-gene and gene-environment interactions	Library	R	http://kooperberg.fhcrc.org/soft.html	Linux, Microsoft Windows, Mac OS X	Free	[133]
80	PSESM	SNP main effects and SNP-SNP interaction effects by using SNP genotyping data (Pseudo standard error method)	Command-line interface	Fortran 90	https://www.shinfu.idv.tw/software	Microsoft Windows, Mac OS X	Free	[134]
81	QCGWAS	Quality control analysis (automated and manual)	Command-line user interface	R	https://cran.r-project.org/web/packages/QCGWAS/	Linux, Microsoft Windows, Mac OS X	Free	[135]
82	QCTOOL	Manipulation and quality control (genetic risk predictor scores, LD between variants, between-sample relatedness, and principal components)	Command-line user interface	C++	https://www.well.ox.ac.uk/~gav/qctool/#overview	Linux	Free	[136]
83	qMSAT	Quality-based Multivariate Score Association Test (two functional areas, MGLL promoter and MGLL 3′-untranslated region)	Script	R	http://qmsat.sourceforge.net/	Linux	Free	[137]

84	QTDT	Quantitative transmission Disequilibrium tests (Linkage disequilibrium mapping of quantitative traits)	Command-line user interface	C++	http://csg.sph.umich.edu/abecasis/qtdt/	Linux, Microsoft Windows, Mac OS X	Free	[138]
85	RAISS	Robust and accurate imputation from summary statistics (suitable for multi-trait analyses)	Command-line user interface	Python	https://gitlab.pasteur.fr/statistical-genetics/raiss	Linux	Free	[139]
86	RaMWAS	To analyze methylome-wide associations (multi-marker analysis and joint analysis of methylation and genotype data)	Command-line user interface, library	R	http://bioconductor.org/packages/release/bioc/html/ramwas.html	Linux, Microsoft Windows, Mac OS X	Free	[140]
87	regioneR	Permutation test to assess the association between genomic region sets (local specificity of the detected association)	Command-line user interface, library	R	http://bioconductor.org/packages/release/bioc/html/regioneR.html	Linux, Microsoft Windows, Mac OS X	Free	[141]
88	RegScan	Association analysis of combinatorial traits in metabolomics (can analyze multiple traits simultaneously and very fast)	Command-line user interface	C++	https://genomics.ut.ee/en/tools/regscan	Linux, Microsoft Windows, Mac OS X	Free	[142]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
89	rqt	Gene-level meta-analysis (Single and Multiple datasets)	Command-line user interface, Library	R	https://github.com/izhbannikov/rqt	Linux, Microsoft Windows, Mac OS X	Free	[143]
90	Scoary	Pan-genome associations to observed phenotypic traits (genes sorted by strength of association per trait)	Command-line user interface	Python	https://github.com/AdmiralenOla/Scoary	Linux	Free	[144]
91	seq2pathway	Pathway analysis of next-generation sequencing data (seq2gene and gene2path)	Command-line user interface, Library	R, Python	http://bioconductor.org/packages/release/bioc/html/seq2pathway.html	Linux, Microsoft Windows, Mac OS X	Free	[145]
92	SEQPower	Sequence-based association data from a number of genetic variation and disease phenotype models (LOGIT, PAR, LNR, ELNR, and BLNR models)	Command-line user interface	C++, Python	http://bioinformatics.org/spower/	Linux, Mac OS X	Free	[146]
93	SIMreg	A similarity-based regression approach for testing gene-environment interactions for quantitative and binary traits.	Command-line user interface	R, C	https://www4.stat.ncsu.edu/~jytzeng/software.php	Linux, Microsoft Windows, Mac OS X	Free	[147]

94	SKAT	SNP-set a gene or region level test (a set of rare or common variants and dichotomous or quantitative phenotypes). Sequence Kernel Association Tests	Command-line user interface, library	R	https://www.hsph.harvard.edu/skat/	Linux, Microsoft Windows, Mac OS X	Free	[148]
95	SNPEVG	A tool for instant global and local viewing and graphing (it includes three programs SNPEVG1, SNPEVG2, and SNPEVG3)	Desktop graphical user interface	–	https://animalgene.umn.edu/snpevg	Microsoft Windows, Mac OS X	Free	[149]
96	snpGeneSets	To make annotation and review more straightforward (genomic mapping annotation for SNPs and gene sets, bidirectional mapping between SNPs and genes sets, and gene effect measures from SNP associations and performance of gene set enrichment analyses to identify functional pathways)	Command-line user interface	R	https://www.umc.edu/SoPH/Departments-and-Faculty/Data-Science/Research/Services/Software.html	Microsoft Windows	Free	[150]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
97	SNPRelate	Relatedness and Principal Component Analysis using Identity by Descent measures	Command-line user interface, library	R	http://bioconductor.org/packages/release/bioc/html/SNPRelate.html	Linux, Microsoft Windows, Mac OS X	Free	[151]
98	SNPStats	Extending snpMatrix to account for genotype instability (Fst, Imputation and meta-analysis, LD, Principal components analysis and Snpmatrix differences)	Command-line interface	R	http://bioconductor.org/packages/release/bioc/html/snpStats.html	Linux, Microsoft Windows, Mac OS X	Free	[152]
99	SNPsyn	Exploration and discovery of synergistic pairs of SNPs (subsequent components and visualization)	Web user interface	C++, Flash	http://snpsyn.biolab.si/	Linux, Microsoft Windows, Mac OS X	Free	[153]
100	SNPTEST	To examine a specific SNP (Binary and several measurable phenotypes, Bayesian and Frequentist measures, an arbitrary collection of covariates and SNPs, and various SNP methods)	Command-line user interface	JavaScript, C++	https://mathgen.stats.ox.ac.uk/genetics_software/snptest/snptest.html	Linux, Mac OS X	Free	[154]

101	SOLAR-eclipse	Genetic variance components analysis, LD, quantitative genetic analysis, SNP association analysis (QTN and QTLD), covariate screening, market-specific analysis, and mega and meta-genetic analyses)	Command-line user interface	C++, Fortran, C	http://solar-eclipse-genetics.org/	Linux	Free	[155]
102	SSEA	SNP Set Enrichment Analysis (SNP genotype and annotation to generate Significant SNPs and together with pathway)	Graphical user interface	Java	https://cbcl.ics.uci.edu/SSEA/	Linux, Microsoft Windows, Mac OS X	Free	[156]
103	STEGO	Similarity Test for Estimating Genetic Outliers (population sub-structure or cryptic relationships)	Command-line user interface	R	https://github.com/dschlauch/stego	Linux, Microsoft Windows, Mac OS X	Free	[157]
104	SurvivalGWAS_SV	Weibull regression model study of genotypes by modelling time to event effects in a dose model. It may account for several covariates and add SNP covariate interaction results	Command-line interface	C#	https://www.liverpool.ac.uk/translational-medicine/research/statistical-genetics/survival-gwas-sv/	Linux, Microsoft Windows, Mac OS X	Free	[158]

(continued)

Table 19.1 (continued)

Sl. No.	Softwares and tools	Description	Software interface	Language	Link	Operating system	Cost	References
		(PLINK and SNPTEST)						
105	TASSEL	Trait Analysis by Association, Evolution, and Linkage (Unified Mixed-Model, GLM, MLM, LD, deletions, diversity statistics, integration of phenotypic and genotypic data, imputing missing data and principal components)	Command-line interface, Desktop graphical user interface	Java	https://www.maizegenetics.net/tassel	Linux, Microsoft Windows, Mac OS X	Free	[47]
106	traseR	Enrichment analyses of trait-associated SNPs in arbitrary genomic intervals	Command-line interface	R	http://bioconductor.org/packages/release/bioc/html/traseR.html	Linux, Microsoft Windows, Mac OS X	Free	[159]
107	treeWAS	A phylogenetic method, clonal population structure, and homologous recombination (cluster-based and dimension-reduction methods)	Command-line user interface, Library	R	https://github.com/caitiecollins/treeWAS	Linux, Microsoft Windows, Mac OS X	Free	[160]
108	Variant Ranker	To identify causal genes (ranking, Case-	Web user interface	R, PHP, JavaScript		Linux, Microsoft	Free	[161]

		Control, and custom Filtering, annotation of coding and noncoding variants, network visualization, and cluster of genes)			http://paschou-lab.mbg.duth.gr/html5up/index.html	Windows, Mac OS X		
109	VSEAMS	Variant set enrichment analysis using multivariate sampling (nonparametric SNP set enrichment method)	Command-line user interface	Perl, R	https://github.com/ollyburren/vseams/wiki	Linux, Mac OS X	Free	[162]
110	wtest	Genetic epistasis monitoring (cis-regulation of SNP and CpG sites and epigenome, post-test diagnostic study, *W*-test, and minor allele frequency)	Command-line user interface	R, C++	https://cran.r-project.org/web/packages/wtest/index.html	Linux, Microsoft Windows, Mac OS X	Free	[163]
111	XGR	eXploring Genomic relations (interpretation of genomic summary, eQTL, SNP-modulated gene networks, and pathways)	Command-line user interface, Web user interface	R	http://xgr.r-forge.r-project.org/	Linux, Microsoft Windows, Mac OS X	Free	[164]

only QTLs at which the two parents differ can be observed in a single segregating mapping population. [174].

To compensate for these disadvantages, a new method have been developed and deployed, known as "Nested association mapping" (NAM) which was established in maize to allow high power and high-resolution mapping via joint "linkage-association research" [175]. Therefore, the NAM population was created by the crossing of 25 diverse inbred lines of maize with "B73" reference lines to produce 5000 RILs population. Because of the fact that the "NAM population" possesses several important characteristics and features, it has been used for large-scale genetic mapping [176–179]. Similarly, Arabidopsis was used as a model plant for the generation of the "Multi-parent Advanced Generation Inter-Cross" (MAGIC) population, which consisted of hundreds of RILs descended from a heterogeneous stock of 19 intermated Arabidopsis accessions [180]. Thus the "MAGIC population" was used in a computational simulation study to identify links; QTLs show 10% phenotypic variation that can be described with an average mapping error of approximately 300 kb. Further, another group was crossed with eight Arabidopsis accessions to produce a set of six RIL populations known as "Arabidopsis multiparent RIL" (AMPRIL) [181]. Thus, the AMPRIL population was used for QTL analysis and revealed that genetic resources could detect QTLs that explained 2% or more of a trait's variance.

There is some difference between the GWAS of rice and maize. The difference arose because of the compromise in power and mapping resolution between "self-pollinated" and cross-pollinated species. Due to "self-fertilization" and small effective population size, The LD decay in the rice genome is at ~100 kb. Similarly, in other "self-pollinated" crops like foxtail millet and soybean, the LD decay rate is slower. Because of extended LD, sequencing with low genome coverage and missing data is quite efficient in conducting successful GWAS in rice. However, since rice's LD decay rate is slower, GWAS cannot address a single gene in rice. In contrast to rice, the maize crop is an outcrossing species because of this difference in flowering time of male and female inflorescence. Therefore, the LD decay rate is rapid in maize within ~2 kb, and it contributes to greater genetic diversity with higher resolution potential in GWAS. Almost all experiments on maize have resolved at the single-gene stage using GWAS. To achieve greater resolution, the GWAS in maize needs tens of millions of SNPs to correctly genotype various varieties, which is a difficult and expensive challenge given the genome's wide-scale abundance of repeats and paralog sequences.

However, in the genetic mapping of basic qualitative traits or mutant mapping, "Bulk segregation analysis" along with "multiple-sample pooling sequencing" may be an alternative to traditional "linkage analysis." For this application, various methods and protocols have been published, including "SHOREmap" (98), "MutMap" [169], "next-generation mapping" [182], and "MutMap-Gap" [170] (105). To prevent possible intervention from diverse genetic origins, James et al. [183] have described the populations produced by backcrossing a mutant line to a nonmutagenized parent for sequencing mapping. In rice, a recessive mutation line was crossed to the parental line used for mutagenesis, and the mutant F2 progeny produced were pooled for sequence mutation analysis [169]. However, this technique and strategy may be used for quantitative trait identification in crops (e.g., Sorting the RILs into distinct sequencing pools based on a specific trait).

19.6 Genome-Wide Association Studies in Crops

The association mapping shows that markers and traits vary because the environmental factors influences the genotype performance through alterations in phenotypes. As a result, the association study lays the groundwork for long-term cumulative impacts to identifying new QTLs, alleles, and genes [184]. The mixed model (Q+K) demonstrated a substantial increase in the goodness of fit in GWAS. The K and Q matrix corrected the correlation between phenotypic traits with QTLs linked to markers [185, 186] at $P < 0.005$ (GLM and MLM) and $P < 0.001$ (FaST). This corresponds to prior studies [35, 186–188]. Some studies have published their findings at a $P < 0.05$–0.01 level of significance, showing that the number of markers is significantly higher [189, 190]. Therefore, the *p*-value plays a significant role in association studies because it influences the degree of false-positive association between traits and markers. It implies that lowering the *p*-value reduces or eliminates the possibility of a false-positive relation [34].

Since the past decade, many GWAS has been conducted successfully in many crops [177, 179, 184, 191–195]. Among all the crops, maximum GWAS has been carried out in maize and rice, producing a vast magnitude of both phenotype and genotype data in multiple environments. In a study, a sample of 1083 cultivated *O. sativa* var. *indica* and *japonica* varieties of 446 wild (*Oryza rufipogon*) rice accessions genome sequencing with low genome coverage was done [196], from which a high-density haplotype map was constructed. More than 1.3 million SNPs were used in the GWAS to detect alleles related to flowering time and ten grain-related traits. Some of the associations match with the previously reported genes. However, the GWAS carried out for leaf sheath color and tiller angle, taking 446 *O. rufipogon* accessions, revealed a higher level of genetic diversity in wild species. In another study, 44,100 SNPs variants in 413 diverse rice accessions revealed the complex hereditary architecture of 34 traits in rice.

Multiple candidate genes for phenotypic traits like grain yield, seed quality, leaf angle, flowering time, leaf size, and disease resistance were identified in a GWAS of maize. Hence, it revealed the genetic heredity and architecture of these traits controlled by multiple QTLs with small effects [176–179]. In another study, Li and the team reported that 368 maize lines were analyzed with over one million genome-wide SNPs to characterize maize kernel and identify QTLs linked with oil composition [193]. They recorded 74 loci associated with the concentration of maize kernel oil and fatty acid composition.

GWAS has also been conducted successfully in other crops. A study that sequenced 916 diverse fox millet varieties with low genome coverage reported several loci for ten agronomic traits tested in five distinct ecosystems [192]. Another research found 0.2 million SNPs in 917 globally diverse accessions and identified many previously identified loci correlated with plant height and inflorescence architecture using GWAS [194]. Despite the low marker density, a panel of 224 spring barley accessions genotyped using a genotyping microarray at 957 SNP sites detected some important candidate genes [62]. In another study, GWAS was carried out in 615 barley cultivars with a very low density of SNPs for 32 morphologic and ten agronomic traits [197]. Transcriptomics-based identification of SNPs from mRNA and its applicability in GWAS is tested in polyploidy crops. Through this investigation, two QTLs with genomic deletions responsible for glucosinolate content of seeds have been detected, leading to the identification of a candidate gene/transcription factor HAG1 [198].

Attempts have been made to perform genome-wide association studies (GWAS) on bread wheat, a polyploid crop with a large genome. GWAS study in "*Triticum Urartu,*" an ancestor species of bread wheat genome has also been made [199]. However, implementing GWAS in wheat is technically difficult and needs great effort to overcome challenges. These experiments demonstrate that GWAS is a versatile strategy capable of mapping several traits genetically simultaneously. However, there is a need to explore further the genetic basis of important agronomic, morphological and physiological traits in other species, which are close to wild relatives of the cultivated crops. Additionally, care should be taken during GWAS for the population structure and the balance between the increased "false-negatives" and decreased "false-positive" rates [26, 185–187]. The highly popular method of GWAS is the "mixed model" (MLM), which was used to detect the genotype–phenotype association in crops [27, 182]. However, this model takes a longer time for computation while analyzing a large population (~1000 individuals), including enormous markers (~1 million SNPs markers).

Similarly, several scientists have reported that the use of the "Efficient Mixed-Model Association eXpedited" (EMMAX) program and the "Compressed Mixed Linear Model Method" has substantially reduced the computation time [29–32, 64, 188]. In addition, "GWAPP" is another web-based application or model used for GWAS, which used a "Linear Mixed Model" known as the "accelerated mixed model" in *A. thaliana* through which SNP detection is done along with population structure [200]. Other techniques such as multiple regression and nonparametric statistics are also used for this purpose [201, 202]. GWAS has less capability to detect rare alleles, which make up a major portion of the natural variation. In rice, low-frequency SNPs make up approximately 44% of the total SNPs (minor allele frequency <0.05). However, rare alleles may be identified by constructing several bi-parental cross populations from large populations (e.g., NAM or MAGIC).

In most cases, only one gene among the several within the GWAS locus contributes to the QTL, which necessitates further analysis through gene annotation and expression profile to identify the causal gene correctly. For example, the causal gene for disease traits identified in GWAS loci contains leucine-rich repeats in their binding sites, and the genes expressed at the grain filling stage are related to the grain-related traits. Similarly, T-DNA mutants, artificial induction of mutations, and analyzing candidate genes through TILLING ("Targeting Induced Local Lesions In Genomes") are efficient methods for validating gene–trait correlations. But through transgenic analysis, the causal genes and their variants can be conclusively identified. More information will be obtained in GWAS analysis, taking diverse germplasm panels, careful evaluation of traits, and through more functional trials that will help to address the persisting biological questions.

GWAS approach has been widely used in rice to identify different QTLs/genes for traits like agromorphologicl traits, yield-related traits, biotic abiotic stresses, Fe, Zn, quality traits, and early seedling vigour [166, 203–212]**.** Several researchers have been reported viz., amylose contents [203], grain yield [166], deep root mass and the number of deep roots [204], grain quality traits [205], seed vigor [206], agronomic traits [207, 208], plant height and grain yield [209], cold tolerance at germination and booting stages [210], salinity tolerance [211], early seedling vigor [213], seedling stage chilling stress [214], grain yield under water deficit [191], grain yield under reproductive drought stress [215], panicle architecture and spikelet's/panicle [216], and salt tolerance in rice [217]. Chilling tolerance study was reported by Schlappi who identified two new tolerance QTLs for low temperature at seedling stage (LTSS)–QTL, *qLTSS3-4* and *qLTSS4-1*, [218], Bollinedi and his team reported 26 QTLs for Fe and Zn localization through GWAS study [219]. Kumar and the team identified QTLs for Fe, Zn, β-carotene, GPC and yield traits in bread wheat using multi-locus and multi-traits GWAS approach [220]. The yield contributing traits would help to break the yield ceiling greatly aided by genes/QTLs that produce high grain yield even under various stress. Furthermore, it would be useful to identify traits-specific donors for designing an effective breeding strategy for crop production [166]. Thus, this book chapter aims to describe the status and prospective of genome-wide association studies in plants.

19.7 Perspectives and Conclusion

The genome-wide association study was first applied in human genetics to unravel the genetic basis of complex medical traits. Later it was applied to model species like Arabidopsis, rice, and maize [185, 221–228]. GWAS has become a powerful tool in studying complex traits in rice and other crops due to advancements in high-throughput sequencing technology and linear mixed model. In contrast to human GWAS, rice takes advantage of being a self-pollinated crop; it can be genotyped once but can be phenotyped multiple times for different traits in multiple environments. Thousands of loci associated with different agronomic and physiological traits are identified in rice using GWAS, and several statistical methods are also developed to improve the computation time. Though many GWAS have been successfully conducted in rice, there are still new challenges [194] like epistatic interactions and G × E interactions important for quantitative traits, which are generally encountered while doing GWAS in rice [224, 225]. Novel statistical methods and experimental designs need to be addressed in the future for these interactions. To date, most of the population used in rice GWAS includes temperate japonica and indica. However, there is a need to include tropical japonica, aus, basmati, and wild relatives in the GWAS panel to address the critical variants and associations present in these collections. The loci identified through GWAS need to be validated through gene annotation and expression analysis. There is also a need to find out the coordinated synergistic relationship between GWAS and genome editing. GWAS identifies the underlying gene for the traits providing targets for genome editing; in return, genome editing helps validate gene function. Ultimately, complementing GWAS with different genomic and phenomic technologies will help in understanding the biological functions of specific alleles involved in biotic and abiotic stress tolerance, improved nutritional quality, and increased grain yield, subsequently improving crop breeding and accelerating genomic-assisted crop breeding.

Conflict of Interest The author does not have any conflict of interest.

References

1. Godfray HCJ, Beddington JR, Crute IR, Haddad L, Lawrence D, Muir JF, et al. Food security: the challenge of feeding 9 billion people. Science. 2010;327:812–8.
2. Tester M, Langridge P. Breeding technologies to increase crop production in a changing world. Science. 2010;327:818–22.
3. Zhang Q. Strategies for developing Green Super Rice. Proc Natl Acad Sci U S A. 2007;104:16402–9.
4. Huang X, Feng Q, Qian Q, Zhao Q, Wang L, Wang A, et al. High-throughput genotyping by whole-genome resequencing. Genome Res. 2009;19:1068–76.
5. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, et al. A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One. 2011;6: e19379.
6. Bevan MW, Uauy C. Genomics reveals new landscapes for crop improvement. Genome Biol. 2013;14:206.
7. Hamblin MT, Buckler ES, Jannink J-L. Population genetics of genomics-based crop improvement methods. Trends Genet. 2011;27:98–106.
8. Zhu C, Gore M, Buckler ES, Yu J. Status and prospects of association mapping in plants. Plant Genome. 2008;1. https://acsess.onlinelibrary.wiley.com/doi/abs/10.3835/plantgenome2008.02.0089.
9. Michael TP, Jackson S. The first 50 plant genomes. Plant Genome. 2013;6. https://onlinelibrary.wiley.com/doi/10.3835/plantgenome2013.03.0001in.
10. Sukumaran S, Yu J. Association mapping of genetic resources: achievements and future perspectives. In: Tuberosa R, Graner A, Frison E, editors. Genomics of plant genetic resources, vol.1: Managing, sequencing and mining genetic resources. Dordrecht: Springer; 2014. p. 207–35. https://doi.org/10.1007/978-94-007-7572-5_9.
11. Visscher PM, Brown MA, McCarthy MI, Yang J. Five years of GWAS discovery. Am J Hum Genet. 2012;90:7–24.
12. Xiao Y, Liu H, Wu L, Warburton M, Yan J. Genome-wide association studies in maize: praise and stargaze. Mol Plant. 2017;10:359–74.
13. Takeda S, Matsuoka M. Genetic approaches to crop improvement: responding to environmental and population changes. Nat Rev Genet. 2008;9:444–57.
14. Mackay TFC, Stone EA, Ayroles JF. The genetics of quantitative traits: challenges and prospects. Nat Rev Genet. 2009;10:565–77.

15. Ersoz ES, Yu J, Buckler ES. Applications of linkage disequilibrium and association mapping in crop plants. In: Varshney RK, Tuberosa R, editors. Genomics-assisted crop improvement, vol. 1: Genomics approaches and platforms. Dordrecht: Springer; 2007. p. 97–119. https://doi.org/10.1007/978-1-4020-6295-7_5.
16. Liu H-J, Yan J. Crop genome-wide association study: a harvest of biological relevance. Plant J. 2019;97:8–18.
17. Varshney RK, Ribaut J-M, Buckler ES, Tuberosa R, Rafalski JA, Langridge P. Can genomics boost productivity of orphan crops? Nat Biotechnol. 2012;30:1172–6.
18. Gupta PK, Kulwal PL, Jaiswal V. Chapter two - Association mapping in plants in the post-GWAS genomics era. In: Kumar D, editor. Advances in genetics. Boston: Academic; 2019. p. 75–154. https://www.sciencedirect.com/science/article/pii/S0065266018300385.
19. Chen E, Huang X, Tian Z, Wing RA, Han B. The genomics of *Oryza* species provides insights into rice domestication and heterosis. Annu Rev Plant Biol. 2019;70:639–65.
20. Chen W, Wang W, Peng M, Gong L, Gao Y, Wan J, et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. Nat Commun. 2016;7:12767.
21. Zhou Y, Srinivasan S, Mirnezami SV, Kusmec A, Fu Q, Attigala L, et al. Semiautomated feature extraction from RGB images for sorghum panicle architecture GWAS. Plant Physiol. 2019;179:24–37.
22. Spindel JE, Dahlberg J, Colgan M, Hollingsworth J, Sievert J, Staggenborg SH, et al. Association mapping by aerial drone reveals 213 genetic associations for *Sorghum bicolor* biomass traits under drought. BMC Genomics. 2018;19:679.
23. Wang S-B, Feng J-Y, Ren W-L, Huang B, Zhou L, Wen Y-J, et al. Improving power and accuracy of genome-wide association studies via a multi-locus mixed linear model methodology. Sci Rep. 2016;6:19444.
24. Zhang Y, Massel K, Godwin ID, Gao C. Applications and potential of genome editing in crop improvement. Genome Biol. 2018;19:210.
25. Hansen TF. The evolution of genetic architecture. Annu Rev Ecol Evol Syst. 2006;37:123–57.
26. Korte A, Farlow A. The advantages and limitations of trait analysis with GWAS: a review. Plant Methods. 2013;9:29.
27. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet. 2006;38:203–8.
28. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, et al. Efficient control of population structure in model organism association mapping. Genetics. 2008;178:1709–23.
29. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. Mixed linear model approach adapted for genome-wide association studies. Nat Genet. 2010;42:355–60.
30. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, et al. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42:348–54.
31. Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. FaST linear mixed models for genome-wide association studies. Nat Methods. 2011;8:833–5.
32. Zhou X, Stephens M. Genome-wide efficient mixed-model analysis for association studies. Nat Genet. 2012;44:821–4.
33. Listgarten J, Lippert C, Kadie CM, Davidson RI, Eskin E, Heckerman D. Improved linear mixed models for genome-wide association studies. Nat Methods. 2012;9:525–6.
34. Wang Q, Tian F, Pan Y, Buckler ES, Zhang Z. A SUPER powerful method for genome wide association study. PLoS One. 2014;9:e107684.
35. Abdurakhmonov IY, Abdukarimov A. Application of association mapping to understanding the genetic diversity of plant germplasm resources. Int J Plant Genomics. 2008;2008:1–18.
36. Zhu C, Yu J. Nonmetric multidimensional scaling corrects for population structure in association mapping with different sample types. Genetics. 2009;182:875–88.
37. Xu S. Theoretical basis of the Beavis effect. Genetics. 2003;165:2259–68.

38. Sallam A, Martsch R. Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (*Vicia faba* L.). Genetica. 2015;143:501–14.
39. Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ, et al. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice. 2013;6:11.
40. Paterson AH, Lander ES, Hewitt JD, Peterson S, Lincoln SE, Tanksley SD. Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. Nature. 1988;335:721–6.
41. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. Nat Genet. 2002;32:650–4.
42. Lander ES, Schork NJ. Genetic dissection of complex traits. Science. 1994;265:2037–48.
43. Risch N, Merikangas K. The future of genetic studies of complex human diseases. Science. 1996;273:1516–7.
44. Thornsberry JM, Goodman MM, Doebley J, Kresovich S, Nielsen D, Buckler ES. *Dwarf8* polymorphisms associate with variation in flowering time. Nat Genet. 2001;28:286–9.
45. Alqudah AM, Koppolu R, Wolde GM, Graner A, Schnurbusch T. The genetic architecture of barley plant stature. Front Genet. 2016;7. https://www.frontiersin.org/articles/10.3389/fgene.2016.00117/full
46. Alqudah AM, Youssef HM, Graner A, Schnurbusch T. Natural variation and genetic make-up of leaf blade area in spring barley. Theor Appl Genet. 2018;131:873–86.
47. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.
48. Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet. 1995;11:241–7.
49. Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucl Acids Res. 2001;29:308–11.
50. Aranzana MJ, Kim S, Zhao K, Bakker E, Horton M, Jakob K, et al. Genome-wide association mapping in Arabidopsis identifies previously known flowering time and pathogen resistance genes. PLoS Genet. 2005;1:e60.
51. Bush WS, Moore JH. Chapter 11: genome-wide association studies. PLoS Comput Biol. 2012;8:e1002822.
52. Digel B, Tavakol E, Verderio G, Tondelli A, Xu X, Cattivelli L, et al. Photoperiod-H1 (Ppd-H1) controls leaf size. Plant Physiol. 2016;172:405–15.
53. Milner SG, Jost M, Taketa S, Mazón ER, Himmelbach A, Oppermann M, et al. Genebank genomics highlights the diversity of a global barley collection. Nat Genet. 2019;51:319–26.
54. Alomari DZ, Eggert K, von Wirén N, Alqudah AM, Polley A, Plieske J, et al. Identifying candidate genes for enhancing grain Zn concentration in wheat. Front Plant Sci. 2018;9. https://www.frontiersin.org/articles/10.3389/fpls.2018.01313/full
55. Camiolo S, Sablok G, Porceddu A. Altools: a user friendly NGS data analyser. Biol Direct. 2016;11:8.
56. Kwak I-Y, Pan W. Adaptive gene- and pathway-trait association testing with GWAS summary statistics. Bioinformatics. 2016;32:1178–84.
57. Günther T, Coop G. Robust identification of local adaptation from allele frequencies. Genetics. 2013;195:205–20.
58. Westreich ST, Nattestad M, Meyer C. BigTop: a three-dimensional virtual reality tool for GWAS visualization. bioRxiv. 2019:650176.
59. Storey JD, Tibshirani R. Statistical significance for genomewide studies. Proc Natl Acad Sci U S A. 2003;100:9440–5.

60. Moore CB, Wallace JR, Frase AT, Pendergrass SA, Ritchie MD. BioBin: a bioinformatics tool for automating the binning of rare variants using publicly available biological knowledge. BMC Med Genomics. 2013;6:S6.
61. Huang M, Liu X, Zhou Y, Summers RM, Zhang Z. BLINK: a package for the next level of genome-wide association studies with both individuals and markers in the millions. GigaScience. 2019;8:giy154. https://doi.org/10.1093/gigascience/giy154.
62. Pasam RK, Sharma R, Malosetti M, van Eeuwijk FA, Haseneyer G, Kilian B, et al. Genome-wide association studies for agronomical traits in a world wide spring barley collection. BMC Plant Biol. 2012;12:16.
63. Rentería ME, Cortes A, Medland SE. Using PLINK for genome-wide association studies (GWAS) and data analysis. In: Gondro C, van der Werf J, Hayes B, editors. Genome-wide association studies and genomic prediction. Totowa: Humana Press; 2013. p. 193–213. https://doi.org/10.1007/978-1-62703-447-0_8.
64. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ, et al. GAPIT: genome association and prediction integrated tool. Bioinformatics. 2012;28:2397–9.
65. Hickey JM, Gorjanc G. Simulated data for genomic selection and Genome-Wide Association Studies Using a Combination of Coalescent and Gene Drop Methods. G3 Genes Genomes Genet. 2012;2:425–7.
66. Zhao J, Ming J, Hu X, Chen G, Liu J, Yang C. Bayesian weighted Mendelian randomization for causal inference based on summary statistics. Bioinformatics. 2020;36:1501–8.
67. Yao Y, Liu Z, Wei Q, Ramsey SA. CERENKOV2: improved detection of functional noncoding SNPs using data-space geometric features. BMC Bioinformatics. 2019;20:63.
68. Millstein J, Chen GK, Breton CV. cit: hypothesis testing software for mediation analysis in genomic applications. Bioinformatics. 2016;32:2364–5.
69. Majumdar A, Haldar T, Bhattacharya S, Witte JS. An efficient Bayesian meta-analysis approach for studying cross-phenotype genetic associations. PLoS Genet. 2018;14:e1007139.
70. Arkin Y, Rahmani E, Kleber ME, Laaksonen R, März W, Halperin E. EPIQ—efficient detection of SNP–SNP epistatic interactions for quantitative traits. Bioinformatics. 2014;30: i19–25.
71. Morris JA, Randall JC, Maller JB, Barrett JC. Evoker: a visualization tool for genotype intensity data. Bioinformatics. 2010;26:1786–7.
72. Svishcheva GR, Belonogova NM, Axenovich TI. Region-based association test for familial data under functional linear models. PLoS One. 2015;10:e0128999.
73. Liu X, Huang M, Fan B, Buckler ES, Zhang Z. Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet. 2016;12: e1005767.
74. Dunham I, Kulesha E, Iotchkova V, Morganella S, Birney E. FORGE: a tool to discover cell specific enrichments of GWAS associated SNPs in regulatory regions. F1000Res. 2015;4:18.
75. Coetzee SG, Rhie SK, Berman BP, Coetzee GA, Noushmehr H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucl Acids Res. 2012;40:e139.
76. Tang Y, Liu X. G2P: a Genome-Wide-Association-Study simulation tool for genotype simulation, phenotype simulation and power evaluation. Bioinformatics. 2019;35:3852–4.
77. Iotchkova V, Ritchie GRS, Geihs M, Morganella S, Min JL, Walter K, et al. GARFIELD - GWAS analysis of regulatory or functional information enrichment with LD correction. bioRxiv. 2016:085738.
78. Yung LS, Yang C, Wan X, Yu W. GBOOST: a GPU-based tool for detecting gene–gene interactions in genome–wide case control studies. Bioinformatics. 2011;27:1309–10.
79. He K, Li Y, Zhu J, Liu H, Lee JE, Amos CI, et al. Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. Bioinformatics. 2016;32:50–7.
80. Song M, Hao W, Storey JD. Testing for genetic associations in arbitrarily structured populations. Nat Genet. 2015;47:550–4.

81. Sung P-Y, Wang Y-T, Hsiung CA, Chung R-H. GCORE-sib: an efficient gene-gene interaction tool for genome-wide association studies based on discordant sib pairs. BMC Bioinformatics. 2016;17:273.
82. Chen W-M, Manichaikul A, Rich SS. A generalized family-based association test for dichotomous traits. Am J Hum Genet. 2009;85:364–76.
83. Lu Q, Yao X, Hu Y, Zhao H. GenoWAP: GWAS signal prioritization through integrated analysis of genomic functional annotation. Bioinformatics. 2016;32:542–8.
84. Zhu S, Qian T, Hoshida Y, Shen Y, Yu J, Hao K. GIGSEA: genotype imputed gene set enrichment analysis using GWAS summary level data. Bioinformatics. 2019;35:160–3.
85. Kowalski J, Dwivedi B, Newman S, Switchenko JM, Pauly R, Gutman DA, et al. Gene integrated set profile analysis: a context-based approach for inferring biological endpoints. Nucl Acids Res. 2016;44:e69.
86. Stanhope SA, Abney M. GLOGS: a fast and powerful method for GWAS of binary traits with risk covariates in related populations. Bioinformatics. 2012;28:1553–4.
87. Hai-Ming X, Li-Feng X, Ting-Ting H, Lin-Feng L, Guo-Bo C, Xi-Wei S, et al. GMDR: versatile software for detecting gene-gene and gene-environment interactions underlying complex traits. Curr Genomics. 2016;17:396–402.
88. Chung D, Yang C, Li C, Gelernter J, Zhao H. GPA: a statistical approach to prioritizing GWAS results by integrating pleiotropy and annotation. PLoS Genet. 2014;10:e1004787.
89. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559–75.
90. Chung D, Kim HJ, Zhao H. graph-GPA: a graphical model for prioritizing GWAS results and investigating pleiotropic architecture. PLoS Comput Biol. 2017;13:e1005388.
91. Balasubramanian R, LaFramboise T, Scholtens D, Gentleman R. A graph-theoretic approach to testing associations between disparate sources of functional genomics data. Bioinformatics. 2004;20:3353–62.
92. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucl Acids Res. 2014;42:D1001–6.
93. McCouch SR, Wright MH, Tung C-W, Maron LG, McNally KL, Fitzgerald M, et al. Open access resources for genome-wide association mapping in rice. Nat Commun. 2016;7:10532.
94. Huang D, Yi X, Zhang S, Zheng Z, Wang P, Xuan C, et al. GWAS4D: multidimensional analysis of context-specific regulatory variant for human complex diseases and traits. Nucl Acids Res. 2018;46:W114–20.
95. gwascat: representing and modeling data in the EMBL-EBI GWAS catalog version 2.22.0 from Bioconductor. [cited 2021 Apr 17]. https://rdrr.io/bioc/gwascat/.
96. Gogarten SM, Bhangale T, Conomos MP, Laurie CA, McHugh CP, Painter I, et al. GWASTools: an R/Bioconductor package for quality control and analysis of genome-wide association studies. Bioinformatics. 2012;28:3329–31.
97. Fuchsberger C, Taliun D, Pramstaller PP, Pattaro C, on behalf of the CKDGen Consortium. GWAtoolbox: an R package for fast quality control and handling of genome-wide association studies meta-analysis data. Bioinformatics. 2012;28:444–5.
98. Patron J, Serra-Cayuela A, Han B, Li C, Wishart DS. Assessing the performance of genome-wide association studies for predicting disease risk. PLoS One. 2019;14:e0220215.
99. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.
100. Dai M, Ming J, Cai M, Liu J, Yang C, Wan X, et al. IGESS: a statistical approach to integrating individual-level genotype data and summary statistics in genome-wide association studies. Bioinformatics. 2017;33:2882–9.
101. Chen CJ, Zhang Z. iPat: intelligent prediction and association tool for genomic research. Bioinformatics. 2018;34:1925–7.

102. Fan Y-H, Song Y-Q. IPGWAS: an integrated pipeline for rational quality control and association analysis of genome-wide genetic studies. Biochem Biophys Res Commun. 2012;422:363–8.
103. Julienne H, Lechat P, Guillemot V, Lasry C, Yao C, Araud R, et al. JASS: command line and web interface for the joint analysis of GWAS results. NAR Genomics Bioinformatics. 2020;2. https://doi.org/10.1093/nargab/lqaa003
104. Terada A, Yamada R, Tsuda K, Sese J. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. Bioinformatics. 2016;32:3513–5.
105. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet. 2015;47:291–5.
106. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics. 2017;33:272–9.
107. Bates D, Mächler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Statist Softw. 2015;67:1–48.
108. Pruim RJ, Welch RP, Sanna S, Teslovich TM, Chines PS, Gliedt TP, et al. LocusZoom: regional visualization of genome-wide association scan results. Bioinformatics. 2010;26:2336–7.
109. Vaez A, van der Most PJ, Prins BP, Snieder H, van den Heuvel E, Alizadeh BZ, et al. lodGWAS: a software package for genome-wide association analysis of biomarkers with a limit of detection. Bioinformatics. 2016;32:1552–4.
110. Mittag F, Büchel F, Saad M, Jahn A, Schulte C, Bochdanovits Z, et al. Use of support vector machines for disease risk prediction in genome-wide association studies: concerns and opportunities. Hum Mutat. 2012;33:1708–18.
111. Grace C, Farrall M, Watkins H, Goel A. Manhattan++: displaying genome-wide association summary statistics with multiple annotation layers. BMC Bioinformatics. 2019;20:610.
112. Haller T, Tasa T, Metspalu A. Manhattan harvester and cropper: a system for GWAS peak detection. BMC Bioinformatics. 2019;20:22.
113. Climente-González H, Azencott C-A. martini: an R package for genome-wide association studies using SNP networks. bioRxiv. 2021;2021.01.25.428047.
114. Childs LH, Lisec J, Walther D. Matapax: an online high-throughput genome-wide association study pipeline. Plant Physiol. 2012;158:1534–41.
115. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis. Bioinformatics. 2016;32:1981–9.
116. Willer CJ, Li Y, Abecasis GR. METAL: fast and efficient meta-analysis of genomewide association scans. Bioinformatics. 2010;26:2190–1.
117. Lin Y-C, Hsiao C-L, Hsieh A-R, Lian I-B, Fann CSJ. Using maximal segmental score in genome-wide association studies. Genet Epidemiol. 2012;36:594–601.
118. Skol AD, Scott LJ, Abecasis GR, Boehnke M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. Nat Genet. 2006;38:209–13.
119. Shim H, Chasman DI, Smith JD, Mora S, Ridker PM, Nickerson DA, et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. PLoS One. 2015;10:e0120758.
120. Xavier A, Xu S, Muir WM, Rainey KM. NAM: association studies in multiple populations. Bioinformatics. 2015;31:3862–4.
121. Fabregat-Traver D, Sharapov SZ, Hayward C, Rudan I, Campbell H, Aulchenko Y, et al. High-performance mixed models based genome-wide association analysis with omicABEL software. F1000Res. 2014;3:200.
122. German CA, Sinsheimer JS, Klimentidis YC, Zhou H, Zhou JJ. Ordered multinomial regression for genetic association analysis of ordinal phenotypes at Biobank scale. Genet Epidemiol. 2020;44:248–60.

123. Wen Y, Wang W, Guo X, Zhang F. PAPA: a flexible tool for identifying pleiotropic pathways using genome-wide association study summaries. Bioinformatics. 2016;32:946–8.
124. Yaspan BL, Bush WS, Torstenson ES, Ma D, Pericak-Vance MA, Ritchie MD, et al. Genetic analysis of biological pathway data through genomic randomization. Hum Genet. 2011;129:563–71.
125. Alonso-Gonzalez A, Calaza M, Rodriguez-Fontenla C, Carracedo A. Gene-based analysis of ADHD using PASCAL: a biological insight into the novel associated genes. BMC Med Genomics. 2019;12:143.
126. Tucker G, Price AL, Berger B. Improving the power of GWAS and avoiding confounding from population stratification with PC-select. Genetics. 2014;197:1045–9.
127. Zhang W, Dai X, Wang Q, Xu S, Zhao PX. PEPIS: a pipeline for estimating epistatic effects in quantitative trait locus mapping and genome-wide association studies. PLoS Comput Biol. 2016;12:e1004925.
128. Cheng Z, Chu H, Fan Y, Li C, Song Y-Q, Zhou J, et al. PExFInS: an integrative post-GWAS explorer for functional indels and SNPs. Sci Rep. 2015;5:17302.
129. George G, Gan S, Huang Y, Appleby P, Nar AS, Venkatesan R, et al. PheGWAS: a new dimension to visualize GWAS across multiple phenotypes. Bioinformatics. 2020;36:2500–5.
130. Denny JC, Ritchie MD, Basford MA, Pulley JM, Bastarache L, Brown-Gentry K, et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. Bioinformatics. 2010;26:1205–10.
131. Grady BJ, Torstenson E, Dudek SM, Giles J, Sexton D, Ritchie MD. Finding unique filter sets in PLATO: a precursor to efficient interaction analysis in GWAS data. Pac Symp Biocomput. 2010:315–26.
132. Slifer SH. PLINK: key functions for data analysis. Curr Protoc Hum Genet. 2018;97:e59.
133. Kooperberg C, LeBlanc M. Increasing the power of identifying gene × gene interactions in genome-wide association studies. Genet Epidemiol. 2008;32:255–63.
134. Tsai S-F, Tung C-W, Tsai C-A, Liao C-T. An exhaustive scan method for SNP main effects and SNP × SNP interactions over highly homozygous genomes. J Comput Biol. 2017;24:1254–64.
135. van der Most PJ, Vaez A, Prins BP, Munoz ML, Snieder H, Alizadeh BZ, et al. QCGWAS: a flexible R package for automated quality control of genome-wide association results. Bioinformatics. 2014;30:1185–6.
136. Wigginton JE, Cutler DJ, Abecasis GR. A note on exact tests of Hardy-Weinberg equilibrium. Am J Hum Genet. 2005;76:887–93.
137. Daye ZJ, Li H, Wei Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. Nucl Acids Res. 2012;40:e60.
138. Abecasis GR, Cardon LR, Cookson WOC. A general test of association for quantitative traits in nuclear families. Am J Hum Genet. 2000;66:279–92.
139. Julienne H, Shi H, Pasaniuc B, Aschard H. RAISS: robust and accurate imputation from summary statistics. Bioinformatics. 2019;35:4837–9.
140. Shabalin AA, Hattab MW, Clark SL, Chan RF, Kumar G, Aberg KA, et al. RaMWAS: fast methylome-wide association study pipeline for enrichment platforms. Bioinformatics. 2018;34:2283–5.
141. Gel B, Díez-Villanueva A, Serra E, Buschbeck M, Peinado MA, Malinverni R. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. Bioinformatics. 2016;32:289–91.
142. Haller T, Kals M, Esko T, Mägi R, Fischer K. RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. Brief Bioinformatics. 2015;16:39–44.
143. Zhbannikov IY, Arbeev KG, Yashin AI. rqt: an R package for gene-level meta-analysis. Bioinformatics. 2017;33:3129–30.
144. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. Genome Biol. 2016;17:238.

145. Wang B, Cunningham JM. (Holly) Yang X. Seq2pathway: an R/Bioconductor package for pathway analysis of next-generation sequencing data. Bioinformatics. 2015;31:3043–5.
146. Wang GT, Li B, Lyn Santos-Cortez RP, Peng B, Leal SM. Power analysis and sample size estimation for sequence-based association studies. Bioinformatics. 2014;30:2377–8.
147. Tzeng J-Y, Zhang D, Pongpanich M, Smith C, McCarthy MI, Sale MM, et al. Studying gene and gene-environment effects of uncommon and common variants on continuous traits: a marker-set approach using gene-trait similarity regression. Am J Hum Genet. 2011;89:277–88.
148. Ionita-Laza I, Lee S, Makarov V, Buxbaum JD, Lin X. Sequence kernel association tests for the combined effect of rare and common variants. Am J Hum Genet. 2013;92:841–53.
149. Wang S, Dvorkin D, Da Y. SNPEVG: a graphical tool for GWAS graphing with mouse clicks. BMC Bioinformatics. 2012;13:319.
150. Mei H, Li L, Jiang F, Simino J, Griswold M, Mosley T, et al. snpGeneSets: an R package for genome-wide study annotation. G3 Genes Genome Genet. 2016;6:4087–95.
151. Zheng X, Levine D, Shen J, Gogarten SM, Laurie C, Weir BS. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics. 2012;28:3326–8.
152. Clayton D. snpStats: SnpMatrix and XSnpMatrix classes and methods version 1.40.0 from Bioconductor. [cited 2021 Apr 18]. https://rdrr.io/bioc/snpStats/.
153. Curk T, Rot G, Zupan B. SNPsyn: detection and exploration of SNP–SNP interactions. Nucl Acids Res. 2011;39:W444–9.
154. Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nat Rev Genet. 2010;11:499–511.
155. Kochunov P, Jahanshad N, Marcus D, Winkler A, Sprooten E, Nichols TE, et al. Heritability of fractional anisotropy in human white matter: a comparison of Human Connectome Project and ENIGMA-DTI data. NeuroImage. 2015;111:300–11.
156. Weng L, Macciardi F, Subramanian A, Guffanti G, Potkin SG, Yu Z, et al. SNP-based pathway enrichment analysis for genome-wide association studies. BMC Bioinformatics. 2011;12:99.
157. Schlauch D, Fier H, Lange C. Identification of genetic outliers due to sub-structure and cryptic relationships. Bioinformatics. 2017;33:1972–9.
158. Syed H, Jorgensen AL, Morris AP. SurvivalGWAS_SV: software for the analysis of genome-wide association studies of imputed genotypes with "time-to-event" outcomes. BMC Bioinformatics. 2017;18:265.
159. Chen L, Qin ZS. traseR: an R package for performing trait-associated SNP enrichment analysis in genomic intervals. Bioinformatics. 2016;32:1214–6.
160. Collins C, Didelot X. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. PLoS Comput Biol. 2018;14:e1005958.
161. Alexander J, Mantzaris D, Georgitsi M, Drineas P, Paschou P. Variant Ranker: a web-tool to rank genomic data according to functional significance. BMC Bioinformatics. 2017;18:341.
162. Burren OS, Guo H, Wallace C. VSEAMS: a pipeline for variant set enrichment analysis using summary GWAS data identifies IKZF3, BATF and ESRRA as key transcription factors in type 1 diabetes. Bioinformatics. 2014;30:3342–8.
163. Sun R, Xia X, Chong KC, Zee BC-Y, WKK W, Wang MH. wtest: an integrated R package for genetic epistasis testing. BMC Med Genomics. 2019;12:180.
164. Fang H, Knezevic B, Burnham KL, Knight JC. XGR software for enhanced interpretation of genomic summary data, illustrated by application to immunological traits. Genome Med. 2016;8:129.
165. Montes JM, Melchinger AE, Reif JC. Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci. 2007;12:433–6.
166. Agrama HA, Eizenga GC, Yan W. Association mapping of yield and its components in rice cultivars. Mol Breed. 2007;19:341–56.

167. Gupta PK, Rustgi S, Kulwal PL. Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol. 2005;57:461–85.
168. Zhao Y, Wang H, Chen W, Li Y. Genetic structure, linkage disequilibrium and association mapping of Verticillium wilt resistance in elite cotton (*Gossypium hirsutum* L.) germplasm population. PLoS One. 2014;9:e86308.
169. Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, et al. Genome sequencing reveals agronomically important loci in rice using MutMap. Nat Biotechnol. 2012;30:174–8.
170. Takagi H, Uemura A, Yaegashi H, Tamiru M, Abe A, Mitsuoka C, et al. MutMap-Gap: whole-genome resequencing of mutant F_2 progeny bulk combined with de novo assembly of gap regions identifies the rice blast resistance gene *Pii*. New Phytol. 2013;200:276–83.
171. Garris AJ, McCOUCH SR, Kresovich S. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics. 2003;165:759–69.
172. Olsen KM, Caicedo AL, Polato N, McClung A, McCouch S, Purugganan MD. Selection under domestication: evidence for a sweep in the rice waxy genomic region. Genetics. 2006;173:975–83.
173. Harushima Y, Yano M, Shomura A, Sato M, Shimano T, Kuboki Y, et al. A high-density rice genetic linkage map with 2275 markers using a single F_2 population. Genetics. 1998;148:479–94.
174. Huang X, Han B. Natural variations and genome-wide association studies in crop plants. Annu Rev Plant Biol. 2014;65:531–51.
175. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q, et al. Genetic properties of the maize nested association mapping population. Science. 2009;325:737–40.
176. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C, et al. The genetic architecture of maize flowering time. Science. 2009;325:714–8.
177. Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, et al. Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet. 2011;43:163–8.
178. Poland JA, Bradbury PJ, Buckler ES, Nelson RJ. Genome-wide nested association mapping of quantitative resistance to northern leaf blight in maize. Proc Natl Acad Sci U S A. 2011;108:6893–8.
179. Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, et al. Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet. 2011;43:159–62.
180. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD, et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis thaliana*. PLoS Genet. 2009;5:e1000551.
181. Huang X, Paulo M-J, Boer M, Effgen S, Keizer P, Koornneef M, et al. Analysis of natural allelic variation in *Arabidopsis* using a multiparent recombinant inbred line population. Proc Natl Acad Sci U S A. 2011;108:4488–93.
182. Austin RS, Vidaurre D, Stamatiou G, Breit R, Provart NJ, Bonetta D, et al. Next-generation mapping of *Arabidopsis* genes. Plant J. 2011;67:715–25.
183. James GV, Patel V, Nordström KJ, Klasen JR, Salomé PA, Weigel D, et al. User guide for mapping-by-sequencing in *Arabidopsis*. Genome Biol. 2013;14:1–13.
184. Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet. 2010;42:961–7.
185. Myles S, Peiffer J, Brown PJ, Ersoz ES, Zhang Z, Costich DE, et al. Association mapping: critical considerations shift from genotyping to experimental design. Plant Cell. 2009;21:2194–202.
186. Platt A, Vilhjálmsson BJ, Nordborg M. Conditions under which genome-wide association studies will be positively misleading. Genetics. 2010;186:1045–52.
187. Vilhjálmsson BJ, Nordborg M. The nature of confounding in genome-wide association studies. Nat Rev Genet. 2013;14:1–2.

188. Zhang Z, Buckler ES, Casstevens TM, Bradbury PJ. Software engineering the mixed model for genome-wide association studies on large samples. Brief Bioinformatics. 2009;10:664–75.
189. Qin H, Chen M, Yi X, Bie S, Zhang C, Zhang Y, et al. Identification of associated SSR markers for yield component and fiber quality traits based on frame map and upland cotton collections. PLoS One. 2015;10:e0118073.
190. Wang Z, Qiang H, Zhao H, Xu R, Zhang Z, Gao H, et al. Association mapping for fiber-related traits and digestibility in alfalfa (*Medicago sativa*). Front Plant Sci. 2016;7:331.
191. Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet. 2012;44:32–9.
192. Jia G, Huang X, Zhi H, Zhao Y, Zhao Q, Li W, et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). Nat Genet. 2013;45:957–61.
193. Li H, Peng Z, Yang X, Wang W, Fu J, Wang J, et al. Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. Nat Genet. 2013;45:43–50.
194. Morris GP, Ramu P, Deshpande SP, Hash CT, Shah T, Upadhyaya HD, et al. Population genomic and genome-wide association studies of agroclimatic traits in sorghum. Proc Natl Acad Sci U S A. 2013;110:453–8.
195. Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, et al. Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. Nat Commun. 2011;2:467.
196. Huang X, Kurata N, Wang Z-X, Wang A, Zhao Q, Zhao Y, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490:497–501.
197. Wang M, Jiang N, Jia T, Leach L, Cockram J, Waugh R, et al. Genome-wide association mapping of agronomic and morphologic traits in highly structured populations of barley cultivars. Theor Appl Genet. 2012;124:233–46.
198. Harper AL, Trick M, Higgins J, Fraser F, Clissold L, Wells R, et al. Associative transcriptomics of traits in the polyploid crop species Brassica napus. Nat Biotechnol. 2012;30:798–802.
199. Ling H-Q, Zhao S, Liu D, Wang J, Sun H, Zhang C, et al. Draft genome of the wheat A-genome progenitor Triticum urartu. Nature. 2013;496:87–90.
200. Seren Ü, Vilhjálmsson BJ, Horton MW, Meng D, Forai P, Huang YS, et al. GWAPP: a web application for genome-wide association mapping in Arabidopsis. Plant Cell. 2012;24:4793–805.
201. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. A mixed-model approach for genome-wide association studies of correlated traits in structured populations. Nat Genet. 2012;44:1066–71.
202. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, Long Q, et al. An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet. 2012;44:825–30.
203. Jin L, Lu Y, Xiao P, Sun M, Corke H, Bao J. Genetic diversity and population structure of a diverse set of rice germplasm for association mapping. Theor Appl Genet. 2010;121:475–87.
204. Courtois B, Audebert A, Dardou A, Roques S, Ghneim-Herrera T, Droc G, et al. Genome-wide association mapping of root traits in a Japonica Rice Panel. Baxter I, editor. PLoS One. 2013;8:e78037.
205. Zhao W-G, Chung J-W, Kwon S-W, Lee J-H, Ma K-H, Park Y-J. Association analysis of physicochemical traits on eating quality in rice (*Oryza sativa* L.). Euphytica. 2013;191:9–21.
206. Dang X, Thi TGT, Dong G, Wang H, Edzesi WM, Hong D. Genetic diversity and association mapping of seed vigor in rice (*Oryza sativa* L.). Planta. 2014;239:1309–19.
207. Lu Q, Zhang M, Niu X, Wang S, Xu Q, Feng Y, et al. Genetic variation and association mapping for 12 agronomic traits in indica rice. BMC Genomics. 2015;16:1–17.

208. Zhang N, Xu Y, Akash M, McCouch S, Oard JH. Identification of candidate markers associated with agronomic traits in rice using discriminant analysis. Theor Appl Genet. 2005;110:721–9.
209. Ma X, Feng F, Wei H, Mei H, Xu K, Chen S, et al. Genome-wide association study for plant height and grain yield in rice under contrasting moisture regimes. Front Plant Sci. 2016;7. http://journal.frontiersin.org/article/10.3389/fpls.2016.01801/full
210. Pan Y, Zhang H, Zhang D, Li J, Xiong H, Yu J, et al. Genetic analysis of cold tolerance at the germination and booting stages in rice by association mapping. PLoS One. 2015;10:e0120590.
211. Kumar V, Singh A, Mithra SVA, Krishnamurthy SL, Parida SK, Jain S, et al. Genome-wide association mapping of salinity tolerance in rice (*Oryza sativa*). DNA Res. 2015;22:133–45.
212. Donde R, Mohapatra S, Baksh SY, Padhy B, Mukherjee M, Roy S, et al. Identification of QTLs for high grain yield and component traits in new plant types of rice. bioRxiv. 2020.
213. Anandan A, Anumalla M, Pradhan SK, Ali J. Population structure, diversity and trait association analysis in rice (*Oryza sativa* L.) germplasm for early seedling vigor (ESV) using trait linked SSR markers. PLoS One. 2016;11:e0152406.
214. Pandit E, Tasleem S, Barik SR, Mohanty DP, Nayak DK, Mohanty SP, et al. Genome-wide association mapping reveals multiple QTLs governing tolerance response for seedling stage chilling stress in *indica* rice. Front Plant Sci. 2017;8:552.
215. Swamy BM, Shamsudin NAA, Abd Rahman SN, Mauleon R, Ratnam W, Cruz MTS, et al. Association mapping of yield and yield-related traits under reproductive stage drought stress in rice (*Oryza sativa* L.). Rice. 2017;10:1–13.
216. Rebolledo MC, Peña AL, Duitama J, Cruz DF, Dingkuhn M, Grenier C, et al. Combining image analysis, genome wide association studies and different field trials to reveal stable genetic regions related to panicle architecture and the number of spikelets per panicle in rice. Front Plant Sci. 2016;7:1384.
217. Yuan J, Wang X, Zhao Y, Khan NU, Zhao Z, Zhang Y, et al. Genetic basis and identification of candidate genes for salt tolerance in rice by GWAS. Sci Rep. 2020;10:9958.
218. Schläppi MR, Jackson AK, Eizenga GC, Wang A, Chu C, Shi Y, et al. Assessment of five chilling tolerance traits and GWAS mapping in rice using the USDA mini-core collection. Front Plant Sci. 2017;8. https://www.frontiersin.org/articles/10.3389/fpls.2017.00957/full
219. Bollinedi H, Yadav AK, Vinod KK, Gopala Krishnan S, Bhowmick PK, Nagarajan M, et al. Genome-Wide Association study reveals novel Marker-Trait Associations (MTAs) governing the localization of Fe and Zn in the rice grain. Front Genet. 2020;11. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7188789/
220. Kumar J, Saripalli G, Gahlaut V, Goel N, Meher PK, Mishra KK, et al. Genetics of Fe, Zn, β-carotene, GPC and yield traits in bread wheat (*Triticum aestivum* L.) using multi-locus and multi-traits GWAS. Euphytica. 2018;214:219.
221. Caicedo AL, Williamson SH, Hernandez RD, Boyko A, Fledel-Alon A, York TL, et al. Genome-wide patterns of nucleotide polymorphism in domesticated rice. PLoS Genet. 2007;3:e163.
222. Zhu Q, Zheng X, Luo J, Gaut BS, Ge S. Multilocus analysis of nucleotide variation of Oryza sativa and its wild relatives: severe bottleneck during domestication of rice. Mol Biol Evol. 2007;24:875–88.
223. Mather KA, Caicedo AL, Polato NR, Olsen KM, McCouch S, Purugganan MD. The extent of linkage disequilibrium in rice (*Oryza sativa* L.). Genetics. 2007;177:2223–32.
224. Clark RM, Schweikert G, Toomajian C, Ossowski S, Zeller G, Shinn P, et al. Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. Science. 2007;317:338–42.

225. Nordborg M, Weigel D. Next-generation genetics in plants. Nature. 2008;456:720–3.

226. Zhang D, Zhang H, Wang M, Sun J, Qi Y, Wang F, et al. Genetic structure and differentiation of *Oryza sativa* L. in China revealed by microsatellites. Theor Appl Genet. 2009;119:1105–17.

227. McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, et al. Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci U S A. 2009;106:12273–8.

228. Tian Z, Qian Q, Liu Q, Yan M, Liu X, Yan C, et al. Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. Proc Natl Acad Sci U S A. 2009;106:21760–5.

Expression Profiling and Discovery of microRNA

20

N. Rajesh, Manoj Kumar Gupta, Ravindra Donde, S. Sabarinathan, Gayatri Gouda, Goutam Kumar Dash, Menaka Ponnana, Lambodar Behera, and Ramakrishna Vadde

Abstract

miRNAs, on average comprised of 22-nucleotides with small non-coding RNAs, regulate the gene expression of targeted genes. Thousands of miRNAs have been identified, having biological significance in many pathways. These microRNAs have also been used as biomarkers for diagnostic and agricultural purposes. Thus, this chapter attempts to describe in brief miRNA biogenesis pathways, miRNA profiling methods, and bioinformatics tools of miRNA profiling. Additionally, we will discuss the role of mi RNA and its applications. miRNA biogenesis can be broadly categorized into canonical and non-canonical pathways. There are various sample types and miRNA extraction procedures. miRNA sequences, once extracted, can be subjected to various computational tools that may aid in understanding its structure and functions. However, few researchers have suggested that there is still scope for developing these tools with appropriate algorithms for avoiding false positive results.

N. Rajesh · R. Vadde (✉)
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

M. K. Gupta · R. Donde · G. Gouda · G. K. Dash · L. Behera
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_20

Keywords

miRNA · Quantitative RT-PCR · Microarray · Next-generation sequencing

Abbreviations

AGO	Argonaut
miRNAs	MicroRNAs
ncRNAs	Non-coding RNAs
RT	Reverse transcription
UTR	Untranslated regions

20.1 Introduction

MicroRNAs (miRNAs) are small non-coding RNAs that have an average length of 22 nucleotides. Mostly, miRNAs are synthesized as primary miRNAs from DNA sequences, which are further processed into precursor miRNAs and finally to mature miRNAs, which interact with 3′UTR to suppress target miRNA expression [1]. However, some authors have also reported that miRNA may also interact with exons, promoters, 5′UTR [2], and activate gene expression under certain conditions [3]. MicroRNA only represents 0.01% of overall RNA, although it is probable that an actual miRNA copy number may be higher than that of mRNA, that is, an average of about 500/cell [4]. This is due to its low molecular weight and diversity. Presently, ~200–300 miRNAs have been estimated in model organisms such as *D. melanogaster, C. elegans*, and *A. thaliana*. In humans, ~1000 miRNAs were estimated. miRNAs regulate transcription and translation by transporting between different subcellular organelles [5]. They play a major role in various biological processes associated with normal development [6] and abnormal miRNA expression [7, 8]. They also mediate various cell–cell communications [9–11]. Since they act as signaling molecules, extracellular secreted miRNAs act as biomarkers for various disease identification [9–11]. Thus, miRNA profiling has developed interest among scientists working in different fields of biology and medicine [12]. Here, in this chapter, the authors describe in brief miRNA biogenesis pathways, miRNA profiling methods, and bioinformatics tools of miRNA profiling. Additionally, the applications of miRNA are also discussed.

20.2 miRNAs Biogenesis

Two organized endonucleolytic cleavages, through both the RNase III enzymes Drosha and Dicer (Fig. 20.1) [14], are involved in the evolutionarily conserved pathway that results in mature miRNA [13]. Drosha transforms the primary miRNA

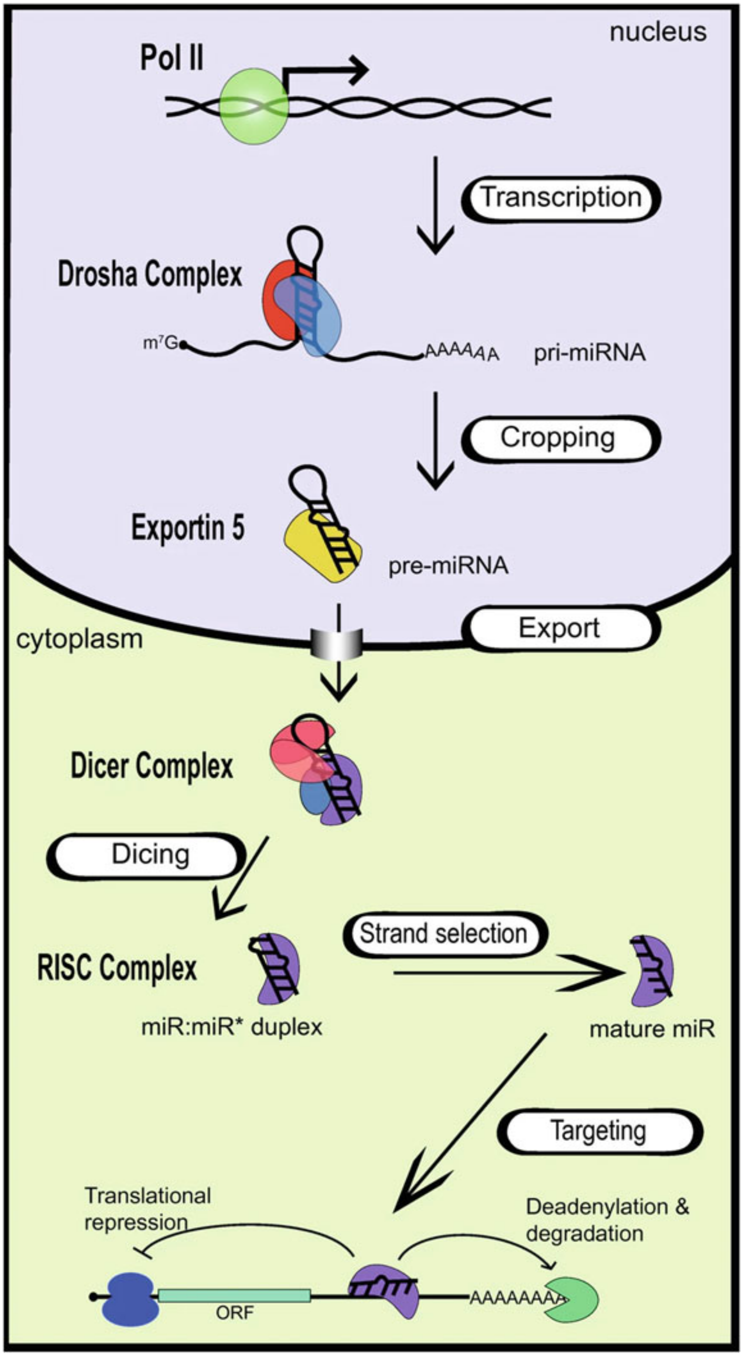


Fig. 20.1 Biogenesis of miRNA. Drosha processes capped as well as polyadenylated pri-miRNAs within the nucleus to produce pre-miRNAs, which are produced by DNA pol II. Pre-miRNAs are processed through Dicer to construct the mature miRNA/miRNA* duplex after being translocated into cytoplasm by exportin 5. MiRNAs are loaded into the RISC complex after they have been

transcript (pri-miRNA) into a 60–100 nt hairpin structure called the precursor-miRNA (pre-miRNA) after transcription by RNA polymerase II (RNA pol II) [15–18]. The pre-miRNA is conveyed into the cytoplasm via exportin-5 and Ran-GTP, where it experiences another round of processing catalyzed through Dicer (Fig. 20.1) [19, 20]. The mature miRNA guide strand and also the miRNA* passenger strand are separated by this cleavage event, resulting in a 22-nt double-stranded product. The passenger strand is degraded when the mature miRNA is loaded further into RNA-induced silencing complex (RISC) (Fig. 20.1) [21]. Despite substantial progress in understanding the underlying mechanism of miRNA biogenesis, little is known about the complex processes that control miRNA expression. Each phase of the generalized biogenesis pathway has also been observed to be differentially regulated, allowing for precise control of miRNA expression, as discussed below. Recent research has revealed that not all miRNAs are created in a similar way and that different mechanisms account for individual miRNA control [13].

20.3 microRNA Profiling

High-quality miRNA can be extracted from a wide range of cell and tissue sources. Although isolation principles of miRNA are the same as that of RNA, a few exceptions are modified for retaining the small fraction of RNA [22]. miRNeasy, mirVana,™ and PureLink™ are the widely used miRNAs commercially available. Checking the quality of isolated RNA is important for its accuracy and to maintain its integrity in miRNA quantification. Since many profiling approaches can be achieved by means of total RNA, a detailed valuation of the miRNA is routine to evaluate the yield, and miRNA integrity is checked through spectrophotometry and automated capillary electrophoresis instruments such as Experion (Bio-Rad) and Bioanalyzer 2100 (Agilent). For the miRNA estimation of its abundance, expressivity as the RNA quantity in the 15–40 nt window, a small RNA chip is required. Overall, the accuracy of this method is high when the overall integrity of RNA is also very high [23].

Nevertheless, numerous properties exclusive to miRNAs pose challenges for their precise discovery and quantification. miRNA's 22 nt length is insufficient for annealing to classical primers designed for reverse transcription. Total RNA mass is represented by a small RNA fraction, and miRNAs must therefore be selectively identified within other diverse RNA species in the background. To overcome these challenges, three major approaches include qRT-PCR, hybridization-based methods, and high-throughput sequencing. In some cases, the genomic location of the dysregulated candidate is difficult to determine.

Fig. 20.1 (continued) processed. The RISC complex is only connected with one strand of the duplex. Within the 3′UTR, the mature miRNA directs repression of mRNA with partially complementary miRNA binding sites. (Adapted from [13])

20.3.1 Quantitative PCR (qPCR) Methods for Pre-miRNAs and Pri-miRNA's Analysis

miRNA analysis by qPCR was first reported by Schnittger et al. [24]. They used three primers for pri- and pre-miRNA analysis. Two forward primers, specifically one binds to the hairpin region outside, and the other targets the pre-miRNA inside the region. The reverse primer specifically binds to hairpin inside the region and amplifies pre-miRNA, whereas another primer set amplifies both pre- and pri-miRNA. SYBR green fluorescence was quantified.

20.3.2 Quantitative PCR Methods for Mature miRNAs

20.3.2.1 Stem-Loop RT-Based Approaches

Chen et al. described the approach of the RT q-PCR stem-loop for miRNA analysis [25], which uses stem-loop RT primer that specifically binds to a mature miRNA. This results in cDNA amplification with miRNA-specific forward and universal reverse primer, with the quantification of mature miRNA levels using miRNA-specific TaqMan probe. The other group uses the pre-amplification step to allow multiplexing, while the other independent group uses the universal TaqMan probe [26] for multiplexing. A stem-loop RT primer (11 bp instead of 6 bp) is a cost-effective approach, resulting in higher specificity by amplifying the longer binding region in the miRNA sequence with the use of SYBR Green [27]. The definite recognition of closely related members of the miRNA family results from the use of different stem-loop RT primers (let-7) [28].

20.3.2.2 Polyadenylation-Based Approaches

For mature miRNA polyadenylation, Shi and Chiang employed poly(A) polymerase and poly(T) adapter in order to produce cDNA [29]. Forward and reverse primer, specific to miRNA, which binds the PCR amplification adapter to Poly(T) and uses SYBR Green for measurement of fluorescence. Another approach uses mature miRNAs polyadenylation and poly(T) adapters for cDNA synthesis and forward primer and reverse universal primers unique to miRNAs. Another approach uses polyadenylation of the miRNA and poly(T) adapters to generate cDNA, with forward and reverse primers specific to miRNA and SYBR green for quantification [30].

20.3.2.3 Ligation-Based Approaches

In this method, miRNAs are circularized using ligase, RT-qPCR of circularized miRNA with SYBR green, and overlapping primers. Another method includes ligating the universal DNA adjuster to mature miRNAs, which uses the universal RT main, which binds to a reverse primer. miRNA-specific forward and universal reverse primers with SYBR Green for fluorescence are used for qPCR [31]. Chen et al. method does not distinguish 5′ and 3′ variants of miRNA results in the

Dumbbell-PCR approach. It uses RT first, universal primer, reverse primer, and miRNA-specific TaqMan probe were used later [32].

20.3.3 Quantitative PCR Methods for Isoforms of miRNA (isomiRs) Analysis

Small RNA high throughput analysis indicates there is diversity in the isomiRs within cells [33, 34]. IsomiR is shaped and divided into 5′ isomiR and 3′ isomiR; 3′ isomiRs being more common [33–35]. multi-insomiRs are of differential functioning and affected by many biological mechanisms and diseases due to changes in the length or sequence of miRNAs [33–35]. The use of experiments designed to analyze canonic miRNAs is found to be quite normal and to be impaired in biological models and diseases [36, 37]. The use of isomiRs is shown to be quite normal. Changes in isomiRs are less important to polyadenylation-based processes [36, 37]. The precise recognization of mature miRNA isomers with the polyadenylation approach was showcased by various miRNA-specified forward primers [38].

20.3.4 Microarray Platforms

20.3.4.1 One-Color Versus Two-Color Arrays

The method of making a microarray has its impression on the kind of investigational plan and scrutiny to be achieved. Numerous microarray formats are categorized as one-color or two-color investigation. Mostly, two-color analysis is produced by spotting oligonucleotide probes onto the array by robotic technology. It involves the shifting of probes by liquid adherence to either single or arrayed pins. There is a characteristically considerable difference among specific microarrays in the quantity of spotted material. As a result, specific microarrays are not straightly analogous without the use of a reference sample. To perform two-color experiments with these custom-spotted microarrays, reference as well as test samples were fluorescently labeled and hybridized on the same array. The primary advantage of spotted arrays is that they can be custom-produced in small batches for a moderate cost, facilitating changes to the probe set as new miRNAs are discovered [39, 40].

20.3.4.2 miRNA Preparation Through Microarrays

Purification of miRNA

Microarray gene expression quality largely depends on the quality of RNA used. Hence, to ensure the quality of RNA, vigorous and reproducible methods are essential for the qualitative extraction of miRNA. Many isolation kits are available, but they concentrated mostly on mRNA and not on miRNA. These mRNAs are contaminated for specific isolation of miRNA. Solid-phase extraction, microfiltration, and reverse-phase or ion-exchange chromatography are used for separating small and large RNA fractions. Still, they didn't provide the levels of recovery, purity, and reproducibility for high-quality analysis. Hence, denaturing polyacrylamide gel electrophoresis (PAGE) is routinely used for miRNA extraction from total RNA [36, 37, 41, 42] but is a time-consuming process. Ambion, Inc. has developed a device (flashPAGE fractionator system) that can extract miRNA rapidly and is reproducible with 80% yield. Relative abundance of flashPAGE-isolated miRNAs analyzed on a microarray has been shown to be 1 ng to 10 g from mammalian tissue.

Labeling of miRNA

Labeling of miRNA was done after its isolation for their detection on arrays. Methods employed for labeling miRNA include direct labeling, random priming, and amplification by PCR [43–50]. In 2003, Krichevsky et al. [51] designed the first oligonucleotide array to detect miRNA in brain tissues by labeling low molecular weight RNAs with radioactive isotopes. The simplest and most typical labeling technique is direct labeling, which employs a tailing strategy with short, tagged, and enzymatically attached miRNAs. Based on Ambion's mirVana miRNA labeling kit, which utilizes poly(A) polymerase (PAP) to add a combination of unmodified and amine-modified nucleotides to the miRNA's 3′ end [52]. Then tailed miRNA was labeled the amine-reactive reagents, including fluorescent dyes, cyor Alex dyes, or NHS biotin for detection with streptavidin coupled to fluorescence dyes. Homogeneous labeling of this method provides miRNA fraction with the highest specific activity without introducing bias. Mature miRNA labeling by this method allows accurate profiling at lower sample input. Experiments in which known amounts of miRNA were spiked into RNA samples indicate that this procedure permits detection of 10 pg (3 fmol) of miRNA in 10 g of total RNA [52].

In the other direct labeling method, 3′ end of each miRNA was labeled with one or two fluorophore-labeled nucleotides with T4 RNA ligase. This labeling procedure leads without prejudice to the maximum activity. Wang et al. added dimethylsulfoxide (DMSO), an important RNA denatures, in the reaction solution and found that T4 RNA ligase activity was enhanced by up to 20% DMSO [53] for reducing interference of structure and sequence variations between miRNAs. This direct labeling method has been shown to be an effective tool for miRNA labeling and to be easily performed under laboratory conditions [53–55].

Normalization Methods for Microarray Studies with miRNA

For differential expression and clustering of labeling genes, microarray data must be normalized and corrected to achieve accurate results. Normalization is also done to reduce bias in the marking of dye and variations in hybridization and screening [56]. It should be highlighted that normalization is the method of eliminating variance between array experiments from non-biological origins. In general, it is recognized that the use of normalization methodologies for microarray experiments has a profound effect on precision, accuracy, and overfitting [57–59]. Consequently, downstream differential expression tests, classifier growth, and data mining are highly dependent on the option of data processing. Even in the well-developed field of mRNA microarray analysis, the acceptable option of standardization methods is still under discussion [60, 61]. This problem is further compounded in the field of miRNA microarray analysis by the fact that the amount of miRNA extracted from biological samples is neither quantitative nor known relative to the original abundance of the total RNA approaches from the mRNA microarray expression review in the expectation that assumptions and approximations are not violated. These methods included median scaling (both global and/or chip-specific), spiked-in control scaling, and logarithmic or other variance-stabilizing normalizations. The most common approach to microarray normalization of miRNA includes the pre-processing of data by means of averaging of technical replicates and background subtraction followed by median normalization and the logarithmical transformation. For profiling miRNA expression in human tissues, Barad et al. [62] took a similar approach, except that post-normalization thresholds based on negative controls were implemented.

Interpretation and Data Analysis

Variability of microarray tests can be accounted for when designing test statistics for differential expression; fold change does not. Preferably, research statistics that provide variation shrinkage are used. Family-wise error rate management methods can usually be prevented by utilizing false-discovery-rate estimate procedures. It is advised to do gene-class testing while conducting differential expression studies. Many concerns arise about how to determine intersections of results while evaluating several similar propositions and how to properly utilize resampling-based inference. It also is necessary to carefully evaluate whether cluster analysis solves the question being posed and whether enough sample sizes can be collected to produce accurate findings before pursuing it. Cross-validation could be carried out on data that had little role in the derivation of the prediction law. Though widely addressed in microarray research, replication of findings demands more attention. Validation, in addition to which parameters decide validation, is a subject that has yet to be thoroughly discussed [60].

20.3.5 Next-Generation Sequencing (NGS)

RNA-seq has many benefits, particularly because its cost has declined in the last few years, compared to conventional techniques for qPCR and microarrays. RNA-seq, for example, is not linked to the predefined genes to be tested, increasing the numerous genes to be identified and discovering novel transcripts [61]. The various Genome Analyzer (GA) systems currently used for the next-generation analyses are Illumina's / Solexa, Rosche/454, ABI / SOLiD. Illumina's GA used for microRNA expression profiling uses massively parallel sequencing of millions of fragments. The low difference among platforms associated with standard strategies has added additional benefit [63].

Earlier, Rao et al. conducted a study to figure out if the RNA-Seq transcriptomic platform provided substantial advantages over microarrays for toxicogenomic analysis. With the aid of both gene expression platforms, "RNA samples from the livers of rats treated for 5 days with five hepatotoxicants (α-naphthylisothiocyanate/ANIT, carbon tetrachloride/CCl_4, methylenedianiline/MDA, acetaminophen/APAP, and diclofenac/DCLF) were examined (RNA-Seq and microarray)." Data were contrasted to see whether there could be some benefit given by RNA-Seq compared to microarrays. When opposed to microarrays, RNA-Seq was able to classify slightly more differentially expressed protein-coding genes, as well as have a greater variety of expression level shifts [64]. APAP- and DCLF-treated rats were shown to have slightly more differentially expressed genes (DEGs) in contrast to ANIT, MDA, and CCl_4 treated rats. Approximately, 78% of the reported DEGs on microarrays are compatible with RNA-Seq results, with a Spearman's correlation coefficient of 0.7–0.83. Both platforms find dysregulation of liver-relevant pathways, such as Nrf2, eiF2, cholestasis, glutathione, cholesterol biosynthesis, and LPS/IL-1-mediated RXR inhibition. In the RNA-Seq results, an excess of liver-related DEGs was observed that not only substantially enriched these pathways but also indicated modulation of additional liver-relevant pathways. Additionally, RNA-Seq permitted the identification of non-coding de novo (or freshly synthesized) expressed sequence tags (ESTs) that could give improved mechanistic clarification. Although these findings suggest that RNA-Seq is an appropriate alternative to microarrays for rat toxicogenomic studies, many advantages have been discovered. Unlike cDNA microarrays, RNA-Seq has a wider functional spectrum and is capable of detecting a larger number of DEGs, thereby resulting in further visibility into the processes of toxicity. It would be important to make use of more comprehensive RNA-Seq data to make maximum use of these additional RNA-Seq data, which is particularly significant for non-coding sequences [64].

In another study, both RNA-Seq and microarray were conducted on RNA samples from a human T-cell activation assay to show the advantages of RNA-Seq over microarray in transcriptome profiling [65]. In comparison to other research, their findings aimed to emphasize the distinction rather than the parallels, between RNA-Seq and microarray transcriptome profiling. Using the same collection of samples, the analysis of RNA-Seq and Affymetrix gene expression datasets showed a high correlation between the gene expression profiles produced by the two platforms. Compared to microarray, RNA-Seq showed a wider dynamic spectrum, allowing for the identification of more differentially expressed genes with higher fold-change. The study of the two datasets uncovered the advantage of preventing technological problems, such as cross-hybridization, non-specific hybridization, and probe detection range. RNA-Seq is devoid of the problems associated with probe redundancy and annotation, which improved the analysis of the results. Given the greater benefits of RNA-Seq, microarrays are nevertheless commonly preferred for transcription profiling studies. Despite this, RNA-Seq was shown to be superior in recognizing low abundance transcripts, differentiating biologically important isoforms and identifying genetic variants. Most researchers are also learning about RNA-Seq sequencing technology, which is more costly than microarray, and data storage is more complicated. The interpretation phase is more complex as well. Based on the current circumstances, we assume that once these obstacles are resolved, the RNA-Seq platform would be the favored approach for transcriptome research [65].

RNA expression from Dicer-positive and Dicer-knockout mouse ES cells was calculated utilizing high-throughput pyrosequencing. An observational link was identified between the sum of microRNAs sequenced and the average number of microRNAs per cell in human embryonic stem cells, of which the bulk can be accounted for by six distinct microRNA loci. Four of these miRNA loci or their human homologues have been shown to play roles in cell cycle control or oncogenesis, indicating that the miRNA pathway may be a central driver of the production of the stem cell-like properties of ES cells. Most of the previously uncharacterized miRNAs were described, of which only a small percentage are articulated at a low level and have less survival features than the well-known miRNAs. There was a low abundance of short RNAs that matched all groups of repetitive elements in cells without Dicer, and in comparison, some SINE- and simple repeat-associated short RNAs were only generated in Dicer-dependent cells. Other Dicer-dependent sequences acted similarly to miRNAs. The only Dicer appears to act as a substrate for miRNAs at a sequencing depth that exceeds the total amount of 5′ phosphorylated short RNAs per cell. This research supports the notion that miRNAs are active in anti-repetitive element protection, a property usually applied to other groups of short RNAs [66].

20.3.5.1 miRNA Preparation Through NGS

The basic method of sequencing RNA [67–69] is the same as DNA sequencing; however, some additional procedures should be performed to sequence RNA. First, the RNA can be removed from cells or tissues. For some forms of RNA, the enrichment strategy is growing to be different. Then the single-stranded RNAs are converted into the double-stranded cDNA by the process of reverse transcription. From the perspective of the cDNA, the process is the same as DNA sequencing.

Wet Experiment

The reads are the unprocessed sequencing data collected from the equipment. The microRNA sequencing [67, 70, 71] protocol is the same as messenger RNA (mRNA) sequencing, except for the library preparation, which is different for the microRNA sample. Usually, microRNAs need enrichment through gel electrophoresis. In the next phase, the appropriate gel section will be cut due to the classical scale of the microRNAs. A library is made from isolated RNA or microRNA from cells or tissues. These ligation strategies may be done sequentially or in parallel: 3′ Adapter Ligation accompanied by 5′ Adapter Ligation. To extract the cDNA, the DNA is first to be reverse transcribed. Next, the cDNA is amplified using PCR. The microRNA has a precise duration, which allows it to purify by using gel electrophoresis. Libraries are typically submitted to the sequencing firm.

Data Analysis

The data obtained by the NGS sequencing method must be evaluated [72]. The reads are also the unprocessed raw data collected from a sequencing machine. The length of the reads will vary based on the sequencing platform used. It could be better to go for longer reads if the aim of the study is to create a genome. A variety of tests can be used to figure out the consistency of sequencing. Performance is the most commonly employed measure. During the phase of base calling, every base is given a quality ranking (base recognition). Q is computed as $-10 \log E$, while Q and E denote the quality score and error rate, respectively. The percentage of bases with quality scores greater than or equal to 10, 20, and 30 is $Q10$, $Q20$, and $Q30$, respectively. $Q20$ shows that the average rate of base calling error is 1%, or that the rate of correct base calling is 99%.

Next, providing more than five bases with consistency scores less than 20 allows sequencing & PCR adaptors to be excluded. Adaptors are usually even longer when it is used in miRNA sequencing to accommodate the short distances between miRNAs. NGS sequencing methods are capable of capturing several reads simultaneously (mixed sequencing). You should classify the source of each sample based on the position of the adaptors. The mature miRNA sequence is usually 20–25 nt in length. Annotations are introduced to the sequence data by aligning them to existing miRNA databases. The miRBase (http://www.mirbase.org/) is the most well-known miRNA annotation database. This comprises known miRNAs from human, mouse, and numerous other species. PMRD [73] is a database of miRNAs expressed exclusively in plants. It comprises the largest number of model plant organisms. (http://bioinformatics.cau.edu.cn/PMRD).

The miRNA discovery method is different when no established archive includes it. The first stage is contrasting the miRNA to various small RNA databases, such as piRNA. Frequently, this method allows use of Rfam [29, 30]. The Rfam database contains information about non-coding RNAs. In case the data cannot be matched, it would be used in the de novo miRNA discovery process. The miRDeep2 program is widely used for miRNA discovery. miRcat is included in the sRNA toolkit and provides the same feature. Sequenced organisms have their genome mapped, and the miRNA precursor sequence can be deduced out from the mapped area. New miRNAs can be represented using the folding model. A new class of miRNA candidates is usually located upon the stem of a stem-loop system. MiRDeep [74], CD-miRNA [75], MiRank [76], and miRCAT [77] are other approaches for miRNA discovery.

Analysis can also be carried out to forecast the interaction between microRNA and mRNA (microRNA: mRNA) using many computational algorithms like PicTar [78], TargetScan [79, 80], and miRanda [79]. Diverse algorithms may produce diverse microRNA: mRNA predictions but use the same general criteria [81]: (1) complementarity among microRNA sequence and the mRNA sequence of 3′-UTR and (2) degree of conservation of the species-to-species microRNA site. Any algorithm provided for a prediction can create both false positives (statistically relevant but not verifiable pair) and false negatives, as well as a compromise between reducing false positive and false negative will be a compromise between tighter versus lowering false negatives. Information about various tools employed in miRNA identification is provided in Table 20.1. NGS has bioinformatic challenges due to the number of samples, but each sample can be bar-coded, helping to

Table 20.1 Tools and approaches for the identification of miRNA

Sl. No.	Softwares/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
1	comTAR	miRNA targets evolution and predicts unknown interactions	Web server, Graphical User Interface	Perl	http://rnabiology.ibr-conicet.gov.ar/comtar/	Windows, Mac OS X, Linux	Free
2	RNA22	A pattern-based approach for microRNA binding and mRNA complexes (interactive and pre-computed predictions)	Web server, Graphical User Interface	–	https://cm.jefferson.edu/rna22/	Windows, Mac OS X, Linux	Free
3	RNAhybrid	The prediction of a non-canonical target site	Web interface	Perl, Java and C/C++	https://bibiserv.cebitec.uni-bielefeld.de/rnahybrid/	Windows, Mac OS X, Linux	Free
4	miRBooking	The stoichiometric mode of action of miRNAs	Web server	Python, JavaScript and Vala	https://major.iric.ca/mirbooking/	Windows, Mac OS X, Linux	Free
5	Cupid	Reconstruction of miRNA target and ceRNA networks	–	MATLAB	http://cupidtool.sourceforge.net/	Windows, Mac OS X, Linux	Free
6	Diana-microT	microRNA target is associated with protein repression levels	Web server	–	https://web.archive.org/web/20101208180159/http://diana.cslab.ece.ntua.gr/microT/	Windows, Mac OS X, Linux	Free
7	microT-CDS	Service integration into miRNA functional analysis (Taverna and advanced multi-step functional miRNA analyses)	Web server	–	http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=microT_CDS/index&threshold=0.7&page=1	Windows, Mac OS X, Linux	Free

8	MicroTar	miRNA target prediction on the basis of complementarity as well as thermodynamic data	Command-line user interface	C	http://tiger.dbs.nus.edu.sg/microtar/	Linux, Unix	Free
9	miRror	Combinatorial analysis for miRNAs and gene targets/proteins	Web server	–	http://www.proto.cs.huji.ac.il/mirror/	Windows, Mac OS X, Linux	Free
10	PicTar	Predicting targets for single and combinations of microRNAs	Web server	–	https://web.archive.org/web/20080724163022/http://pictar.bio.nyu.edu/	Windows, Mac OS X, Linux	Free
11	PITA	Site accessibility in microRNA target recognition (UTRs and microRNAs)	Command-line user interface	Perl	https://genie.weizmann.ac.il/pubs/mir07/mir07_prediction.html	Linux	Free
12	Sylamer	The analysis visualization and interpretation of miRNA and siRNA from expression data	Web server	Java	https://www.ebi.ac.uk/research/enright	Windows, Mac OS X, Linux, Unix	Free
13	TAREF	Target prediction through mRNA UTR sequence (TArget REFiner)	Standalone	Python	https://scbb.ihbt.res.in/TAREF/programchoice.html	Linux	Free
14	p-TAREF	Animal and plant miRNA target system identification with limited development	Standalone	Python, Perl, Java and C	https://scbb.ihbt.res.in/SCBB_dept/Software.php	Linux	Free
15	TargetScan	Canonical targeting model (predicted microRNA targets in mammals)	Web server	Perl	http://www.targetscan.org/vert_72/	Windows, Mac OS X, Linux	Free
16	MiRscan	Predicting microRNA genes from pairs of conserved sequences with potential to form RNA foldbacks	Web server	–	http://hollywood.mit.edu/mirscan/index.html	Windows, Mac OS X, Linux	Free

(continued)

Table 20.1 (continued)

Sl. No.	Softwares/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
17	DoRiNA	RNA interactions in post-transcriptional regulation	Web server	Python API	https://dorina.mdc-berlin.de/	Windows, Mac OS X, Linux	Free
18	TargetFinder	To predict plant small RNA binding target sites from a sequence database (position-weighted scoring matrix)	Command-line user interface	Perl	https://github.com/carringtonlab/TargetFinder	Windows, Mac OS X, Linux	Free
19	microCLIP	The identification of transcriptome wide miRNA-target interactions (CLIP-guided detection of miRNA interactions)	Command-line user interface	R, Java	http://carolina.imis.athena-innovation.gr/diana_team/microCLIP/index.html	Uinux	Free
20	MicroPC	A detailed resource for predicting and comparing plant microRNAs	Web server	–	http://www3a.biotec.or.th/micropc/	Windows, Mac OS X, Linux	Free
21	miRanalyzer	The identification as well as analysis of microRNAs in high-throughput sequencing experiments	Standalone	Weka	https://bioinfo2.ugr.es/ceUGR/miranalyzer/	Windows, Mac OS X, Linux	Free
22	MatureBayes	Finding a miRNA precursor sequence using a naive bays classifier (discovering mature miRNAs)	Web server	Python	http://mirna.imbb.forth.gr/MatureBayes.html	Linux	Free
23	TarBase	miRNA: mRNA gene interactions, advanced information ranging from the binding site location	Web interface	–	http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=tarbasev8%2Findex	Windows, Mac OS X, Linux	Free

24	LncBase	To predict microRNA targets on long non-coding RNAs	Web interface	–	http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=lncbasev2%2Findex	Windows, Mac OS X, Linux	Free
25	miRGen	Accurate depiction of microRNA promoters as well as their regulators	Web interface	PHP framework	http://carolina.imis.athena-innovation.gr/diana_tools/web/index.php?r=mirgenv3%2Findex	Windows, Mac OS X, Linux	Free
26	mirExTra	Differential expression analysis as well as central microRNA discovery module	Web interface	–	http://carolina.imis.athena-innovation.gr/mirextra/	Windows, Mac OS X, Linux	Free
27	psRNATarget	High-throughput analysis of next-generation data	User-friendly interfaces, HTML5 APIs	–	http://plantgrn.noble.org/psRNATarget/	Linux	Free
28	ViennaRNA Web Services	Detection of minimum free energy structures, RNA–RNA hybrids/non-coding RNA detection	Standalone program, Web server	C, Perl	http://rna.tbi.univie.ac.at/	Windows, Mac OS X, Linux	Free
29	miRBase	A searchable database of information on the location and sequence of the mature miRNA sequence	Web server	–	http://www.mirbase.org/search.shtml	Windows, Mac OS X, Linux	Free
30	miRNEST	An integrative collection of plant, animal, as well as virus microRNA data	Web server, user interface	HuntMi	http://rhesus.amu.edu.pl/mirnest/copy/home.php	Windows, Mac OS X, Linux	Free
31	mirPath	To predict miRNA targets (in CDS or 3′-UTR regions)	Web interface	R	http://snf-515788.vm.okeanos.grnet.gr/	Windows, Mac OS X, Linux	Free

(continued)

Table 20.1 (continued)

Sl. No.	Softwares/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
32	PASmiR	A literature-curated database for miRNA molecular regulation in plant abiotic stress	Web server	Apache, JavaServer Pages, MySQL, Struts2, Spring, Hibernate	http://hi.ustc.edu.cn:8080/PASmiR	Windows, Mac OS X, Linux	Free
33	Rfam	A collection of RNA sequence families of structural RNAs including non-coding RNA genes as well as cis-regulatory elements (multiple sequence alignment and a covariance model)	Web server	MySQL database	https://rfam.org/	Windows, mac OS X, Linux	Free
34	mirPub	For searching publications related to microRNA molecules	Web server	–	http://diana.imis.athena-innovation.gr/DianaTools/index.php?r=mirpub	Windows, Mac OS X, Linux	Free
35	PsRobot	To analyze plant small RNA data (stem-loop small RNA and small RNA target prediction)	Command-line user interface	Perl	http://omicslab.genetics.ac.cn/psRobot/index.php	Linux	Free
36	miRPathDB	A new dictionary on microRNAs as well as target pathways	Web interface, HTML5	JavaScript	https://mpd.bioinf.uni-sb.de/	Windows, Mac OS X, Linux	Free
37	AtmiRNET	Reconstructing regulatory networks of microRNAs	Web server	R	http://atmirnet.itps.ncku.edu.tw/	Windows, Mac OS X, Linux	Free
38	PmiRExAt	Plant miRNA expression atlas database	Web server, application program interface	Java EE 6	http://pmirexat.nabi.res.in/searchdb.html	Windows, Mac OS X, Linux	Free

39	miR-PREFeR	microRNA prediction from small RNAseq data	Command-line user interface	C/C++	https://github.com/hangelwen/miR-PREFeR	Windows, Mac OS X, Linux	Free
40	PMRD	Plant microRNA database (multiple search tools)	Web interface	Perl CGI	http://bioinformatics.cau.edu.cn/PMRD/	Windows, Mac OS X, Linux	Free
41	pssRNAMiner	A plant short small RNA regulatory Cascade analysis server	Web interface (SQLite database)	Java, PERL, PHP	http://bioinfo3.noble.org/pssRNAMiner/	Windows, Mac OS X, Linux	Free
42	TAPIR	Predicts targets for plant miRNAs and the detection of miRNA–mRNA duplexes	Web interface, Standalone	–	http://bioinformatics.psb.ugent.be/webtools/tapir/	Linux	Free
43	PMTED	Plant MiRNA target expression database	Web interface	Perl	https://tools4mirs.org/software/mirna_databases/pmted/	Windows, Mac OS X, Linux	Free
44	AHD	A comprehensive genetic and phenotypic information database for plant hormone research in Arabidopsis	Web interface	Perl	http://ahd.cbi.pku.edu.cn/	Windows, Mac OS X, Linux	Free
45	miSolRNA	Tomato fleshy fruit development and ripening studies	Web server	Python	http://www.misolrna.org/	Windows, Mac OS X, Linux	Free
46	PmiRKB	Plant microRNA Knowledge Base	Web server	PHP, Javascript	http://bis.zju.edu.cn/pmirkb/	Windows, Mac OS X, Linux	Free
47	PlantDARIO	The analysis of plant small non-coding RNAs from small RNA-Seq data	Web server	Perl	http://plantdario.bioinf.uni-leipzig.de/index.py	Windows, Mac OS X, Linux	Free

(continued)

Table 20.1 (continued)

Sl. No.	Softwares/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
48	OmiRas	Differential expression analysis of miRNAs derived from small RNA-Seq data	Web server	Python, PostgreSQL, R	http://tools.genxpro.net/omiras/	Windows, Mac OS X, Linux	Free
49	Tools4miRs	One place to gather all the tools for miRNA analysis	Web server, meta-server	Python	https://tools4mirs.org/	Windows, Mac OS X, Linux	Free
50	seqcluster	Analysis of specific small RNA sequencing data	Command-line user interface	R	https://seqcluster.readthedocs.io/	Linux	Free

minimize costs. Low-cost solutions could potentially be offered by the third-generation sequencing technologies in progress [82, 83].

20.4 Application

These profiling techniques have potential applications to primarily classify new miRNAs predicted by bioinformatics methods, dissect the upregulated and downregulated profiling of diverse miRNAs in the same cells, and to compare different tissues or cells with miRNA expression profiles. These arrays have been used to decipher miRNA expression patterns during growth, differentiation, cancer, as well as disease conditions.

20.4.1 Identification with New Small RNA (sRNA) Arrays

Northern blot analysis or cloning methods often detect new miRNAs in animal and plant species. These techniques involve a significant quantity of total RNA as the starting material and often do not detect low-abundance miRNAs. A combined method to address the barriers by using computational prediction and microarray analysis has been proposed earlier [84]. Ten of the established KSHV pre-miRNAs were discovered through this process, along with a previously undetected novel pre-miRNA. Using additional computational methods, the authors established a total of 18 new Epstein-Barr viruses (EBV) pre-miRNAs that generate 22 mature miRNA molecules. Thereby more than quadrupling, the total number of hitherto reported EBV miRNAs [84]. Another computationally predicted >800 novel candidates for mammalian miRNA [85]. To establish the survival properties of miRNA genes, the authors sequenced 122 miRNAs in ten primate organisms. Good conservation is observed in hairpin stems of miRNA, and further variety in loop sequences is noticed. There was a surprising decrease in conservation at the nucleotide stage in sequences immediately flanking the miRNA hairpins. This cross-species comparative profile was used to estimate novel miRNAs. Nine-hundred seventy-six candidate miRNAs were discovered by scanning whole-genome human–mouse and human–rat alignment assemblies. The bulk of the applicant's novel vertebrates are conserved throughout other vertebrates (dog, cow, chicken, opossum, zebrafish). The outcomes of a northern blot study demonstrated that 16 of the 69 candidates had correctly converted into mature miRNAs. The expression of 179 novel candidates is supported by the inclusion of these candidates in gene clusters, as well as in literature reported since these predictions were made. These observations indicate that the human genome harbors considerably larger amounts of miRNAs than traditionally estimated [85].

20.4.2 Tissue-Specific miRNAs

Arrays of miRNA may provide gene expression "profiles" that reflect complex miRNA expression patterns which are typical to cells or tissue's microenvironmental responses. Various tissue-specific miRNAs and, in addition, certain miRNAs exhibit varying levels of expression across different tissues. miR-181, miR-155, miR-142, and miR-223 were directly expressed in 17 malignant hematopoietic cell lines [86]. They contrasted the expression profiles of malignant hematopoietic cell line-specific microRNAs (miRNAs; miR-181, miR-155, miR-142, and miR-223) to those of regular human B, T, monocytic, and granulocytic cell lineages and noticed that they differed. Despite appearing to have identical expression patterns to "normal human hematopoietic lineages," malignant cell lines reported miRNA expression patterns that were distinct from "normal human hematopoietic lineages," showing the important influence of miRNAs in human hematopoietic diseases. Their studies substantiated the importance of miRNA expression in human hematopoiesis and oncogenesis [86].

Other authors have demonstrated that some typical sRNAs could be exchanged between different cells and express their unique sRNAs [53]. They employed a new class of sRNAs that was established by integrating dynamic programming prediction, enrichment of sRNAs, and microarray analysis. Information on the laboratory techniques that their laboratories use to design capture probes and mark enriched small RNAs was included in the study. The microarray findings indicate that their tailored technologies are important in improving the array's sensitivity and specificity, discovering expression trends through various cells and discovering the differentially expressed sRNAs during the bone marrow stem cell differentiation process. That study supported that computational prediction and microarray analysis could be an advantageous way of studying known and predicted small RNAs [53].

20.4.3 miRNA's Function in Stem Cells

In tissue development, maintenance and repair of stem cells play a major role. miRNA array technology has aided in unmasking stem cell function regulation by comparing sRNA expression of various developmental stages of stem cells. An important enzyme in miRNA biogenesis has been reported in mouse embryonic stem (ES) cells with defects in differentiation and proliferation due to the loss of Dicer, while Dicer re-expression in ES mouse cells free these functions [87, 88]. Earlier Murchison et al. developed embryonic stem cell lines where the Dicer gene may be inactivated under certain conditions. Dicer deficiency threatens microRNA maturation and causes a dsRNA-triggered gene silencing error. In the absence of Dicer, small interfering RNAs are not negatively impaired in their capacity to inhibit gene expression. Interesting to remember, Dicer failure can clarify the phenotype previously observed in Dicer-null animals. Decreased expression of Dicer often reduces the abundance of transcripts from mammalian centromeres but does so without greatly influencing histone alteration status or DNA methylation at pericentromeric repeats. The models presented in that article aid us in dissecting the biological functions of the RNAi machinery in cultured mammalian cells [87]. To research Dicer feature, another group of researchers knocked out the Dcr-1 gene in embryonic stem cells (ES cells) through conditional gene targeting. They produced Dcr-null ES cells. Although these cells were defective in RNA interference (RNAi) and microRNA generation, they were still viable (miRNAs). But the ES cells were absolutely unable to distinguish in vitro or in vivo. Centromeric repeat sequences, as well as the expression of homologous small dsRNAs, are suppressed epigenetically. Restoration of Dicer expression in knockout cells reversed these phenotypes. Our results point to Dicer's presence in essential biological processes, such as stem cell differentiation and centromeric heterochromatin structure and silencing, in a mammalian organism [88]. Thus, understanding the novel insights of stem cell regeneration and differentiation have been provided by the identification of patterns of miRNA expression in stem cells, but these are in their infancy [53]. Effective combination with miRNA analysis of stem cell studies may have scientific and medical uses. miRNA biology, self-renewal, and differentiation being vital aspects of stem cell function.

20.4.4 miRNA and Cancer

In most respects, cancer is a genetic disease that includes both protein code and non-protein-coding genetic alterations. Augmenting real-time quantitative RT-PCR, miRNA arrays are used to classify and compare standard cell and tissue miRNA expression profiles with those in tumors. Different miRNA patterns of expression were connected by numerous kinds of tumors, and the miRNA profiles were different from normal cancerous tissues. For example, in >80% of tumor samples, miR-126, miR-143, and miR-145 were downregulated compared to associated normal tissues, while in 80% of tumor samples, miR-21 was found to be overexpressed. miRNA arrays were able to establish a correct diagnosis with slightly higher accuracy of poorly separated samples. The benefit of miRNA over mRNA profiling is that diagnostic profiles are clearly demonstrated in contrast to the mRNA-based classifier [42]. Previous studies show that approximately half of all miRNA genes were present in cancer-associated genomic regions and developed a key nodal point in cancer growth pathways. This indicates that miRNAs may play an important role in the pathogenesis of human cancer.

Using Affymetrix miRNA microarrays, a group of researchers examined the miRNA profiles of samples that exhibited global miRNA declines as a consequence of inducible Dicer1 deletion. An unusually high percentage of deregulated miRNAs were up-regulated, and thus even up to a quarter of deregulated miRNAs found in response to Dicer1 degradation displayed substantial up-regulation after "robust multichip average" (RMA) context correction and quantile normalization, suggesting a normalization bias. We discovered that when non-miRNA small RNAs were used in place of miRNAs, the use of cyclic loess improved the accuracy and efficiency of miRNA recognition. The findings were verified in samples from patients with prostate cancer, where cyclic loess normalization and array weights were used along with rigorous normal-exponential history correction to appropriately distinguish the largest number of decreased miRNAs and the lowest volume of false-positive up-regulated miRNAs. The observation highlighted the reality that global miRNA decreases must be compensated for when using microarrays for the identification of miRNAs that are differentially represented. The usage of cyclic loess made of non-miRNA small RNAs was found to help increase the sensitivity and specificity of miRNA profiling in cancer samples with global miRNA decrease [89].

Another group of researchers tested for miRNAs and their target genes for new markers of tumor subtype. "Gene Expression Omnibus" (GEO) database was used to view the miRNA expression profiles from breast cancer GSE38867, which included seven "ductal carcinomas *in situ* breast" (DCIS) cancer samples, seven invasive breast cancer samples, seven metastatic breast cancer samples, and seven normal breast samples. The limma package was used to classify the differentially expressed miRNAs in various subtypes of breast cancer. Using MicroRNA.org as a database source, they predicted the target genes of the miRNAs that were observed to be differentially expressed. To conduct the GO feature and KEGG pathway studies, we inserted the goal genes and their interacting genes (which STRING predicted) into DAVID [90]. The numbers of differentially expressed miRNAs found in DCIS, invasive, and metastatic breast cancer, respectively, were 21, 47, and 107. Three subtype-specific miRNAs ("hsa-miR-99a and hsa-miR-151-3p for DCIS breast cancer, hsa-miR-145 and hsa-miR-210 for invasive breast cancer, and has-miR-205 and has-miR-361-5p metastatic breast cancer") were found to be differentially expressed. A large number of different GO functions and KEGG pathways had enriched in their miRNA target and gene networks that interacted with them [90].

20.5 Limitation

There are a variety of properties peculiar to miRNAs that cause difficulties when trying to detect and measure them [12]. As an example, mature miRNAs usually have a length of 22 nt, which is too short for annealing to standard primers designed for PCR and reverse transcription. Additionally, in contrast to messenger RNAs, miRNAs lack a universal primer binding site, including a poly(A) tail, which could be used for selective enrichment or reverse transcription. miRNAs are an exceptionally unusual (~0.01%) part of the overall RNA mass, and it is, therefore, crucial that they be narrowly identified in the form of other, diverse RNA populations, including pre-and pri-miRNA precursors that also comprise the mature miRNA series. Additionally, miRNAs in a family (for example, the let-7 family) may vary by a single nucleotide, causing the ability to differentiate among the types of single-nucleotide variations to be particularly significant. Even for a single miRNA, biological samples may exhibit any degree of sequence duration heterogeneity. Oftentimes, this is attributed to so-called "isomiRs" [91–93], by the addition of nucleotides to the 3′ ends of mature miRNAs post-transcriptionally; also, post-transcriptional cleavage at the 3′ end of mature miRNAs results in truncated sequences that are shorter than the standard miRNA. Changes to the 5′ end of the miRNA can have a significant impact on its role since they may modify the sequence of the seed region, which is usually identified as nucleotides 2–8 of the miRNA and is the primary determinant of mRNA target selection [93–95]. Nucleotide additions to the 3′ end have little effect on the seed region but can still impact miRNA stability and mRNA-targeting efficacy. It is necessary to bear in mind that the degree of sequence duration heterogeneity varies between miRNAs, and the overwhelming majority of miRNAs demonstrate only mild length heterogeneity; therefore, the importance of measuring different types depends on the aims of a miRNA profiling experiment. Another issue with profiling hundreds of miRNAs in parallel is that, due to their small duration, GC content variations create a significant variability in annealing temperatures (T_m). This influences the resultant profiles, which have prejudices peculiar to miRNAs. Three well-established techniques exist, namely quantitative reverse transcription PCR (qRT-PCR), hybridization-based approaches (e.g., DNA microarrays), and high-throughput sequencing (RNA sequencing).

20.6 Conclusion and Future Perspective

In conclusion, high-throughput approaches are used for gene expression analysis, including miRNA profiling. It would be helpful to perform further analysis on profiling pre-miRNAs and pri-miRNAs, as well as convergence with broader datasets, in order to further grasp the functions of miRNAs in gene regulation and disease [12]. Additionally, miRNAs are increasingly valued to be compartmentalized within cells (e.g., miRNA within the cytoplasm versus the nucleus or miRNAs present within different protein complexes). As a consequence, novel details can be revealed by miRNA profiling in unique subcellular compartments instead of whole-cell extracts, as has historically been the case. Characterization of miRNAs associated with their targets (e.g., through PAR-CLIP or HITS–CLIP) is expected to increase as experimental approaches and sequencing technologies advance. This may have a significant payback for studying biological mechanisms in which miRNAs play a prominent role. Additionally, we should discuss that so many non-coding RNAs, like PIWI-interacting RNAs (piRNAs), and long intergenic non-coding RNAs (lincRNAs), are gradually recognized as playing important roles in cellular physiology, although it is possible that other non-coding RNA groups would be discovered. Profiling approaches like RNA-seq, which are capable of detecting all forms of RNA, are likely to shed light on the whole transcriptome in the future.

Conflicts of Interest None

Other Information Figure 20.1 (CC BY 2.0) [13] has been reused under the terms of the Creative Commons Attribution License.

References

1. Ha M, Kim VN. Regulation of microRNA biogenesis. Nat Rev Mol Cell Biol. 2014;15 (8):509–24.
2. Broughton JP, Lovci MT, Huang JL, Yeo GW, Pasquinelli AE. Pairing beyond the seed supports MicroRNA targeting specificity. Mol Cell. 2016;64(2):320–33.
3. Vasudevan S. Posttranscriptional upregulation by microRNAs. Wiley Interdiscip Rev RNA. 2012;3(3):311–30.
4. Quantitative prediction of miRNA-mRNA interaction based on equilibrium concentrations [Internet]. [cited 2021 Feb 13]. https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1001090.
5. Makarova JA, Shkurnikov MU, Wicklein D, Lange T, Samatov TR, Turchinovich AA, et al. Intracellular and extracellular microRNA: an update on localization and biological role. Prog Histochem Cytochem. 2016;51(3–4):33–49.
6. Fu G, Brkić J, Hayder H, Peng C. MicroRNAs in human placental development and pregnancy complications. Int J Mol Sci. 2013;14(3):5519–44.
7. Tüfekci KU, Oner MG, Meuwissen RLJ, Genç S. The role of microRNAs in human diseases. Methods Mol Biol. 2014;1107:33–50.
8. Paul P, Chakraborty A, Sarkar D, Langthasa M, Rahman M, Bari M, et al. Interplay between miRNAs and human diseases. J Cell Physiol. 2018;233(3):2007–18.
9. Hayes J, Peruzzi PP, Lawler S. MicroRNAs in cancer: biomarkers, functions and therapy. Trends Mol Med. 2014;20(8):460–9.
10. Wang J, Chen J, Sen S. MicroRNA as biomarkers and diagnostics. J Cell Physiol. 2016;231 (1):25–30.
11. Huang W. MicroRNAs: biomarkers, diagnostics, and therapeutics. Methods Mol Biol. 2017;1617:57–67.
12. Pritchard CC, Cheng HH, Tewari M. MicroRNA profiling: approaches and considerations. Nat Rev Genet. 2012;13(5):358–69.
13. Davis BN, Hata A. Regulation of MicroRNA biogenesis: a miRiad of mechanisms. Cell Commun Signal. 2009;7(1):18.
14. Kim VN. MicroRNA biogenesis: coordinated cropping and dicing. Nat Rev Mol Cell Biol. 2005;6(5):376–85.
15. Lee Y, Kim M, Han J, Yeom K-H, Lee S, Baek SH, et al. MicroRNA genes are transcribed by RNA polymerase II. EMBO J. 2004;23(20):4051–60.
16. Lee Y, Ahn C, Han J, Choi H, Kim J, Yim J, et al. The nuclear RNase III Drosha initiates microRNA processing. Nature. 2003;425(6956):415–9.
17. Gregory RI, Yan K-P, Amuthan G, Chendrimada T, Doratotaj B, Cooch N, et al. The microprocessor complex mediates the genesis of microRNAs. Nature. 2004;432(7014):235–40.
18. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ. Processing of primary microRNAs by the microprocessor complex. Nature. 2004;432(7014):231–5.
19. Lee Y, Jeon K, Lee J-T, Kim S, Kim VN. MicroRNA maturation: stepwise processing and subcellular localization. EMBO J. 2002;21(17):4663–70.
20. BOHNSACK MT, CZAPLINSKI K, GÖRLICH D. Exportin 5 is a RanGTP-dependent dsRNA-binding protein that mediates nuclear export of pre-miRNAs. RNA. 2004;10 (2):185–91.
21. Gregory RI, Chendrimada TP, Cooch N, Shiekhattar R. Human RISC couples MicroRNA biogenesis and posttranscriptional gene silencing. Cell. 2005;123(4):631–40.
22. A dicer-independent miRNA biogenesis pathway that requires Ago catalysis - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/20424607/.
23. mRNA and microRNA quality control for RT-qPCR analysis - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/20079844/.
24. A high-throughput method to monitor the expression of microRNA precursors - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/14985473/.

25. Chen C, Ridzon DA, Broomer AJ, Zhou Z, Lee DH, Nguyen JT, et al. Real-time quantification of microRNAs by stem-loop RT-PCR. Nucleic Acids Res. 2005;33(20):e179.
26. A universal TaqMan-based RT-PCR protocol for cost-efficient detection of small noncoding RNA - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/24149841/.
27. Improved RT-PCR assay to quantitate the pri-, pre-, and mature microRNAs with higher efficiency and accuracy - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/26294305/.
28. Quantification of distinct let-7 microRNA family members by a modified stem-loop RT-qPCR [Internet]. [cited 2021 Feb 13]. https://www.spandidos-publications.com/10.3892/mmr.2017.8297.
29. Facile means for quantifying microRNA expression by real-time PCR - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/16235564/.
30. Specific and sensitive quantitative RT-PCR of miRNAs with DNA primers. BMC Biotechnol [Internet]. [cited 2021 Feb 13]. https://bmcbiotechnol.biomedcentral.com/articles/10.1186/1472-6750-11-70.
31. miRNA length variation during macrophage stimulation confounds the interpretation of results: implications for miRNA quantification by RT-qPCR - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/30487268/.
32. miR-ID: a novel, circularization-based platform for detection of microRNAs [Internet]. [cited 2021 Feb 13]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3022285/.
33. A challenge for miRNA: multiple isomiRs in miRNAomics - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/24768184/.
34. IsomiRs: expanding the miRNA repression toolbox beyond the seed - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/30953728/.
35. A comprehensive, cell specific microRNA catalogue of human peripheral blood [Internet]. [cited 2021 Feb 13]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5766192/.
36. Elbashir SM, Lendeckel W, Tuschl T. RNA interference is mediated by 21- and 22-nucleotide RNAs. Genes Dev. 2001;15(2):188–200.
37. Johnson SM, Grosshans H, Shingara J, Byrom M, Jarvis R, Cheng A, et al. RAS is regulated by the let-7 microRNA family. Cell. 2005;120(5):635–47.
38. Nejad C, Pépin G, Behlke MA, Gantier MP. Modified polyadenylation-based RT-qPCR increases selectivity of amplification of 3′-MicroRNA Isoforms. Front Genet [Internet]. 2018 [cited 2021 Feb 13];9. https://www.frontiersin.org/articles/10.3389/fgene.2018.00011/full.
39. Oberthuer A, Juraeva D, Li L, Kahlert Y, Westermann F, Eils R, et al. Comparison of performance of one-color and two-color gene-expression analyses in predicting clinical endpoints of neuroblastoma patients. Pharmacogenomics J. 2010;10(4):258–66.
40. Schwarz R, Joseph B, Gerlach G, Schramm-Glück A, Engelhard K, Frosch M, et al. Evaluation of one- and two-color gene expression arrays for microbial comparative genome hybridization analyses in routine applications. J Clin Microbiol. 2010;48(9):3105–10.
41. Cummins JM, He Y, Leary RJ, Pagliarini R, Diaz LA, Sjoblom T, et al. The colorectal microRNAome. Proc Natl Acad Sci U S A. 2006;103(10):3687–92.
42. Lu J, Getz G, Miska EA, Alvarez-Saavedra E, Lamb J, Peck D, et al. MicroRNA expression profiles classify human cancers. Nature. 2005;435(7043):834–8.
43. Babak T, Zhang W, Morris Q, Blencowe BJ, Hughes TR. Probing microRNAs with microarrays: tissue specificity and functional inference. RNA. 2004;10(11):1813–9.
44. Bentwich I. Identifying human MicroRNAs. In: Paddison PJ, Vogt PK, editors. RNA interference [internet]. Berlin: Springer; 2008 [cited 2021 Feb 13]. p. 257–69. (Current topics in microbiology and immunology). https://doi.org/10.1007/978-3-540-75157-1_12.
45. Baskerville S, Bartel DP. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA. 2005;11(3):241–7.
46. An oligonucleotide microchip for genome-wide microRNA profiling in human and mouse tissues - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/15210942/.

47. Microarray analysis of microRNA expression in the developing mammalian brain - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/15345052/.
48. Development of a micro-array to detect human and mouse microRNAs and characterization of expression in human organs - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/15616155/.
49. MicroRNA expression profiling of single whole embryonic stem cells [Internet]. [cited 2021 Feb 13]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1351374/.
50. Thomson JM, Parker J, Perou CM, Hammond SM. A custom microarray platform for analysis of microRNA gene expression. Nat Methods. 2004;1(1):47–53.
51. A microRNA array reveals extensive regulation of microRNAs during brain development - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/13130141/.
52. An optimized isolation and labeling platform for accurate microRNA expression profiling - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/16043497/.
53. Yin JQ, Zhao RC. Identifying expression of new small RNAs by microarrays. Methods. 2007;43(2):123–30.
54. Wang H, Ach RA, Curry B. Direct and sensitive miRNA profiling from low-input total RNA. RNA. 2007;13(1):151–9.
55. Kloosterman WP, Wienholds E, de Bruijn E, Kauppinen S, Plasterk RHA. In situ detection of miRNAs in animal embryos using LNA-modified oligonucleotide probes. Nat Methods. 2006;3(1):27–9.
56. Davison TS, Johnson CD, Andruss BF. [2] Analyzing micro-RNA expression using microarrays. In: Methods in enzymology [Internet]. Academic; 2006 [cited 2021 Feb 13]. p. 14–34. (DNA Microarrays, Part B: Databases and Statistics; vol. 411). https://www.sciencedirect.com/science/article/pii/S0076687906110022.
57. Operational criteria for selecting a cDNA microarray data normalization algorithm [Internet]. [cited 2021 Feb 13]. https://www.spandidos-publications.com/or/15/4/983.
58. Microarray data analysis: from disarray to consolidation and consensus - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/16369572/.
59. Quackenbush J. Microarray data normalization and transformation. Nat Genet. 2002;32(Suppl):496–501.
60. Allison DB, Cui X, Page GP, Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nat Rev Genet. 2006;7(1):55–65.
61. Evaluation of quantitative miRNA expression platforms in the microRNA quality control (miRQC) study. Nat Methods [Internet]. [cited 2021 Feb 13]. https://www.nature.com/articles/nmeth.3014.
62. Barad O, Meiri E, Avniel A, Aharonov R, Barzilai A, Bentwich I, et al. MicroRNA expression detected by oligonucleotide microarrays: system establishment and expression profiling in human tissues. Genome Res. 2004;14(12):2486–94.
63. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat Biotechnol [Internet]. [cited 2021 Feb 13]. https://www.nature.com/articles/nbt.2957.
64. Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, et al. Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Front Genet [Internet]. 2019 [cited 2021 Feb 20];9. https://www.frontiersin.org/articles/10.3389/fgene.2018.00636/full.
65. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One [Internet]. 2014 [cited 2021 Feb 20];9(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3894192/.
66. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. Fast folding and comparison of RNA secondary structures. Monatshefte Chem. 1994;125(2):167–88.
67. Friedländer MR, Chen W, Adamidi C, Maaskola J, Einspanier R, Knespel S, et al. Discovering microRNAs from deep sequencing data using miRDeep. Nat Biotechnol. 2008;26(4):407–15.

68. Auer PL, Doerge RW. Statistical design and analysis of RNA sequencing data. Genetics. 2010;185(2):405–16.
69. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. Nat Rev Genet. 2011;12(2):87–98.
70. Bar M, Wyman SK, Fritz BR, Qi J, Garg KS, Parkin RK, et al. MicroRNA discovery and profiling in human embryonic stem cells by deep sequencing of small RNA libraries. Stem Cells. 2008;26(10):2496–505.
71. Creighton CJ, Reid JG, Gunaratne PH. Expression profiling of microRNAs by deep sequencing. Brief Bioinform. 2009;10(5):490–7.
72. Hu Y, Lan W, Miller D. Next-generation sequencing for MicroRNA expression profile. Methods Mol Biol. 1617;2017:169–77.
73. Zhang Z, Yu J, Li D, Zhang Z, Liu F, Zhou X, et al. PMRD: plant microRNA database. Nucleic Acids Res. 2010;38(Database issue):D806–13.
74. Discovering microRNAs from deep sequencing data using miRDeep - PubMed [Internet]. [cited 2021 Feb 13]. https://pubmed.ncbi.nlm.nih.gov/18392026/.
75. Tyagi S, Vaz C, Gupta V, Bhatia R, Maheshwari S, Srinivasan A, et al. CID-miRNA: a web server for prediction of novel miRNA precursors in human genome. Biochem Biophys Res Commun. 2008;372(4):831–4.
76. Xu Y, Zhou X, Zhang W. MicroRNA prediction with a novel ranking algorithm based on random walks. Bioinformatics. 2008;24(13):i50–8.
77. Moxon S, Schwach F, Dalmay T, Maclean D, Studholme DJ, Moulton V. A toolkit for analysing large-scale plant small RNA datasets. Bioinformatics. 2008;24(19):2252–3.
78. Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, et al. Combinatorial microRNA target predictions. Nat Genet. 2005;37(5):495–500.
79. Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell. 2005;120(1):15–20.
80. Most mammalian mRNAs are conserved targets of microRNAs [Internet]. [cited 2021 Feb 13]. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2612969/.
81. Sethupathy P, Megraw M, Hatzigeorgiou AG. A guide through present computational approaches for the identification of mammalian microRNA targets. Nat Methods. 2006;3(11):881–6.
82. Ansorge WJ. Next-generation DNA sequencing techniques. New Biotechnol. 2009;25(4):195–203.
83. Mardis ER. Anticipating the $1,000 genome. Genome Biol. 2006;7(7):112.
84. Grundhoff A, Sullivan CS, Ganem D. A combined computational and microarray-based approach identifies novel microRNAs encoded by human gamma-herpesviruses. RNA. 2006;12(5):733–50.
85. Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, Cuppen E. Phylogenetic shadowing and computational identification of human microRNA genes. Cell. 2005;120(1):21–4.
86. Ramkissoon SH, Mainwaring LA, Ogasawara Y, Keyvanfar K, McCoy JP, Sloand EM, et al. Hematopoietic-specific microRNA expression in human cells. Leuk Res. 2006;30(5):643–7.
87. Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ. Characterization of dicer-deficient murine embryonic stem cells. Proc Natl Acad Sci U S A. 2005;102(34):12135–40.
88. Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, Jenuwein T, et al. Dicer-deficient mouse embryonic stem cells are defective in differentiation and centromeric silencing. Genes Dev. 2005;19(4):489–501.
89. Wu D, Hu Y, Tong S, Williams BRG, Smyth GK, Gantier MP. The use of miRNA microarrays for the analysis of cancer samples with global miRNA decrease. RNA. 2013;19(7):876–88.
90. Sun E-H, Zhou Q, Liu K-S, Wei W, Wang C-M, Liu X-F, et al. Screening miRNAs related to different subtypes of breast cancer with miRNAs microarray. Eur Rev Med Pharmacol Sci. 2014;18(19):2783–8.

91. Podolska A, Kaczkowski B, Litman T, Fredholm M, Cirera S. How the RNA isolation method can affect microRNA microarray results. Acta Biochim Pol. 2011;58(4):535–40.
92. Cloonan N, Wani S, Xu Q, Gu J, Lea K, Heater S, et al. MicroRNAs and their isomiRs function cooperatively to target common biological pathways. Genome Biol. 2011;12(12):R126.
93. Wyman SK, Knouf EC, Parkin RK, Fritz BR, Lin DW, Dennis LM, et al. Post-transcriptional generation of miRNA variants by multiple nucleotidyl transferases contributes to miRNA transcriptome complexity. Genome Res. 2011;21(9):1450–61.
94. Katoh T, Sakaguchi Y, Miyauchi K, Suzuki T, Kashiwabara S, Baba T, et al. Selective stabilization of mammalian microRNAs by 3′ adenylation mediated by the cytoplasmic poly (A) polymerase GLD-2. Genes Dev. 2009;23(4):433–8.
95. Wei C, Salichos L, Wittgrove CM, Rokas A, Patton JG. Transcriptome-wide analysis of small RNA expression in early zebrafish development. RNA. 2012;18(5):915–29.

21 Computational Approaches in Identifying Long Non-coding RNA

Manoj Kumar Gupta, N. Rajesh, S. Sabarinathan, Gayatri Gouda, Ravindra Donde, Menaka Ponnana, Goutam Kumar Dash, Pallabi Pati, Sushil Kumar Rathore, Ramakrishna Vadde, and Lambodar Behera

Abstract

Long non-coding RNA (lncRNA) is the largest non-protein and functional RNA. The majority of lncRNAs are functionally uncharacterized. Thus, researchers are employing both experimental as well as computational approaches to characterize unknown lncRNAs. Recently, the majority of lncRNAs have been characterized using transcriptome sequencing datasets under different conditions. The information on these transcripts was mainly restricted to genomic loci and expression

S. Sabarinathan, Gayatri Gouda, Ravindra Donde, Menaka Ponnana and Goutam Kumar Dash contributed equally with all other contributors.

M. K. Gupta · G. Gouda · R. Donde · G. K. Dash · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

M. Ponnana
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Plant Physiology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_21

patterns. This discrepancy in lncRNAs' functional understanding has largely been due to the lack of methods to classify basic genome-scale bio-molecular interactions and resources that systematically archive these interactions. Additionally, experimental methods are time-consuming and cost-effective. To overcome this, various bioinformatics tools have been developed to predict lncRNA. Thus, in this chapter, the authors attempted to understand the underlying principle of various computational approaches to detect lnRNAs. Information obtained reveals that lncRNA may be annotated by employing its coding potential and sequence conservation, folding algorithms, and interactions. Additionally, several databases have also been developed to help researchers detect lncRNAs from the large genomic dataset. Though the result obtained from these tools and databases is useful, a systematic integrative and metadata analysis is also required to understand diverse lncRNA regulatory mechanisms of action at various levels. Further attempts would be made to annotate their features, which is beneficial in understanding the underlying cell biology.

Keywords

Bioinformatic tools · Computational approaches · Database · Long non-coding RNA

Abbreviations

CSF	Codon substitution score
eRNAs	Enhancer-related RNAs
lincRNAs	Long intergenic RNAs
lncRNA	Long noncoding RNA
ncRNA	Non-encoding RNA
Profile-HMM	Profile hidden Markov models

21.1 Introduction

A non-encoding RNA (ncRNA) as the name suggests is a non-encoding, non-protein, functional RNA, categorized into microRNA (miRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and long ncRNA (lncRNA) depending on their function and length [1–3]. Since the last decade, ncRNAs have been one of the most discussed topics in the genomics field. Non-coding RNAs have been primarily categorized into small and lncRNAs based on their size. This repertory has been substantially extended to improve technology for records with single nucleotide polymorphism [4, 5]. However, there has been no detailed examination of the functionality of these findings. The lncRNA class also contains transcripts identified as macroRNAs and vlincRNAs by some authors. The MacroRNAs are long, unspliced, and can readily form secondary structures and transcribed RNAs with RNA pol II. MacroRNAs like Airn [6] or KCNQ1OT1 play a key role in modulating imprinting in normal tissues. VlincRNAs, a subset of human transcriptome covering the space between 50 and 700 kb, represent or rather long intergenic RNAs, on the other hand. These transcripts are found in different tumors and demonstrate cellular expression. Pluripotent or the malignancy level in certain tumors has also been shown to be associated with them [7]. High-level techniques have allowed large-scale non-code transcripts known as circular RNAs to be detected. Transcripts, which have distinct molecular roles, are often included in the long non-coding RNA repertoire. Those functional attributes can, depending on the type of molecular interaction with other cell biomolecules, be defined as guides, decoys, signals, and scaffolds [8].

ncRNAs have major roles in living organisms, like peptide bond formation during translation mediated by rRNA [9], transcription and post-transcriptional gene expression regulation mediated by miRNA of gene expression [10], and imprinting, X inactivation, and epigenetic marks regulation mediated by lncRNA [11–13]. In addition, they also have immense significance in the sense of numerous diseases. The cluster miR-17-92 acts as an oncogene, while miR-15a–miR-16-1 acts as a tumor suppressor [14].

Even if they aren’t translated into proteins, lncRNAs are functional. Indeed, as early research shows the central role of Xist in X-chromosome inactivation, an increasing number of studies have identified various roles for lncRNAs in several cellular processes, including gene imprinting, differentiation and growth, antiviral reaction, and vernalization in plants [15]. Some lncRNAs have been shown to interact with chromatin-modifying complexes, to be active in the conformation of nuclear domains or in the functioning of transcriptional enhancers. Others have been seen to intervene with the transcriptional machinery and preserve the stability of nuclear speckles [16–19]. Several human conditions, such as heart diseases, diabetes, and intracranial aneurysm, are associated with lncRNA named as ANRIL [20]. Considering this, to date, numerous experimental approaches have been developed to detect lncRNA. However, the major drawbacks of the experimental methods are increased time and cost. Considering these, numerous bioinformatics approaches were developed to predict ncRNAs, including lncRNA. Computational methodologies combine genome-scale databases, expression patterns, motifs, and structure annotations. These offer vast opportunities for interpreting the functional role of lncRNAs. Thus, in this chapter, we present an overview of computational approaches to classify and annotate lncRNAs functionally.

21.2 Long Non-coding RNAs and Their Mechanisms

LncRNAs are a diverse class of ncRNA comprising transcripts longer than 200 nucleotides length but not encoding proteins [21]. Within this lncRNA grouping, they are often categorized based on their genomic position and majorly as “enhancer-related RNAs” (eRNAs), intronic RNAs, transcribed ultraconserved RNAs, natural antisense transcripts (NATs), and long intergenic RNAs (lincRNAs) [21]. LncRNA genes are encoded either in the antisense or sense DNA strand and are found within a protein-coding gene or inside the introns of genes [22]. Unlike mRNA, however, lncRNAs are translated by RNA polymerase II. Many transcripts are multi-exonic, have a poly-A tail, have a $5'$ cap, and experience RNA splicing. Active promoters of such genes are typically labeled with H3K4me3, and gene bodies show H3K36me3 histone modifications [23]. Unlike protein-coding genes, lncRNAs do not have usable initiation and termination codons, and thus do not contribute to the translation of protein [24]. The proteins are expressed at far lower amounts than the protein-coding equivalents and have comparatively poor evolutionary conservation. While not being strongly conserved, the expression pattern of lncRNAs has been shown to be highly relevant to cell and tissue forms.

Regarding the mode of action of these regulatory factors, two researchers suggested four large categories of action [8]. Signaling LncRNAs constitute a class in which there is a large degree of spatial and temporal precision. They play a role in signal transduction. After transcription, these signaling lncRNAs trigger signaling pathways in response to a stimulus. Their existence can also signify a specific cell developmental state, disorder, or behavior (Fig. 21.1) [8]. Another way lncRNAs control their targets is by serving as decoy molecules that inhibit RNA binding factors from binding to their partners. By impeding chromatin remodeling, transcription factors, and microRNAs in their target genes, decoy lncRNAs may negatively affect downstream results [8]. Interestingly, miRNAs can attack lncRNAs directly and affect their modulation of transcription and vascular function. Assisted by their ability to bind protein and also base pair with target sequences, lead lncRNAs are responsible for detecting transcriptional regulators to particular areas. In like manner to the roles performed by guide lncRNAs, scaffold lncRNAs can mediate protein–protein interactions by creating intricate protein–protein complexes [8]. LncRNAs are a distinct group of regulatory elements influencing transcription that share common characteristics.

21.3 lncRNA Annotation Strategies

Several different algorithms are used to categorize, describe, annotate, and identify RNA molecules.

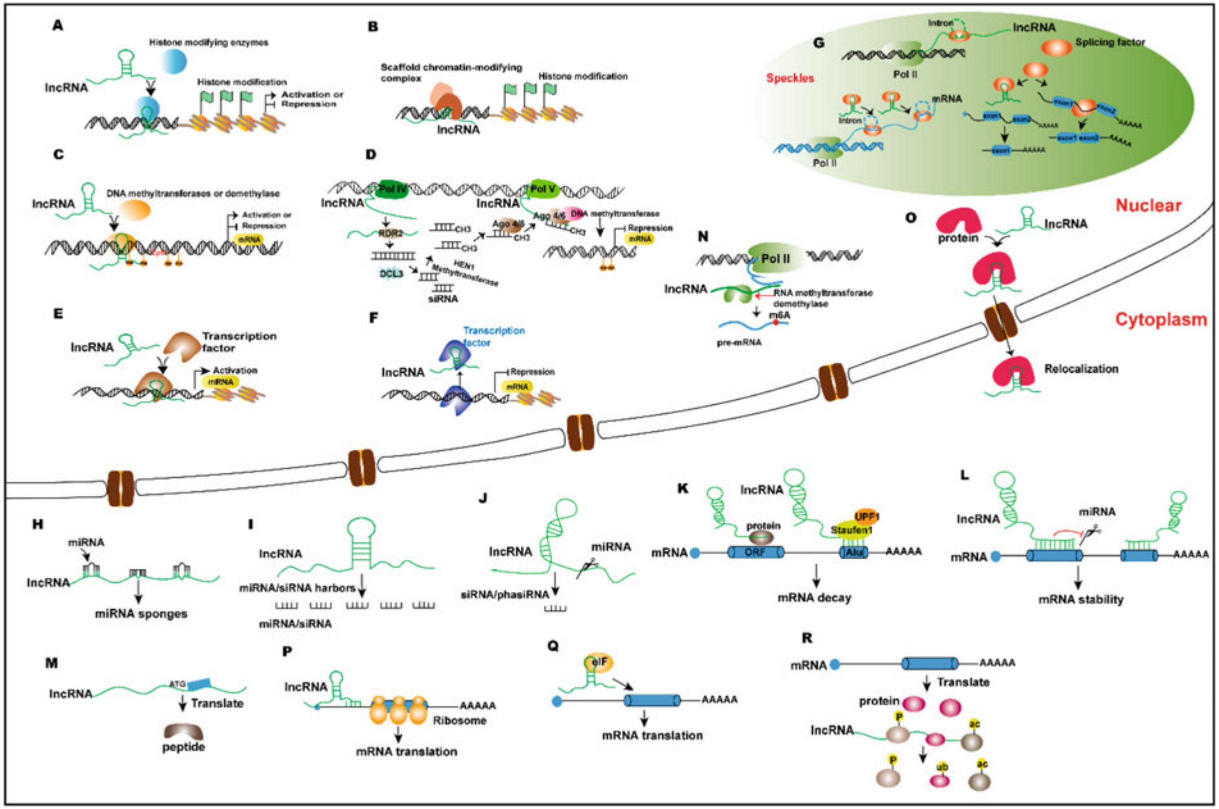


Fig. 21.1 Regulatory mechanisms of lncRNAs at the genome level. (*A*) lncRNAs interact with histone-modifying enzymes to activate or repress gene transcription. (*B*) lncRNAs recruit histone-modified complexes or act as scaffolds for multiple histone modifiers to regulate histone modification of genes and

Fig. 21.1 (continued) thereby regulate gene transcription. (*C*) lncRNAs recruit DNA methyltransferases or demethylases to regulate the target gene transcription. (*D*) Pol IV/V transcribed lncRNAs are involved in RNA-dependent DNA methylation, thus activating or repressing gene transcription. (*E*, *F*) lncRNAs interact with transcription factors to activate or repress gene expression. (*G*) lncRNAs interact with splicing factors or proteins to regulate the mRNA alternative splicing; splicing factors directly regulate the lncRNA's alternative splicing in speckles. (*H*) lncRNAs act as miRNA sponges that regulate target gene expression. (*I*) lncRNAs act as miRNA or small interfering RNAs (siRNA) precursors. (*J*) miRNAs target lncRNAs to produce siRNA or phased small-interfering RNAs (phasiRNAs). (*K*) lncRNAs are involved in the Staufen1-mediated mRNA decay, and lncRNAs bind to proteins and mediate mRNA decay. (*L*) lncRNAs directly bind to mRNA and regulate mRNA stability or competitively bind to mRNA to improve mRNA stability. (*M*) lncRNAs can be translated to peptides. (*N*) lncRNAs interact with RNA methyltransferases or demethylases, and thus regulate mRNA expression. (*O*) lncRNAs combine with proteins to regulate protein localization. (*P*) lncRNAs interact with mRNAs and affect mRNA translation. (*Q*) lncRNAs bind the translation initiation complex eIF (eukaryotic initiation factor) to regulate mRNA translation. (*R*) lncRNAs interact with proteins and control protein phosphorylation, acetylation, and ubiquitination at the post-translation level. (Adapted from [25])

21.3.1 Coding Potential and Sequence Conservation

Statistical methods for gene finding have sought to quantify small differences in DNA sequences among non-coding and coding regions of DNA. ORF sequences are calculated by identifying the start and stop codon position in the translated protein. Gene identification methods depend on codon-explicit features in several genomes regions, which represent its function. A codon use array is employed to help in evaluating the possible codon use of each gene [26]. LncRNA annotation may be complicated by the presence of certain transcripts, but in reality are coding for proteins. A significant number of lncRNAs identified with the "ambiguous ORF" biotype are thought to be protein-coding. Still, these RNAs have more than one ORF, and are therefore erroneously categorized as lncRNAs. Another cause of false positives is those brief ORF and protein codon occurrences that can be quickly ignored and the biological complexity of transcripts that function at both protein and RNA level [27]. The collection of a small number of protein sequences is also used to boost the specificity of motifs and the reliability of screening for lncRNAs. The "codon substitution score" (CSF) is also employed to classify uncommon transcripts and identify lncRNAs. The CSF scores are focused on nucleotide substitution trends found in protein-coding transcripts, as well as employ an empirical codon substitution matrix of Ka/Ks ratios, defined as "the number of non-synonymous substitutions per non-synonymous site (Ka) to the number of synonymous substitutions per synonymous site (Ks)" [28]. Lower levels of non-synonymous modifications or silent changes in codons are indicative for the conservation of protein coding genes. This matrix combines both transversion and transformation probabilities to assess evolution for conserved coding areas [28, 29]. The CSF score analyses the mutation frequencies of non-coding and coding sequences, while the substitution rates may be used to evaluate the molecular evolution. From one perspective, it is an approximation of how likely the exonic sequences potentially evolved through substitutions.

Assuming the null hypothesis, the CSF helps one to find the genuinely special individual lncRNAs. Alignment and homology identification utilizing sequence data are commonly employed in bioinformatics and can identify homologous sequences. "Profile hidden Markov models (Profile-HMM)" can generate position-specific profiles sequences from a multiple sequence alignment. These profiles may be scanned to identify other matches to which they are identical. HMMER [30] is a widely employed algorithm for creating Profile-HMMs: it allows for rigorous modeling of coding regions with differential conservation and indels rates, culminating in the more sensitive detection of exonic regions [30, 31]. The Profile-HMM method suggests that new non-coding RNA sequences may be detected by similarities to previously identified sequence profiles as well as common motifs, like CGIs, Alu repeats, and T-UCRs [32].

"Coding Potential Calculator" (CP Calculator) is a machine-learning classifier built on a support vector machine (SVM) that uses a number of sequence-based features (ORF volume and sequence similarity) to identifying potential lncRNAs. CPC is a system that computes ORF duration by utilizing the hexamer frequency as well as using multiple dynamic frameshift models. The frameshift model employs penalties in the design of a splice site to regulate unnecessary insertions or deletions, while edit distance is used to allow for mutations within lncRNA transcripts. The program then used BlastX to classify recognized protein sequences, prior to actually outputting a binary classification and trust score by SVM.

21.3.2 Folding Algorithms

The identification of lncRNAs might not be necessary to evaluate their role; however, higher order knowledge, including other genomic features, may be more useful. In reality, thermodynamic folding can be employed to categorize lncRNA sequences into categories and recognize common motifs among folded lncRNAs. The reason for using folding algorithms in lncRNA annotation is the belief that lncRNA plays the tasks of a scaffold by assisting in the creation of RNA/DNA hybrids [33, 34]. The scaffold assumption means that the secondary structures of their scaffold proteins are close enough to enable scaffold component proteins to attach to the scaffold. Fold identification was an excellent method to find a similar collection of secondary and tertiary motifs to classify short non-coding RNAs. Using these detailed studies, thousands of intergenic lncRNA genes were identified in mouse cells [23]. Pseudoknots are key functional motifs in deciding the secondary structure of several intergenic introns [35]. During transcription, RNA folds into pseudoknots as well as hairpin motifs to provide minimum free energy. Algorithms like Randfold can integrate hairpins [36] and pseudoknots [37] to achieve higher precision structure and aid in detecting increasingly complicated motifs.

21.3.3 LncRNA Interactions

The variations between different lncRNAs and how they communicate with RNA-binding proteins (RBP) can be defined by their interactions with these proteins. Mass spectra of lncRNA-associated peptides may be used to track map, control, and quantity of lncRNA–ribosomal protein complexes interaction, which in turn promote the creation of a "coding/non-coding gene co-expression network (CNC)." The localization of lncRNA binding sites and motifs gives insight into the genetic associations responsible for hereditary disorders and polygenic cancers. Recent research found, on average 1.1% of RBP binding sites in human lncRNAs, totaling 21,073 RBP–lncRNA associations in 14 cancer forms [38, 39] which indicate a crucial regulatory relationship between RBP and lncRNAs, independent of biotype. These functional analysis approaches may be used to forecast and evaluate a particular lncRNA–protein complex. Preprocessing, peak calling, differential analysis applications, and StarBase employ the RBP–lncRNA relation for prediction, as well as an annotation.

It is also established that a regulatory layer occurs in the targeting of lncRNAs by microRNAs (miRNAs) [33, 40–46]: it is owing to the existence of "miRNA recognition elements (MREs)" sites in certain lncRNAs. MREs' motifs are untranslated into miRNA-binding regions situated in both 3′-UTR as well as the "coding region of gene" (CDS) regions, which in turn leads to mRNA degradation or translational repression [47]. These miRNA–lncRNA associations are thought to play a vital role in biological processes like organogenesis, development [48], and disease pathophysiologies [49, 50]. For instance, lncRNAs, MRAK081523 and MRAK088388, play a modulatory function through regulating N4bp2 as well as Plxna4 expression via let-7i-5p and miR-29b-3p sequestration [51]. MiRcode is a method initially developed for miRNA binding sites prediction. It uses seed complementarity, evolutionary survival, and Gencode annotation to forecast lncRNA. The microRNA (miRNA) code comprises over 10,000 annotations in the human genome.

21.4 Databases and Public Repositories

To date, numerous databases and tools have been developed for the identification of lncRNA (Table 21.1). One of the first databases of ncRNAs with proven or suspected regulatory roles was created by Barciszewski's group in 2003 [52]. Their non-coding RNA database was the first archive comprising nucleotide sequences (obtainable in FASTA format), brief explanations of the activities of individual ncRNAs, literature references and GenBank accession numbers. According to the details, the number of ncRNAs in the database was less than 40, besides homologs and miRNAs. At present, the database comprising ~30,000 human sequences from 99 distinct organisms and three different realms of life, namely Eukaryota, Bacteria, and Archaea. The primary source of sequences used in the database was GenBank, and the additional annotation details for human and mouse ncRNAs were also obtained from "H-inviational Integrated Database of Annotated Human Genes version 3.4" and FANTOM3, respectively.

Another database named Rfam has been created in the same year. It conducts several sequence alignments and covariance models describing non-coding RNA families [53]. At first, the Rfam 1.0 database comprises over 50,000 RNA families belonging to 25 RNA families. After incorporation with more advanced RNA databases like miRBase, IRESite, Pseudobase, snoRNABase, the plant snoRNA database, TransTerm, and the Yeast snoRNA database, the next edition (Rfam 9.1) would include more than 700 completely new families, hitting a total of more than 1300 [54]. In 2005, Mattick's group revealed the creation of RNAdb (http://research.imb.uq.edu.au/RNAdb), a robust human ncRNA database comprising over 800 specific experimentally characterized ncRNAs, correlated with diseases and/or developmental processes. This database was further applied in 2007 with the RNAdb 2.0 where the authors presented even nucleotide sequences and annotations for tens of thousands of non-housekeeping ncRNAs, including a wide variety of mammalian microRNAs, small nucleolar RNAs, as well as ncRNAs expected to occur utilizing structural features and alignments.

Table 21.1 List of tools for identification o lncRNA

Sl. No.	Software's/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
1	NONCODE	An integrated knowledge database dedicated to non-coding RNAs	Web server	MySQL	http://www.noncode.org/	Linux, Mac OS X, Microsoft Windows	Free
2	AnnoLnc2	To systematically annotate novel lncRNAs for human and mouse	Web server, standalone	–	http://annolnc.gao-lab.org/	Linux	Free
3	COME	Coding potential calculator based on multiple features (predicts the coding potential for a given transcript)	Command line user interface	R	https://github.com/lulab/COME	Linux, Mac OS X, Microsoft Windows	Free
4	CPAT	RNA coding potential assessment tool	Command line user interface, Web server	R, Python, NumPy	http://code.google.com/p/cpat/	Linux, Mac OS X, Microsoft Windows	Free
5	Pfamscan	Predicting active site residue annotations	Web server	Perl, MySQL	https://www.ebi.ac.uk/Tools/pfa/pfamscan/	Linux, Mac OS X, Microsoft Windows	Free
6	phyloCSF	To distinguish protein coding and non-coding regions	Command line user interface	Caml	https://github.com/mlin/PhyloCSF/wiki/	Linux, Mac OS X	Free
7	PNRD	Plant non-coding RNA database	Web server	LAMP	http://structuralbiology.cau.edu.cn/PNRD/	Linux, Mac OS X, Microsoft Windows	Free
8	LncReg	A database for lncRNA-associated regulatory networks	Web server, user-friendly interface	HTML/CSS and JavaScript	http://bioinformatics.ustc.edu.cn/lncreg/	Linux, Mac OS X, Microsoft Windows	Free

9	RNAsoft	A suite of RNA secondary structure prediction and design	Web server, standalone	C++, Perl	http://www.RNAsoft.ca/	Linux	Free
10	RNA shapes	Secondary structure prediction tools (RNA alishapes and PKNOTS RG)	Command line user interface	B ELLMAN’S GAP	http://bibiserv.cebitec.uni-bielefeld.de/rnashapesstudio/	Linux, Mac OS X, Microsoft Windows	Free
11	RNA Movies 2	Sequential animation of RNA secondary structures	Command line user interface	Java	https://bibiserv.cebitec.uni-bielefeld.de/rnamovies	Linux, Mac OS X, Microsoft Windows	Free
12	Infernal	Inference of RNA Alignment (hidden Markov model)	Command line user interface	C, Perl	http://eddylab.org/infernal/	Linux, Mac OS X	Free
13	PlantcircBase	Plant Circular RNA Database	Web server	–	http://ibi.zju.edu.cn/plantcircbase/	Linux, Mac OS X, Microsoft Windows	Free
14	PlantCircNet	A database for plant circRNA–miRNA–mRNA regulatory networks (circRNA information and expression profiles)	Web server	LAMP	http://bis.zju.edu.cn/plantcircnet/index.php	Linux, Mac OS X, Microsoft Windows	Free
15	PlantNATsDB	A database of plant natural antisense transcripts	Web server, Graphical interface	MySQL, JavaScript	http://bis.zju.edu.cn/pnatdb/	Linux, Mac OS X, Microsoft Windows	Free
16	RNAcentral	The non-coding RNA sequence database	Web server	Perl	http://rnacentral.org	Linux, Mac OS X, Microsoft Windows	Free

(continued)

Table 21.1 (continued)

Sl. No.	Software's/ tools	Description	Software interface	Language developed/ using/working	Link	Operating system	Cost
17	RNAInter	RNA interactome repository with increased coverage and annotation (RNA-RNA and RNA-protein interactions)	Web server	–	http://www.rna-society.org/raid/	Linux, Mac OS X, Microsoft Windows	Free
18	LncRNASNP	Functional SNPs and mutations in human and mouse lncRNAs	Web interface	Python	http://bioinfo.life.hust.edu.cn/lncRNASNP/#!/	Linux, Mac OS X, Microsoft Windows	Free
19	LGC	Characterization and identification of long non-coding RNAs based on feature relationship	Web server	–	https://bigd.big.ac.cn/lgc/calculator	Linux, Mac OS X, Microsoft Windows	Free
20	CPAT	Coding-Potential Assessment Tool (an alignment-free logistic regression model)	Web server	C, Python	http://lilab.research.bcm.edu/cpat/	Linux, Mac OS X, Microsoft Windows	Free
21	LncExpDB	Expression database of human long non-coding RNAs	Web interface	Java Script Pages, MySQL, Asynchronous JavaScript, and XML	https://bigd.big.ac.cn/lncexpdb/		
22	lncRScan-SVM	Predicting Long Non-Coding RNAs Using Support Vector Machine	Web interface	Python	https://sourceforge.net/projects/lncrscansvm/?source=directory	Linux, Mac OS X, Microsoft Windows	Free
23	PLEK	Predict lncRNAs and mRNAs based on k-mer scheme	Command line user interface	C, C++, Python	https://sourceforge.net/projects/plek/	Linux	Free

24	FEELnc	Flexible Extraction of Long non-coding RNAs	Command line user interface	R, Perl	https://github.com/tderrien/FEELnc	Linux, Mac OS X, Microsoft Windows	Free
25	RNAplonc	Pattern recognition analysis on long non-coding in plants	Command line user interface	Perl, Java	https://github.com/TatianneNegri/RNAplonc/	Linux	Free
26	PLncPRO	Plant lncRNA Prediction by Random forests (discovery of abiotic stress-responsive lncRNAs in rice and chickpea)	Web server	Python	http://ccbb.jnu.ac.in/plncpro/	Linux, Mac OS X, Microsoft Windows	Free
27	slncky	Filters a high-quality set of lncRNA from reconstructed RNA-seq data	Command line user interface	Python	http://slncky.github.io/	Linux, Mac OS X, Microsoft Windows	Free

circRNA circular RNA, *miRNA* microRNA, *lncRNAs* long non-coding RNAs, *LAMP* Linux Apache MySQL and PHP

A second curated database, the "H-Invitational Database (H-InvDB)" (http://www.h-invitational.jp/), is a collaborative effort of researchers working on the "Human Full-length cDNA Annotation project" [55]. The H-InvDB (release 3.4, August 2006) database contains more than 1700 putative ncRNAs identified by the omission of any open reading frame as well as by not belonging to the pseudogene category.

Over a period of time, many other databases appeared to fill the gaps in classifying other ncRNAs like SRP RNAs, tmRNAs, or RNase P RNAs, as well as other ncRNAs defined as per cellular localization, feature, or sedimentation factors (i.e., 6S RNA, 5.3S RNA, etc.). Thus, to standardize the ncRNA database, NONCODE (http://www.noncode.org/) has been developed. The most recent edition of NONCODE (v1.0) includes non-redundant 5339 eukaryotic and archaebacterial sequences. Over the past few years, a substantial increase in the volume of data on ncRNAs has contributed to the NONCODE v.2.0 where the sum of obtained ncRNAs surpassed over 206,226 sequences from 861 distinct species [56]. In this edition, novel groups of ncRNAs, such as snRNA-like RNAs (snlRNAs), Piwi-interacting RNAs (piRNAs), and stem-bulge RNAs (sbRNAs) [57], have been included along with other unclassified ncRNAs. To date, NONCODE has entered version 3.0 and now includes over 42,000 public sequences from 125 species representing all the life types. One of the main attractions of this database is the provision of usable data for a particular lncRNA of interest.

By merging current databases like H-invDB, 5.0 [58], FANTOM3, miRBase [59], NONCODE [60], snoRNA (LBME db rel.3) [61], Rfam [53], RNAdb [62], and GEO [63], another Japanese community built a framework to mine/annotate usable RNA candidates from non-coding RNA sequences and named it fRNAdb [64]. fRNAdb is a database offering support for computational analysis of EST support assessment, RNA secondary structure motif discovery, cis-regulatory factor quest, and protein homology search. Besides, the fRNAdb database is connected to a customized "UCSC genome browser (RNA-specific custom tracks)." The upgrade fRNAdb 3.0 helps users to annotate anonymous RNA transcripts and recognize novel non-coding RNAs. In comparison, as in the NONCODE database, the fRNAdb database has increased the number of sequences to over three times the number of sequences previously available.

As previously mentioned, non-coding RNAs may be represented in a parent-dependent fashion and are referred to as "imprinted" ncRNAs. An increasing number of studies indicate that dysregulation of imprinted ncRNAs was related to several human diseases like Silver–Russell syndrome, Beckwith–Wiedemann syndrome, Prader–Willi syndrome, and multiple tumors [65–67]. Members of this group comprise microRNAs, antisense ncRNAs, small nucleolar RNAs, piRNAs, small interfering RNAs, mRNA-like ncRNAs, and piwi-interacting RNAs [68]. To examine the data in a more organized way, Zhang et al. cataloged all of the imprinted non-coding RNAs (ncRNAs) in a systematic database known as ncRNAimprint [69]. This database includes 7094 entries, of which 6612 are piRNA, accompanied by 187 microRNAs, 129 snoRNAs, 107 siRNAs, and 129 antisense ncRNAs [23]. Just 33 documents are actually accessible for mammalian organisms (http://

rnaqueen.sysu.edu.cn/ncRNAimprint). Another public repository of ncRNA expression data is the Non-coding RNA Expression Database which includes details of more than 5000 long ncRNAs found in humans and mice. NRED combines both evolutionary conservation and secondary structure proof to reflect lncRNAs, rendering it a valuable archive. Therefore, NRED provides important knowledge on characterizing lncRNAs and their transcriptome distributions.

The researchers also found that the NRED resource has been associated with another reference database for lncRNAs, namely lncRNAdb. This extensive database maintains a listing for all the lncRNAs with a regulatory role, as well as connected with some of the biological functions within eukaryotes. The LncRNADB provides details regarding the sequences, genetic background, subcellular localization, and expression. It is related to the UCSC genome viewer for visualization and expression information from a number of sources. The database includes over 150 separate lncRNAs from 60 distinct animals. 80% of them are human, and most of them are mammalian.

To reduce the task of determining the role of latent non-coding RNAs, a computational method was developed in which RNA sequencing data was combined with already existing annotation tools. The authors calculated lncRNAs based on gene similarity, structural similarity, transcriptional similarity, and orthology. They accumulated a huge number of human lincRNAs, with a substantial proportion of them not detected by GENCODE, RefSeq, or UCSC. Next, the authors created a database of 8195 human lincRNAs from integrating RNA-seq data from 24 tissues and cell types, using publicly accessible transcript annotations. They showed that the expression of lncRNAs is tissue-specific, owing to which adjacent genes are often expressed in the same tissue. The findings have been compiled into the Human Body Chart lincRNAs. This interconnected and detailed catalog could help to determine the global properties of lncRNAs enabling more studies on their role.

21.5 Conclusion and Future Directions

A large chunk of the human transcriptome is made up of lncRNAs and poses various molecular structures, functions, and mechanisms of action. Thus, there is a demand for the development of tools and techniques that can detect lncRNAs more precisely. Cantered on automated sequence annotation, experimental verification, disease association, and interaction studies, we also have a range of emerging reference datasets. Though to date, several experimental approaches have been developed, they demand huge capital and time. Hence, to overcome this, nowadays, several computational modeling approaches have been developed to answer unusual queries from transcriptome data. There are currently a variety of methods available for elucidating whether a transcript encodes or does not encode using possible standards for computer coding that include measuring multiple sequential characteristics. In general, non-coding RNA sequences work by folding into secondary structures, which are necessary for their function. Some approaches often influence these structures in the annotation process by calculating the minimum free energy and conservation of these characteristics. Completely these magnitudes contribute greatly to accurately assessing the functional capacity of these transcripts. The additional vital feature is to recognize the RNA mechanisms. These characteristics need to be included and tested comprehensively when designing the models. We may be unaware this RNA folds into 2^0 and 3^0 structures and dictates complex interactions with proteins and nucleic acid in different pipelines due to restricted structural features. To find the optimal folding stage of an RNA, a majority of the secondary structure methods used minimum free energy algorithms. At times, the same transcript could serve as RNA coding as well as non-coding. Therefore, to establish the conditions leading to gene expression, the co-factors influencing transcription must be further studied. A systematic integrative and metadata analysis is also required in order to understand diverse lncRNA regulatory mechanisms of action at various levels. It is also vital to remember that it is necessary to experimentally validate all these computational predictions, but they may minimize the search space for experimentation to a large extent. Finally, many novel lncRNAs are expected since most of the statistical methods predict lncRNAs with exactness. There are gaps among the non-coding transcript identification and their mechanisms. Further attempts would be made to annotate their features, which is beneficial in understanding the underlying cell biology.

Conflicts of Interest None

Other Information Figure 21.1 (CC BY 4.0) [25] has been reused under the terms of the Creative Commons Attribution License.

References

1. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, et al. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. Science. 2005;308(5725):1149–54.
2. ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, Dutta A, Guigó R, Gingeras TR, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799–816.
3. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, Hackermüller J, et al. Structured RNAs in the ENCODE selected regions of the human genome. Genome Res. 2007;17 (6):852–64.
4. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods [Internet]. [cited 2021 Feb 12]. https://www.nature.com/articles/nmeth.1226.
5. Nagalakshmi U, Waern K, Snyder M. RNA-Seq: a method for comprehensive transcriptome analysis. Curr Protoc Mol Biol. 2010;89(1):4.11.1–4.11.13.
6. Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. Genome Biol | Full Text [Internet]. [cited 2021 Feb 12]. https://genomebiology.biomedcentral.com/articles/10.1186/gb-2009-10-11-r124.
7. St Laurent G, Shtokalo D, Dong B, Tackett MR, Fan X, Lazorthes S, et al. VlincRNAs controlled by retroviral elements are a hallmark of pluripotency and cancer. Genome Biol. 2013;14(7):R73.
8. Wang KC, Chang HY. Molecular mechanisms of long noncoding RNAs. Mol Cell. 2011;43 (6):904–14.

9. Nissen P, Hansen J, Ban N, Moore PB, Steitz TA. The structural basis of ribosome activity in peptide bond synthesis. Science. 2000;289(5481):920–30.
10. Chen K, Rajewsky N. The evolution of gene regulation by transcription factors and microRNAs. Nat Rev Genet. 2007;8(2):93–103.
11. Pandey RR, Kanduri C. Transcriptional and posttranscriptional programming by long noncoding RNAs. Prog Mol Subcell Biol. 2011;51:1–27.
12. Heo JB, Sung S. Vernalization-mediated epigenetic silencing by a long intronic noncoding RNA. Science. 2011;331(6013):76–9.
13. Plessy C, Pascarella G, Bertin N, Akalin A, Carrieri C, Vassalli A, et al. Promoter architecture of mouse olfactory receptor genes. Genome Res. 2012;22(3):486–97.
14. MicroRNA signatures in human cancers. Nat Rev Cancer [Internet]. [cited 2021 Feb 12]. https://www.nature.com/articles/nrc1997.
15. Marchese FP, Raimondi I, Huarte M. The multidimensional mechanisms of long noncoding RNA function. Genome Biol. 2017;18(1):206.
16. Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat Rev Mol Cell Biol. 2016;17(12):756–70.
17. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Annu Rev Biochem. 2012;81:145–66.
18. Sunwoo H, Dinger ME, Wilusz JE, Amaral PP, Mattick JS, Spector DL. MEN ε/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. Genome Res. 2009;19(3):347–59.
19. Clemson CM, Hutchinson JN, Sara SA, Ensminger AW, Fox AH, Chess A, et al. An architectural role for a nuclear non-coding RNA: NEAT1 RNA is essential for the structure of Paraspeckles. Mol Cell. 2009;33(6):717–26.
20. Broadbent HM, Peden JF, Lorkowski S, Goel A, Ongen H, Green F, et al. Susceptibility to coronary artery disease and diabetes is encoded by distinct, tightly linked SNPs in the ANRIL locus on chromosome 9p. Hum Mol Genet. 2008;17(6):806–14.
21. Bhat SA, Ahmad SM, Mumtaz PT, Malik AA, Dar MA, Urwat U, et al. Long non-coding RNAs: mechanism of action and functional utility. Non Coding RNA Res. 2016;1(1):43–50.
22. Shi X, Sun M, Liu H, Yao Y, Song Y. Long non-coding RNAs: a new frontier in the study of human diseases. Cancer Lett. 2013;339(2):159–66.
23. Guttman M, Amit I, Garber M, French C, Lin MF, Feldser D, et al. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature. 2009;458 (7235):223–7.
24. Bhartiya D, Scaria V. Genomic variations in non-coding RNAs: structure, function and regulation. Genomics. 2016;107(2–3):59–68.
25. Zhang X, Wang W, Zhu W, Dong J, Cheng Y, Yin Z, et al. Mechanisms and functions of long non-coding RNAs at multiple regulatory levels. Int J Mol Sci. 2019;20(22):5573.
26. Supek F, Vlahovicek K. INCA: synonymous codon usage analysis and clustering by means of self-organizing map. Bioinformatics. 2004;20(14):2329–30.
27. Computational analysis of noncoding RNAs - Washietl - 2012 - WIREs RNA - Wiley Online Library [Internet]. [cited 2021 Feb 12]. https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1134.
28. Empirical codon substitution matrix. BMC Bioinformatics | Full Text [Internet]. [cited 2021 Feb 12]. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-6-134.
29. Stadler MB, Shomron N, Yeo GW, Schneider A, Xiao X, Burge CB. Inference of splicing regulatory activities by sequence neighborhood analysis. PLoS Genet. 2006;2(11):e191.
30. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(Web Server issue):W29–37.
31. Sonnhammer EL, Eddy SR, Durbin R. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins. 1997;28(3):405–20.
32. Wistrand M, Sonnhammer ELL. Improving profile HMM discrimination by adapting transition probabilities. J Mol Biol. 2004;338(4):847–54.

33. Jalali S, Bhartiya D, Lalwani MK, Sivasubbu S, Scaria V. Systematic transcriptome wide analysis of lncRNA-miRNA interactions. PLoS One. 2013;8(2):e53823.
34. Guttman M, Rinn JL. Modular regulatory principles of large non-coding RNAs. Nature. 2012;482(7385):339–46.
35. Reeder J, Steffen P, Giegerich R. pknotsRG: RNA pseudoknot folding including near-optimal structures and sliding windows. Nucleic Acids Res. 2007;35(Web Server issue):W320–4.
36. Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L, Vos J, et al. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. Genome Res. 2006;16(10):1289–98.
37. Machado-Lima A, del Portillo HA, Durham AM. Computational methods in noncoding RNA research. J Math Biol. 2008;56(1–2):15–49.
38. Li J-H, Liu S, Zheng L-L, Wu J, Sun W-J, Wang Z-L, et al. Discovery of Protein–lncRNA Interactions by Integrating Large-Scale CLIP-Seq and RNA-Seq Datasets. Front Bioeng Biotechnol [Internet]. 2015 [cited 2021 Feb 12];2. https://www.frontiersin.org/articles/10.3389/fbioe.2014.00088/full.
39. Wang T, Xiao G, Chu Y, Zhang MQ, Corey DR, Xie Y. Design and bioinformatics analysis of genome-wide CLIP experiments. Nucleic Acids Res. 2015;43(11):5263–74.
40. Fritah S, Niclou SP, Azuaje F. Databases for lncRNAs: a comparative evaluation of emerging tools. RNA. 2014;20(11):1655–65.
41. Li J-H, Liu S, Zhou H, Qu L-H, Yang J-H. starBase v2.0: decoding miRNA-ceRNA, miRNA-ncRNA and protein-RNA interaction networks from large-scale CLIP-Seq data. Nucleic Acids Res. 2014;42(Database issue):D92–7.
42. Gene Ontology-based function prediction of long non-coding RNAs using bi-random walk | SpringerLink [Internet]. [cited 2021 Feb 12]. https://link.springer.com/article/10.1186/s12920-018-0414-2.
43. Paraskevopoulou MD, Georgakilas G, Kostoulas N, Reczko M, Maragkakis M, Dalamagas TM, et al. DIANA-LncBase: experimentally verified and computationally predicted microRNA targets on long non-coding RNAs. Nucleic Acids Res. 2013;41(Database issue):D239–45.
44. Yang J-H, Li J-H, Jiang S, Zhou H, Qu L-H. ChIPBase: a database for decoding the transcriptional regulation of long non-coding RNA and microRNA genes from ChIP-Seq data. Nucleic Acids Res. 2013;41(Database issue):D177–87.
45. Chen X. Predicting lncRNA-disease associations and constructing lncRNA functional similarity network based on the information of miRNA. Sci Rep [Internet]. 2015 Aug 17 [cited 2021 Feb 12];5. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4538606/.
46. Jeggari A, Marks DS, Larsson E. miRcode: a map of putative microRNA target sites in the long non-coding transcriptome. Bioinformatics. 2012;28(15):2062–3.
47. miREE: miRNA recognition elements ensemble. BMC Bioinformatics | Full Text [Internet]. [cited 2021 Feb 12]. https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-454.
48. Scruggs BS, Michel CI, Ory DS, Schaffer JE. SmD3 regulates intronic noncoding RNA biogenesis. Mol Cell Biol. 2012;32(20):4092–103.
49. Zinkin NT, Grall F, Bhaskar K, Otu HH, Spentzos D, Kalmowitz B, et al. Serum proteomics and biomarkers in hepatocellular carcinoma and chronic liver disease. Clin Cancer Res. 2008;14(2):470–7.
50. Weiland M, Gao X-H, Zhou L, Mi Q-S. Small RNAs have a large impact: circulating microRNAs as biomarkers for human diseases. RNA Biol. 2012;9(6):850–9.
51. Song X, Cao G, Jing L, Lin S, Wang X, Zhang J, et al. Analysing the relationship between lncRNA and protein-coding gene and the role of lncRNA as ceRNA in pulmonary fibrosis. J Cell Mol Med. 2014;18(6):991–1003.
52. Szymański M, Erdmann VA, Barciszewski J. Noncoding regulatory RNAs database. Nucleic Acids Res. 2003;31(1):429–31.
53. Griffiths-Jones S, Bateman A, Marshall M, Khanna A, Eddy SR. Rfam: an RNA family database. Nucleic Acids Res. 2003;31(1):439–41.

54. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, Lindgreen S, et al. Rfam: updates to the RNA families database. Nucleic Acids Res. 2009;37(Database issue):D136–40.
55. Integrative annotation of 21,037 human genes validated by full-length cDNA Clones [Internet]. [cited 2021 Feb 12]. https://journals.plos.org/plosbiology/article?id=10.1371/journal.pbio.0020162.
56. He S, Liu C, Skogerbø G, Zhao H, Wang J, Liu T, et al. NONCODE v2.0: decoding the non-coding. Nucleic Acids Res. 2008;36(Database issue):D170–2.
57. Deng W, Zhu X, Skogerbø G, Zhao Y, Fu Z, Wang Y, et al. Organization of the Caenorhabditis elegans small non-coding transcriptome: genomic features, biogenesis, and expression. Genome Res. 2006;16(1):20–9.
58. Yamasaki C, Koyanagi KO, Fujii Y, Itoh T, Barrero R, Tamura T, et al. Investigation of protein functions through data-mining on integrated human transcriptome database, H-invitational database (H-InvDB). Gene. 2005;364(1–2):99–107.
59. Griffiths-Jones S, Grocock R, Dongen S, Bateman A, Enright A, Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. Nucleic Acids Res. 2006;34(Database issue):D140–4.
60. Liu C, Bai B, Skogerbø G, Cai L, Deng W, Zhang Y, et al. NONCODE: an integrated knowledge database of non-coding RNAs. Nucleic Acids Res. 2005;33(Suppl_1):D112–5.
61. Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res. 2006;34(Database issue):D158–62.
62. Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, Mattick JS. RNAdb 2.0—an expanded database of mammalian non-coding RNAs. Nucleic Acids Res. 2007;35(Database issue):D178–82.
63. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, et al. NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res. 2007;35(Database issue):D760–5.
64. Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, et al. fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. Nucleic Acids Res. 2006;35(Database issue):D145–8.
65. Sahoo T, del Gaudio D, German JR, Shinawi M, Peters SU, Person RE, et al. Prader-Willi phenotype caused by paternal deficiency for the HBII-85 C/D box small nucleolar RNA cluster. Nat Genet. 2008;40(6):719–21.
66. Bliek J, Terhal P, van den Bogaard M-J, Maas S, Hamel B, Salieb-Beugelaar G, et al. Hypomethylation of the H19 gene causes not only Silver-Russell syndrome (SRS) but also isolated asymmetry or an SRS-like phenotype. Am J Hum Genet. 2006;78(4):604–14.
67. Zhang X, Zhou Y, Mehta KR, Danila DC, Scolavino S, Johnson SR, et al. A pituitary-derived MEG3 isoform functions as a growth suppressor in tumor cells. J Clin Endocrinol Metab. 2003;88(11):5119–26.
68. Koerner MV, Pauler FM, Huang R, Barlow DP. The function of non-coding RNAs in genomic imprinting. Development. 2009;136(11):1771–83.
69. Zhang Y, Guan D-G, Yang J-H, Shao P, Zhou H. Qu L-H. ncRNAimprint: a comprehensive database of mammalian imprinted noncoding RNAs. RNA. 2010;16(10):1889–901.

Circular RNA in Rice (Oryza sativa)

22

Maryam Moazzam-Jazi, Vahideh Hedayati, and Sohrab Moradi

Abstract

Circular RNAs are a group of non-coding RNAs with a closed-loop structure that produce via the atypical alternative splicing event in all eukaryotic species. As a result of current advances in sequencing technologies and data analysis tools, an enormous number of circRNAs have been recognized in different animal and plant species. In contrary to the extensive researches of circRNAs in animals, circRNA investigation in plants is in its early phase. However, the genome-wide prediction of circRNAs in plants is rapidly emerging, indicating the important biological roles of circRNAs in plant growth, development, and stress responses. In this chapter, we describe the circRNA biosynthesis and its features in plants and compare it with animals when required. We also provide the bioinformatics resources for plant circRNA discovery and review the putative biological functions of circRNA in plants.

Keywords

Circular RNA · circRNA · Plants · Rice

M. Moazzam-Jazi (✉)
Cellular and Molecular Endocrine Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran

V. Hedayati
Dana Gene Pajouh Company, Tehran, Iran

S. Moradi
Biotechnology Department, National Institute of Genetic Engineering and Biotechnology, Tehran, Iran

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_22

Abbreviation

ceRNA	Competing endogenous RNAs
CircRNA	Circular RNA
DEC	Differentially expressed circular RNAs
lncRNA	Long noncoding RNAs
IRES	Internal ribosomal entry site
LLERCPs	Reverse complementary pairs
LLEs	LINE1-like elements
miRNAs	MicroRNAs
ORF	Open reading frames
Psa	Pseudomonas syringae pv. Actinidiae
PSY1	Phytoene synthase 1
PTGMS	Photo-thermosensitive genic male sterile
RBPs	RNA-binding proteins
RNAase	Ribonuclease
RNA-seq	RNA-sequencing
snRNP	Small nuclear ribonucleoprotein
SEP3	SEPALLATA3

22.1 Introduction

A group of single-stranded non-coding RNAs with a closed-loop structure named circular RNAs (CircRNAs) are produced via the back-splicing reaction. In this non-canonical alternative splicing event, 5′ and 3′ ends of pre-mRNA transcripts are connected by a covalent bond [1, 2]. Unlike linear RNAs, the closed-loop structure of circRNAs with neither 5′ cap nor 3′ poly-A tail provided greater stability and resistance to degradation by a variety of RNAase. Hence, eukaryotic total RNA digestion using RNAase can be beneficial for circRNAs segregation and enrichment from other RNA types [3, 4]. In 1976, Sanger et al. discovered the first circRNAs, viroid's RNA that propagated in tomato and Gynura [5]. As a result of emerging modern deep sequencing technologies along with the improved circRNA enrichment approaches and data analysis tools, thousands of circRNAs were discovered in many organisms, including archaea, humans, mouse, Arabidopsis, rice, and zebrafish. Consequently, circRNAs are commonly found in all eukaryotes. According to CIRCpedia, 183,000 circRNAs have been recognized in different human samples as of July 2018. Similarly, various research teams have reported over 95,000 circRNAs in different plant species through June of this year (2018), in which the highest number of identified circRNAs belonged to *Oryza sativa* and *Arabidopsis thaliana* [6]. After miRNAs and lncRNA, a trending research topic in molecular biology can be assigned to circRNAs. In contrast to the thorough and extensive circRNAs characterization in animals, little attention has been paid to the plant circRNA discovery. Therefore, our knowledge of circRNAs was mostly originated from researches in animals and humans. Based on these studies, numerous roles are proposed for circRNAs, including acting as miRNA sponge, the vehicle for RNA-binding proteins, and the regulator of gene expression and translation. Although circRNAs are mainly categorized as non-coding RNAs, some of them appear to serve as the template for protein synthesis as demonstrated by pieces of evidence obtained from humans and animals. For example, a circRNA derived from beta-catenin can translate to protein and promote liver cancer cell growth [7]. Nevertheless, the plant circRNAs might produce and regulate through different mechanisms and have distinct functional roles as compared with animal circRNAs. In this chapter, we have presented in brief the current studies in characterizing plant circRNAs, with more emphasis on rice (*Oryza Sativa*). The biogenesis and discovery of circRNAs as well as their characteristics in plants are discussed. Next, we offered the available computational pipelines and circRNA databases that can be applied to identify circRNAs across RNA sequencing reads. Finally, the putative functions of plant circRNAs are surveyed.

22.2 Biogenesis and Discovering of Plant circRNAs

Diverse mechanisms are contributed to form the circRNAs, however, the most well-known mechanism is the alternative back-splicing phenomenon, which occurred via covalently joining the upstream 3′ splice site (acceptor) to a downstream 5′ splice site (donor) of pre-mRNA transcripts. The circRNA molecules possess a distinctive exon–exon junction that is not observed in the linear transcript. Remember that the 5′ splice site of an upstream exon is joined to the 3′ splice site of a downstream exon during the normal alternative splicing event. *Occurring* RNA circulation at the splicing sites implies that the back-splicing process requires the spliceosome machinery, which is typically catalyzed by the linear pre-mRNA splicing [8]. As depicted in Fig. 22.1, five different pathways are involved in the alternative back-splicing event. First, lariat-driven circularization, in which normal alternative splicing event can create circRNAs. The exon-skipping event and intronic sequence removal in pre-mRNA transcripts can form the exon-harboring lariat and usual intronic lariat, respectively, which might result in circRNA molecules generation (Fig. 22.1a). Second, RBP-driven circularization in which RBPs are attached to the specific motif in the flanking introns of the circularized exons, which bring the splice sites closer and call the spliceosome in a back-splicing site to form the exonic circRNAs (ecircRNA) (Fig. 22.1b). Therefore, a set of proteins catalyzed the circRNA production via modulating the availability of back-splicing signals for the spliceosome. Third, intron pairing-driven circularization, in which base-pairing happens among the reverse complementary and repetitive bases at the intronic regions of the back-spliced exons. The ecircRNA or exonic-intronic circRNAs (EIciRNAs) will be generated in the case of removing or retaining the introns, respectively (Fig. 22.1c). The circularization efficiency might be influenced by the secondary RNA structure, the existence of mispairing in intronic repetitive elements,

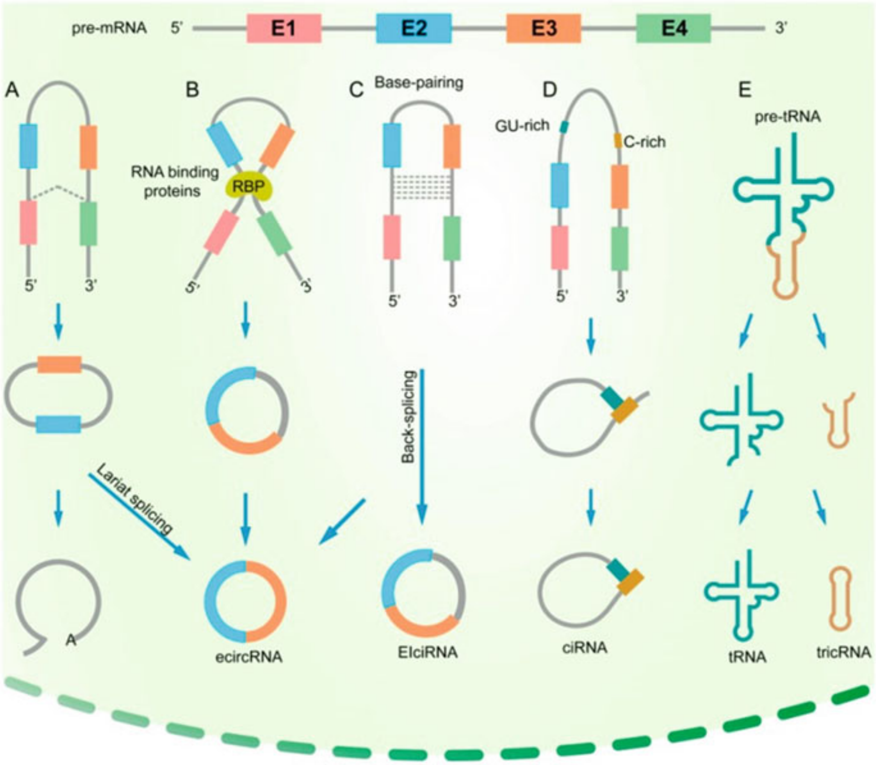


Fig. 22.1 Schematic illustration of five pathways involved in circular RNA biosynthesis. (Adapted from [9])

and low complexity regions. In the fourth, architecture of the ciRNA relies on a 7-nt guanine-rich element and an 11-nt cytosine-rich element to prevent cleavage and exonucleolytic degradation (Fig. 22.1d). Fifth is the tricRNA formation, in which tRNA splicing enzymes divide the pre-tRNA into two parts: tricRNAs are generated by a 3′–5′ phosphodiester bond and the other part generates tRNAs (Fig. 22.1e) [9–12].

CircRNAs can be classified into ten categories according to the genomic origin of back-splicing signals location. (1) exonic and (2) intronic circRNAs that are created in the case of locating both acceptor and donor back-splice sites within single exon and intron, respectively; while (3) exonic-intronic circRNAs generated if one back-splice site exists at exon and another one located at intron. (4) intronic-exonic circRNAs in which two back-splice sites are located at two diverse introns across single or multiple exons. (5) While UTR circRNAs have resulted from two back-splice sites positioned at the UTR region, (6) UTR-exonic and (7) UTR-intronic circRNAs are produced if one back-splice site is situated at the UTR region and another one located at exonic or intronic sequences, respectively. Similarly,

(8) intergenic circRNAs, which are produced when two back splice sites are located at a specified intergenic region, whereas (9) intergenic-genic circRNAs are formed in the case of positioning one back splice site at the intergenic region and another one at the genic region. Finally, (10) across-genic circRNAs are derived from two back splice sites positioned at two distinct genes. All types of circRNAs, except for intergenic-derived circRNAs, completely or partially cover the genic region of the genome. Most discovered circRNAs in plants have originated from annotated genes, both exonic and intronic parts, proposing a direct or indirect role in gene expression regulation. It has been reported that 86% and 92% of circRNAs originated from protein-coding genes in *Arabidopsis thaliana* and *Oryza sativa*, respectively [6].

Owing to the lack of poly $(A)^+$ tail and non-linear structure, the circRNAs are hardly captured by the RNA-seq technology that the corresponding libraries are normally enriched for poly-adenylated RNAs. Circular RNAs can be detected using rRNA-depleted RNA-seq libraries or the library enrichment using RNase R, an enzyme that especially breaks the linear RNA transcripts [13, 14]. CircRNA recognition requires adequate sequencing coverage, ideally hundreds of millions of reads, even for the circRNA-enriched libraries. Ye et al. utilized the available rRNA-depleted RNA-seq datasets for discovering circRNAs in Arabidopsis and rice; this study was the first report on the plant circRNAs detection at the genome level that led to the identification of 6012 circRNAs from the leaves of *Arabidopsis thaliana* and 12,037 circRNAs from the roots and shoots of *Oryza sativa* [15]. Given the circular conformation and a rather low expression of circRNAs, the treatment of RNA-seq library with RNase R for the enrichment of circRNAs is a widely used and preferred method, leading to characterize the high confidence circRNAs in various plants, including grape, tomato, bamboo, and soybean [14, 16–19]. Nevertheless, some circRNAs have been successfully recognized in the polyadenylated-enriched libraries. Lu et al. separately analyzed the circRNAs in poly (A)-selected and poly (A)-depleted samples obtained from leaf and panicle tissues of *Oryza Sativa* ssp. Japonica Nipponbare [20]. They reported that the number of leaf and panicle circRNAs identified in poly (A)-selected libraries were marginally more than poly (A)-depleted libraries, which might be referred to as the more high-quality transcriptomic data derived from the poly (A)-selected libraries. However, the poly (A)-depleted libraries display much higher detection efficiency of circRNA compared to poly (A)-selected samples [20]. The up-to-date list of identified circRNAs in rice has been demonstrated in Table 22.1.

22.3 Characteristics of Plant Circular RNAs

In addition to the nuclear genome, circRNAs are derived from the mitochondrial and chloroplast genomes, implying their possible role in the photosynthesis and respiration processes [23]. The length of circRNAs is usually <1 kbp; however, their length distribution ranges from shorter than 200 bp to longer than 100 kbp. The size of plant circRNAs is mostly between 200 and 600 bp, only a few of them are longer than 2 kbp. The possibility of using different splice donor and acceptor sites during back

Table 22.1 List of the genome-wide identification of circRNAs in *Oryza sativa*

Tissues/ developmental stages	CircRNA count	Approach	Biological process/stress	References
Roots and shoots at the flowering time	12,037	rRNA-depleted RNA-Seq	Pi-starvation stress	[15]
Panicles and leaves at the flowering time	2354	rRNA-depleted/poly (A)$^+$-selected RNA-Seq	Normal conditions	[20]
Roots of seedlings	2806	rRNA-depleted/RNase R-treated RNA-Seq	Normal conditions	[21]
Young panicles at the flowering time	9994	rRNA-depleted RNA-seq	Fertility transition	[18]
Young leaves	2932	rRNA-depleted RNA-seq	Magnaporthe oryzae inoculation	[22]

splicing event gives rise to the generation of multiple circRNAs from a single gene locus; only a few of the circRNAs are produced from various loci of the same gene. For instance, it was predicted that the parental gene, LOC_Os12g02040, can create 38 circRNA isoforms in rice, in which the full-length sequence of eight isoforms was also experimentally validated [24]. Similar to animals, the abundance of most plant circRNAs is relatively low in different plant species and represents the stress- and developmental-specific expression pattern [25]. Lu et al. observed that from 30 experimentally verified circRNAs in rice, four and three circRNAs exhibited the leaf- and panicle-specific expression pattern in this plant [20]. In comparison to the linear RNAs, circRNAs exhibit a lower expression that appears to correlate with the expression of their parental genes. For instance, there is a negative correlation between the expression of Os08circ16564 and the expression of its derived gene, AK064900, in rice [20]. Similarly, the positive correlation between the transcript level of Ac_ciRNA_04842 and its parental gene, Achn372061, was discovered in kiwifruit [26]. Therefore, the expression level of circRNAs may modulate by the interior regulatory mechanisms that have not been completely clarified, yet. Furthermore, circRNAs are conserved among various plant species, which may refer to their important biological functions in plants. The researchers reported that plenty of circRNAs originated from more than 700 orthologous genes between Arabidopsis and rice. However, the bracketing introns of the preserved circRNAs show no sequence similarity or common motifs, indicating that other conserved mechanisms might be contributed to the biosynthesis of plant circRNAs [15]. In both animals and plants, the transcription and back splicing reactions of pre-mRNAs are mediated by RNA polymerase II. In animals, there are substantial reverse complementary sequences in the flanking introns of circularized exons that are indispensable for circRNA generation in these organisms. For example, long introns containing Arthrobacter luteus (Alu) elements enclosed circularized exons in humans [27, 28]. On the contrary, the repetitious regions or reverse complementary sequences appeared not to be enriched in the bracketing sequences of exonic circRNAs in plants, proposing that intron pairing-driven circularization might not be the main pathway for plant circRNA biosynthesis. For example, in rice, Arabidopsis, and soybean, the percentage of reverse complementary sequences in the intronic regions surrounding exonic circRNAs was only 6.2%, 0.3%, and 2.7%, respectively [29]. However, a circRNA-Seq study on maize revealed that LLEs and their LLERCPs are considerably increased in the enclosing sequences of circRNAs, demonstrating that transposons may be contributing to form the plant circRNAs [19]. The major spliceosome, U2-dependent spliceosome, has catalyzed the removal of intronic sequences with splicing signals of GT and AG at the 5′ and 3′ splice sites, respectively, in both animals and plants. The recognized circRNAs in different organisms are mainly surrounded by the standard splice sites, a GT at the 5′ donor site, and an AG at the 3′ acceptor site [15, 30]. However, the splice signals of circRNAs differ among plant species. It has been experimentally proved that the back-splicing sites of many circRNAs (92.7%) identified in the root tissue of *Oryza sativa* are flanked by different non-GT/AG splicing signals, such as GC/CG, CT/GC, and GC/GT [24]. The non-canonical splicing signals surrounded the circRNA were also specified in *Cucumis sativus* [31] and chloroplast of *Arabidopsis thaliana* [32]. In contrast, based on the studies conducted on Arabidopsis, grape, and cotton, most of the characterized circRNAs in these plants are spliced by the common GT/AG splicing signals [16, 33, 34]. The investigation of circRNAs in 12 plant species, including *Arabidopsis thaliana, Gossypium arboretum, Gossypium hirsutum, Glycine max, Gossypium raimondii, Hordeum vulgare, Oryza sativa, Poncirus trifoliata, Solanum lycopersicum, Solanum tuberosum*, *Triticum aestivum*, and *Zea mays,* demonstrated that the majority of circRNAs contain the standard splicing signals (GT/AG or CT/AC); however, the percentage of circRNAs with atypical splicing signals is different among various species. The distribution of different junction sites of circRNAs in *O. sativa* and *A. thaliana* were detected by Chu et al. [6] which, only 6% and 9% of all discovered circRNAs have the canonical splicing sites in rice and Arabidopsis, respectively. The discrepant splicing signals across different species might result from diverse RNA-Seq approaches, circRNA prediction tools, and filtering criteria used in the identification of circRNAs. All of these could greatly influence the number and characteristics of the recognized circRNA [6].

22.4 Bioinformatics Resources for Investigating circRNAs in Plants

As millions of circRNA sequencing reads have been produced using high-throughput sequencing technologies, the development of efficient computational pipelines is essential to handle the growing RNA-seq datasets and profound perception of circRNA characteristics. The core approach for predicting circRNA from rRNA-depleted RNA-Seq libraries (Ribominus-Seq) is based on the existence of back-splice spanning reads. Note that the genome duplication, tandem repeat of the genome, trans-splicing, and transcription template switching can also generate the

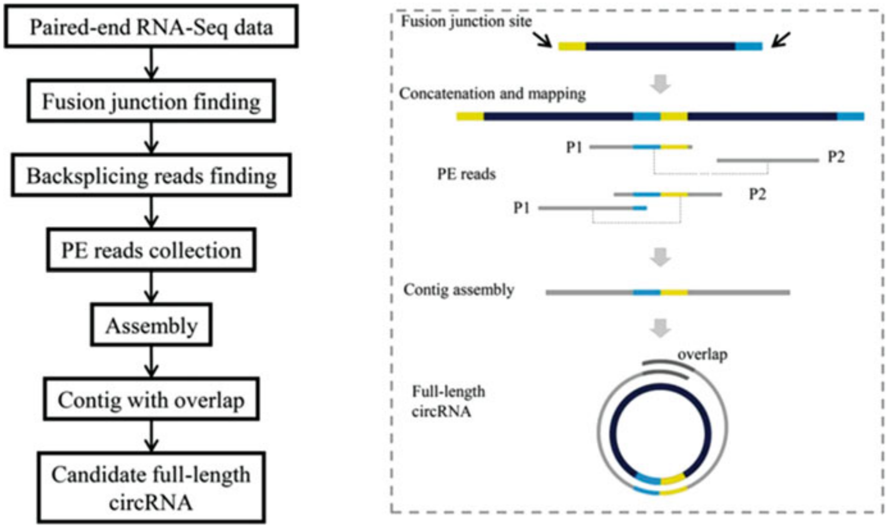


Fig. 22.2 Identification of full-length sequences of circRNAs in rice. The computational pipeline of circseq_cup for recognizing the full-length circRNA sequences using paired-end RNA-Seq data. (Adapted from [24] with permission)

back–splice junctions. Hence, using accurate methods is essential to avoid the discovery of false-positive circRNA molecules [35]. Several tools, including circRNA finder [36], CIRCexplorer [37], CIRI [38], find circ [39], Mapsplice [40], and PcircRNA_finder [41], have been developed for detecting circular RNAs. Among them, PcircRNA_finder was specially designed for detecting circRNA in plants. Furthermore, the circseq_cup pipeline was developed for detecting the full-length circRNAs in rice [24]. At the first step of this pipeline, the paired-end RNA-seq reads are aligned to a reference genome by publicly available tools, like STAR-Fusion, TopHat-Fusion, and segemehl, to identify possible fusion splice junction sites; distance between two fusion junction sites was considered to be $\leq$5 kb for rice. At the second step and after specifying the possible fusion splice junction sites, back-splicing sites, where the corresponding circRNA reads can be contiguously mapped, are determined. The back-splicing sites are corroborated by aligning the unmapped reads obtained from the alignment in the first step. At the third step, back-splicing paired-end reads were collected for contig assembly using Cap3 software. An assembled contig with $\geq$5 nucleotides overlap at its two ends was referred to the potential full-length circRNA sequence (Fig. 22.2). A single mismatch is acceptable for an overlap region equal to or more than ten nucleotides. Additionally, in the case of a short contig length (<100 bp), the overlap length should be $\geq$20 nucleotides [24].

Very recently, the CircPlant tool was also developed for identifying the plant circular RNAs using multiple plant-related criteria. CircPlant is composed of four modules: (1) circRNA identification, (2) circRNA–mRNA interplay detection,

(3) circRNA–miRNA–mRNA network drawing, and (4) circRNA annotation. The authors of the tool conducted several comparison tests on simulated and real RNA-seq data of Arabidopsis and rice and claimed that CircPlant is more accurate and efficient than all evaluated programs [42]. For further information on circRNA computational pipelines and analysis, we recommend readers to refer to two comprehensive reviews [43, 44]. Generally, current algorithms applied by the circRNA prediction programs can be categorized into two main groups, pseudo-reference- and fragmented-based strategies. In the pseudo-reference-based strategy, all possible back-spliced junctions are constructed and used as a reference for mapping RNA sequencing reads; those reads that aligned with the predicted back-spliced junctions are retained. Considering the requirement of gene annotation for executing this strategy, it cannot be used for circRNA discovery in the poorly annotated or non-model genomes. However, KNIFE, a pseudo-reference-based tool for detecting circRNAs, contains a de novo analysis module for discovering circRNAs produced from back-splicing sites with no previous annotations [45]. In the fragmented-based approach, RNA-sequencing reads are fragmented into small segments with about 25 nucleotides and aligned to a reference genome; segments that are aligned at back-splicing sites are reserved [12]. Different computational pipelines apply various methods, like the aligner and the usage of chimeric/fusion reads to predict circRNAs. These pipelines recognize the back-splicing sites within unmapped reads or sequencing reads that do not map end-to-end using the information of mapped/unmapped reads within the sequence alignment mapped (SAM) file. Several alignment tools, such as Bowtie, BWA, Star, Kallisto, and TopHat are widely used for mapping reads with reference [35]. According to the comparison of various circRNA detection tools by Hansen et al., different tools can create highly diverse results with high false-positive fraction; but, merging the output of the several programs can produce a much more reliable output [46]. Similarly, another researcher group comprehensively evaluated different circRNA prediction tools in terms of accuracy, sensitivity, F1 score and area under the curve, RAM usage, running time, and the required physical disk space; they found that CIRI, CIRCexplorer, and KNIFE outperform according to the establishment of a balance between accuracy and sensitivity [47].

Multiple circRNA databases have been designed for the effective management and organization of a growing number of circRNAs that are continuously being characterized. In comparison with animals, there are only two comprehensive plant circRNA databases, PlantcircBase [6] and PlantCircNet [48] that comprise the circRNA data of various plant species, data browsing, potential mRNA–miRNA–circRNA interplay networks, and BLAST. Moreover, while the tissue-specific circRNAs in Arabidopsis have been collected in the AtCircDB database [49], the data of expressed circRNA transcripts in response to abiotic stresses in maize and rice are accessible in the CropCircDB database [50]. The information of experimentally confirmed circRNA in both animals and plants, including Arabidopsis, rice, wheat, tomato, cotton, and kiwifruit along with their computationally predicted functions were collected at CircFunBase database; the circRNA–miRNA interaction networks can be also visualized at this database [51]. Recently, the GreenCircRNA database was developed that provides the information of 213,494 circRNAs from

69 plant species, in which the related miRNA information for 38 plants are available at the miRNA database (miRBase). Hence, the circRNAs that act as miRNA sponge were identified in these 38 plants, including rice [52]. This resource can greatly simplify the further analysis of circRNAs that operate as competitive endogenous RNAs (ceRNA). Likewise, the miRNA-binding sites localizing at the alternative splicing regions of linear and circular RNA transcripts at 11 plant species are available at the ASmiR database [53]. A summary of bioinformatics resources for plant circRNAs is presented in Table 22.2.

22.5 Plant circRNAs Functions

The expression of most circRNAs is confined to the specific cells, tissues, developmental phases, and adverse conditions that are extremely controlled, indicating their programmed generation by the eukaryotic cells under various conditions [55]. The putative functions of plant circRNAs are miRNA trapping (sponging), transcription regulation, translation into peptides/proteins, and modifying protein functions via direct interaction with them [56]. All of these activities are critical for plant development and growth, as well as appropriate reactions to diverse stressful situations. Here, we explain the plant circRNAs functions in detail.

22.5.1 CircRNAs as miRNA Sponges

According to prior investigations in animals, the most considerable function of circRNAs is to operate as miRNA trapping (sponging) or regulate gene expression in a miRNA-dependent manner (Fig. 22.3IV). The miRNA sponges or ceRNAs are the transcripts with several miRNA-binding positions that can impede the miRNA function [28]. Some circRNAs can attach to the RBP through their specific binding sites, proposing that they can also act as RBP sponges [55]. Moreover, several researchers have reported that circRNAs can be potential miRNA sponges in plants; however, the number of plant circRNAs functioning as miRNA sponges as well as their miRNA-binding sites is fewer than animal circRNAs. Ye et al. detected the potential miRNA-binding sites in only 5% and 6.6% of circRNAs in Arabidopsis and rice, respectively [15]. Based on diverse studies in different plant species, multiple circRNAs were identified as miRNA sponges in several plants including pepper, Chinese cabbage, cucumber, and sea buckthorn. Similarly, five circRNAs originated from *Arabidopsis thaliana* might operate as miRNA sponges, which only one of them has been experimentally confirmed [28]. Investigation of the ceRNA network and differential expression of mRNA, circRNA, and miRNA transcripts in Arabidopsis leaves offered that circRNAs can play important regulatory roles during leaf senescence [51]. A group of researchers have revealed that out of 235 exonic circRNAs with potential miRNA-binding sites in *Oryza sativa*, just 31 circRNA harbors two or more miRNA-binding sites, which is significantly less than that of human circRNAs. According to this study, the miRNA-binding sites are not

Table 22.2 A summary of bioinformatics resources for detecting circRNAs in plants

Name	Description	Weblinks	Latest release	References
pcircRNA_finder	A pipeline for identifying plant circRNAs	http://ibi.zju.edu.cn/bioinplant/tools/manual.htm	2017	[41]
Circseq-cup	A pipeline to predict the circRNAs and assembly as the full-length sequence	https://github.com/bioinplant/circseq-cup	2016	[24]
CircPlant	An integrated tool for circRNA detection and functional prediction in plants	http://bis.zju.edu.cn/circplant/	2019	[42]
CircCode	A tool for recognizing the coding ability of circRNAs	https://github.com/PSSUN/CircCode	2020	[54]
AtCircDB	A tissue-specific database of circRNAs for Arabidopsis	http://genome.sdau.edu.cn/circRNA	2018	[49]
PlantCircNet	A database of plant circRNA–miRNA–gene regulatory networks	http://bis.zju.edu.cn/plantcircnet/	2018	[48]
ASmiR	A comprehensive database of miRNA targets in alternatively spliced linear and circRNAs of plants	http://forestry.fafu.edu.cn/bioinfor/db/ASmiR	2019	[53]
CropCircDB	A database of plant circRNAs in response to abiotic stresses.	http://deepbiology.cn/crop/	2019	[50]
PlantcircBase	A comprehensive database of plant circRNAs	http://ibi.zju.edu.cn/plantcircbase/index.php	2020	[6]
CircFunBase	A database for functional circular RNAs	http://bis.zju.edu.cn/CircFunBase/index.php	2019	[51]
GreenCircRNA	A comprehensive database of plant circRNAs	http://greencirc.cn/	2019	[52]

considerably elevated among rice circRNAs, inferring that miRNA sponging is not probably the major role of circRNAs in this plant [20]. Another study in the PTGMS rice line demonstrated that 17% of all identified circRNAs have the miRNA-binding sites, in which the count of binding positions is highly variable. While 56% and 15% of circRNAs possessed the miRNA-binding positions for one miRNA and two miRNAs, respectively, 29% of circRNAs bear the putative binding positions for more than two miRNAs. Moreover, this study has also reported that about 23% of

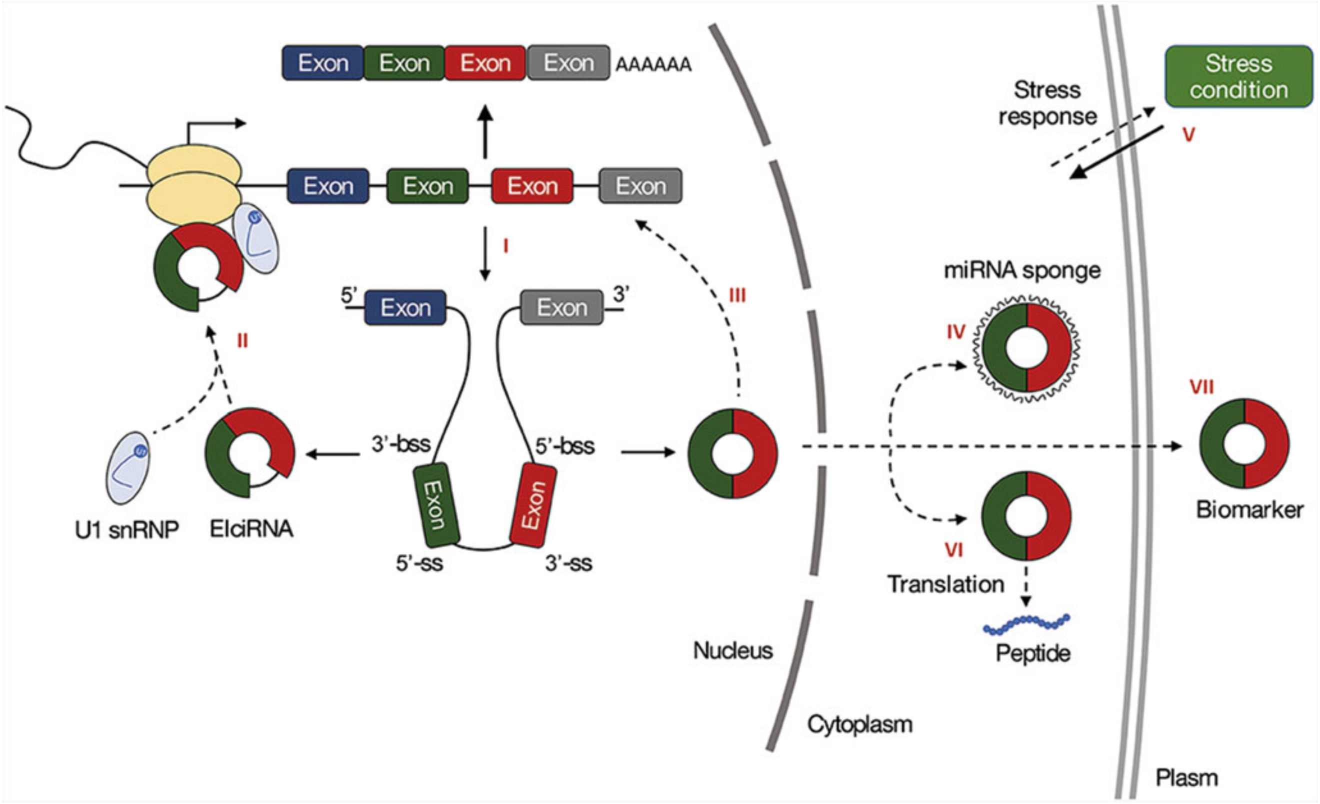


Fig. 22.3 CircRNA functions in plants. (I) CircRNAs processing can affect the splicing of their linear counterparts. (II) CircRNAs can regulate the transcription of their parental genes. (III) CircRNAs can regulate the splicing of their linear cognates. (IV) CircRNAs can act as miRNA sponges. (V) CircRNAs can regulate gene expression in response to biotic or abiotic stresses. (VI) CircRNAs can be translated. (VII) CircRNAs are promising biomarkers. (Adapted from [28])

differentially expressed circRNAs with miRNA-binding sites in rice can act as miRNA sponges, proposing that the interaction of miRNA and circRNA can lead to the circRNAs act as miRNA sponges as well as the expression level of circRNAs may be modulated via miRNA-mediated target circRNA cleavage [57]. Note that a part of the miRNA-binding site can be hidden by the secondary structure of circRNAs, which is mostly neglected during the identification of the possible miRNA-binding sites in plant circRNAs [25]. It has been reported that the overexpression of Os08circ16564 in rice, a circRNA imagined as a target imitator of OsmiR172, can decrease the transcript level of its derived gene with no modification on the transcript level of OsmiR172, representing that the Os08circ16564 cannot play the miR172 sponge role in vivo. It may be assumed that the association of OsmiR172 with the circRNA is blocked through the preservation of miR172-binding position in the stem-loop zone of Os08circ16564 [20, 25]. Generally, studies in both plants and animals propose that operating as a miRNA sponge is not the key function of circRNAs. Instead, it has been speculated that the plant circRNA degradation via the miRNA-mediated cleavage pathway can regulate the expression level of circRNAs [25]. Therefore, circRNAs may play regulatory roles in different processes by interacting with miRNA, which should be further explored in future researches.

22.5.2 CircRNAs in Stress Response

Prior investigations have proven that the expression of plant circRNAs is induced under various adverse environmental conditions, comprising high temperature, salt, chilling, dehydration, nutrient shortage, and invasion of a pathogen (Fig. 22.3V) [28].

Several sophisticated biological pathways have been evolved in plants to deal with the environmental challenges during their growth and development cycle, in which enormous gene expression reprogramming has a central role. Non-coding RNA transcripts, like miRNAs and lncRNAs, have been revealed to play central functions in plant gene expression modulation in response to different stresses. The circRNAs may also operate like other ncRNAs. It has been reported that the interaction between DEC and miRNAs can regulate the expression of stress-responsive genes and facilitate the survival of plants under damaging conditions [28, 55]. The induced circRNAs under biotic stresses were firstly identified during the bacterial (*Pseudomonas syringae*) invasion to Arabidopsis leaves; the biological roles of these circRNAs remain to be disclosed [33]. The differentially expressed circRNAs were also recognized during a pathogen attack in kiwifruit. In total, 584 DECs have been detected in response to the *Pseudomonas syringae pv. actinidiae* (Psa) infection that their transcripts level depends on the infection phase [26]. Later studies revealed the responsive circRNAs to the MIMV (maize Iranian mosaic virus) infection in maize [58], the Verticillium in cotton [59], and the TYLCV (tomato yellow leaf curl virus) in tomato [60] that can act as the negative regulators of virus interaction and tomato. Recently, Feng et al. elucidated that circRNAs are contributing to the interaction of rice with rice blast fungus (*Magnaporthe oryzae*) by circRNA-sequencing and transgenic methods. They discovered 636 circRNAs that were specially produced in response to *M. oryzae* contamination [22]. This study uncovers a new level of gene modulation and immunity in rice under the fungal infection. Besides, many differentially expressed circRNAs were deciphered in plants under abiotic stresses. Ye et al. characterized the stress-specific expression of 27 circRNAs in rice roots under phosphate-starvation conditions, of which 21 were down-regulated and 6 were up-regulated [15]. Also, 163 chilling-responsive circRNAs were identified in tomato fruit [14] and 62 circRNAs related to the photosynthesis and hormone signal pathways were detected to be differentially expressed under drought stress in wheat [61]. CircRNA data analysis in cucumber uncovered that similar to rice, non-GT/AG splicing sites are popular among identified circRNAs in this plant, and pairing-driven circularization is not the key pathway for circRNAs biogenesis. In this research, numerous cucumber circRNAs were detected in response to salt stress that were involved in the modulation of transcription, metabolism adaptation, ion homeostasis related pathways, and the regulation of proline metabolism by modulating the corresponding biosynthesis and degradation genes under salinity conditions [31]. Besides, differentially expressed circRNAs have been observed in pear [62], Arabidopsis, and maize [63] under dehydration conditions. It has been reported that the overexpression of a specific circRNA in tomato, *PSY1-circ1*, which is originated from *PSY1* gene, resulted in a substantial reduction of lycopene and β-carotene in the transgenic tomato, proposing the circRNAs function during the plant developmental stage [64]. Moreover, researchers found that the numbers of circularized exons, alternative circularization events, and lengths of circRNAs can alter under heat stress in Arabidopsis [65]. In contrast to the detection of numerous circRNAs in different plant species under biotic and abiotic stresses, the detailed mechanisms underlying their regulatory roles under these conditions are still poorly understood. As mentioned earlier, the closed-loop structure of circRNAs with no free 5′ or 3′ end protects them against degradation by exoribonucleases. It can be speculated that the long half-life of circRNAs can allow them to operate as slow-responding regulators to different stresses. Moreover, plant circRNAs may act as signaling molecules transferred in a lengthy distance through the xylem and phloem or a short distance via cell to cell. The first identified circRNAs, plant viroids, have been shown to control long-distance and cell-to-cell transmitting via diverse circRNA motifs during plant development [25]. Consequently, it can be assumed that there are probably some functional similarities between the plant innate circRNAs and viroids.

22.5.3 CircRNAs in Gene Expression Regulation

According to increasing pieces of evidence, circRNA can influence the splicing process of their corresponding linear transcripts (Fig. 22.3I) and control the transcription of their derived genes [28] (Fig. 22.3II). During circRNA generation, introns can be kept between circularized exons that create a type of circRNAs called exon-intron circRNAs or EIciRNAs. The EIciRNAs are mostly restricted to the nucleus where their interaction with U1 snRNP can promote their parental gene transcription, indicating the direct gene expression regulation by circRNAs. In soybean, 293 EIciRNAs, composed of 183 and 175 have been identified in resistant and sensitive species during the biotic stress imposed by cotton bollworm feeding, implying the EIcircRNAs contribution to plant stress tolerance [34]. The CircSEP3 in Arabidopsis, a circRNA originated from the exon 6 of the *SEP3* gene, has been determined to control transcription and splicing of the corresponding linear cognates [66] (Fig. 22.3III). This circRNA can strongly bind to its related DNA region and create the RNA: DNA hybrid results in pausing its parental gene expression and reducing the related transcript, whereas the level of a transcript without the exon 6, produced during the exon skipping process, is increased. Thus, circRNAs can operate as trans-acting elements to regulate the expression of their parental genes at both transcription and splicing stages [55, 66]. Similarly, the overexpression of Os08circ16564 in transgenic rice plants can greatly diminish the transcript abundance of its parental gene (AK064900) in different tissues, contrary to non-transgenic plants. Since there were many linear transcripts in the transgenic rice, it has been proposed that circRNAs and their related linear transcripts might function as post-transcriptional regulators of their parental genes [20, 55]. Considering the circRNA functions are strictly linked to their subcellular localization, further studies on circRNA locality may be critical for disclosing their biological roles in plants. Investigating the putative functions of specific plant circRNAs involved in modulating gene expression and alternative splining processes, as well as specifying the circRNA impacts at the phenotypic level are crucial challenges [55].

22.5.4 CircRNAs Can Be Protein-Coding and Act as a Biomarker

In eukaryotic cells, the common translation process is initiated in a cap-related mode, wherein the 40S ribosomes have been recruited to the cap structure at the 5′ end of mRNA transcripts [67, 68]. Due to the lack of 5′ and 3′ ends in the closed-loop structure of circRNAs, they are considered untranslatable transcripts. Nevertheless, recent studies in mammals explained that circRNA can be translated in a cap-independent pathway; actually, IRES elements and m6A (*N*6-Methyladenosine) RNA alteration can trigger the circRNAs translation [21, 69] (Fig. 22.3VI). Though several circRNAs contained potential ORFs with IRES elements, a small number of them have been demonstrated to translate into proteins/peptides, which their biological functions have not been yet uncovered [70, 71]. The IRES factors are deciphered in different organisms, including plants, animals, and viruses. Rice yellow mottle virus is a kind of virus with a small circular RNA genome harboring the IRES elements that can directly translate into 16 kDa proteins [72]. Additionally, genetic engineering-based researches have revealed that in the case of inserting an IRES element upstream of the initiator AUG codon, the circRNAs can be translated to proteins in vitro [73]. As a result of the circular structure of circRNAs, they can have unlimited open reading frames, and long repeated-sequence proteins can be originated from circRNAs in vitro and in vivo. Therefore, we can hypothesize that the IRES elements might also be detected in some plant endogenous circRNAs and trigger their translation. The epigenetic alteration of m6A can also stimulate the 5′ cap-free translation process of mRNA transcripts at both 3′-UTR and 5′-UTR in animals [25]. In plants, the enrichment of m6A has been found not only near the stop codon and within 3′-untranslated regions but also near the start codon of mRNAs in Arabidopsis, unlike mammals. This specific distribution suggests that m6A might also participate in the translation process in plants [74]. Recently, a group of researchers developed a CircCode tool, a computational pipeline written in python language, for predicting the coding capability of circRNAs. The investigation of potentially translated circRNAs from Arabidopsis and humans using CircCode has demonstrated that this tool has an appropriate performance in terms of high sensitivity and high accuracy in both organisms and could identify 1569 and 3610 translated circRNAs in Arabidopsis and humans, respectively [54]. Additionally, the unique characteristics of circRNAs, such as long half-lives, resistance to degradation, and their facility and specificity of detection enable these RNA molecules to function as an appropriate biomarker. In Arabidopsis, circRNAs have proved to act as a valid biomarker of exon-skipped alternative splicing variants, including in the homeotic MADS-box transcription factor family [66]. In contrast to animals, the identification of plant circRNAs, acting as a biomarker, is quite a new and interesting topic that requires further researches.

22.6 Conclusion and Future Perspectives

With the arrival of deep-sequencing technologies, numerous circRNAs have been recognized in diverse plant species. The specific expression patterns of circRNAs under various plant developmental stages and adverse conditions imply their vital roles in plant growth and development, which can be considered as a new level of transcriptional and post-transcriptional gene regulation. However, the mechanism of plant circRNA biogenesis and biological functions is still in its infancy and requires further studies. Therefore, it is worth characterizing circRNAs in more plant species in response to diverse developmental periods and stress situations. Similarly, the expression regulation of parental genes by their derived circRNAs directly or indirectly via the mRNA–circRNA–miRNA network opens a new scope for circRNA researches in plants. Contrarily to animals, the coding ability of circRNAs and their role as biomarkers have not been much surveyed in plants. Therefore, determining and understanding the functional roles of circRNAs in plants can be a promising research topic in the future.

Conflict of Interest None

Additional Information Figure 22.1 (CC BY 4.0) [9] and Fig. 22.3 (CC BY 4.0) [28] and have been reused under Creative Commons Attribution licenses. Figure 22.2 is used with permission (File provided).

References

1. Jeck WR, Sharpless NE. Detecting and characterizing circular RNAs. Nat Biotechnol. 2014;32 (5):453–61.
2. Chen L-L, Yang L. Regulation of circRNA biogenesis. RNA Biol. 2015;12(4):381–8.
3. Salzman J, Gawad C, Wang PL, Lacayo N, Brown PO. Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. PLoS One. 2012;7(2): e30733.
4. Suzuki H, Tsukahara T. A view of pre-mRNA splicing from RNase R resistant RNAs. Int J Mol Sci. 2014;15(6):9331–42.
5. Sanger HL, Klotz G, Riesner D, Gross HJ, Kleinschmidt AK. Viroids are single-stranded covalently closed circular RNA molecules existing as highly base-paired rod-like structures. Proc Natl Acad Sci. 1976;73(11):3852–6.
6. Chu Q, Bai P, Zhu X, Zhang X, Mao L, Zhu Q-H, et al. Characteristics of plant circular RNAs. Brief Bioinform. 2020;21(1):135–43.
7. Mo D, Li X, Raabe CA, Cui D, Vollmar J-F, Rozhdestvensky TS, et al. A universal approach to investigate circRNA protein coding function. Sci Rep. 2019;9(1):1–13.
8. Starke S, Jost I, Rossbach O, Schneider T, Schreiner S, Hung L-H, et al. Exon circularization requires canonical splice signals. Cell Rep. 2015;10(1):103–11.
9. Zhao X, Cai Y, Xu J. Circular RNAs: biogenesis, mechanism, and function in human cancers. Int J Mol Sci [Internet]. 2019 [cited 2021 Feb 14];20(16). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6720291/.
10. Conn SJ, Pillman KA, Toubia J, Conn VM, Salmanidis M, Phillips CA, et al. The RNA binding protein quaking regulates formation of circRNAs. Cell. 2015;160(6):1125–34.
11. Kelly S, Greenman C, Cook PR, Papantonis A. Exon skipping is correlated with exon circularization. J Mol Biol. 2015;427(15):2414–7.
12. Aufiero S, Reckman YJ, Pinto YM, Creemers EE. Circular RNAs open a new chapter in cardiovascular biology. Nat Rev Cardiol. 2019;16(8):503–14.
13. Li X, Yang L, Chen L-L. The biogenesis, functions, and challenges of circular RNAs. Mol Cell. 2018;71(3):428–42.
14. Zuo J, Wang Q, Zhu B, Luo Y, Gao L. Deciphering the roles of circRNAs on chilling injury in tomato. Biochem Biophys Res Commun. 2016;479(2):132–8.
15. Ye C-Y, Chen L, Liu C, Zhu Q-H, Fan L. Widespread noncoding circular RNAs in plants. New Phytol. 2015;208(1):88–95.
16. Gao Z, Li J, Luo M, Li H, Chen Q, Wang L, et al. Characterization and cloning of grape circular RNAs identified the cold resistance-related Vv-circATS1. Plant Physiol. 2019;180(2):966–85.
17. Luo Z, Han L, Qian J, Li L. Circular RNAs exhibit extensive intraspecific variation in maize. Planta. 2019;250(1):69–78.
18. Wang Y, Gao Y, Zhang H, Wang H, Liu X, Xu X, et al. Genome-wide profiling of circular RNAs in the rapidly growing shoots of Moso bamboo (Phyllostachys edulis). Plant Cell Physiol. 2019;60(6):1354–73.
19. Chen L, Ding X, Zhang H, He T, Li Y, Wang T, et al. Comparative analysis of circular RNAs between soybean cytoplasmic male-sterile line NJCMS1A and its maintainer NJCMS1B by high-throughput sequencing. BMC Genomics. 2018;19(1):1–14.
20. Lu T, Cui L, Zhou Y, Zhu C, Fan D, Gong H, et al. Transcriptome-wide investigation of circular RNAs in rice. RNA. 2015;21(12):2076–87.
21. Yang Y, Fan X, Mao M, Song X, Wu P, Zhang Y, et al. Extensive translation of circular RNAs driven by N 6 -methyladenosine. Cell Res. 2017;27(5):626–41.

22. Fan J, Quan W, Li G-B, Hu X-H, Wang Q, Wang H, et al. CircRNAs are involved in the rice-Magnaporthe oryzae interaction. Plant Physiol. 2020;182(1):272–86.
23. Darbani B, Noeparvar S, Borg S. Identification of circular RNAs from the parental genes involved in multiple aspects of cellular metabolism in barley. Front Plant Sci. 2016;7:776.
24. Ye C-Y, Zhang X, Chu Q, Liu C, Yu Y, Jiang W, et al. Full-length sequence assembly reveals circular RNAs with diverse non-GT/AG splicing signals in rice. RNA Biol. 2017;14 (8):1055–63.
25. Li Q-F, Zhang Y-C, Chen Y-Q, Yu Y. Circular RNAs roll into the regulatory network of plants. Biochem Biophys Res Commun. 2017;488(2):382–6.
26. Wang Z, Liu Y, Li D, Li L, Zhang Q, Wang S, et al. Identification of circular RNAs in Kiwifruit and their species-specific response to bacterial canker pathogen invasion. Front Plant Sci [Internet]. 2017 [cited 2021 Feb 14];8. https://www.frontiersin.org/articles/10.3389/fpls.2017.00413/full.
27. Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, et al. Circular RNAs are abundant, conserved, and associated with ALU repeats. RNA. 2013;19(2):141–57.
28. Zhang P, Li S, Chen M. Characterization and function of circular RNAs in plants. Front Mol Biosci [Internet]. 2020 [cited 2021 Feb 14];7. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7248317/
29. Zhao W, Cheng Y, Zhang C, You Q, Shen X, Guo W, et al. Genome-wide identification and characterization of circular RNAs by high throughput sequencing in soybean. Sci Rep. 2017;7 (1):5636.
30. Shen T, Han M, Wei G, Ni T. An intriguing RNA species—perspectives of circularized RNA. Protein Cell. 2015;6(12):871–80.
31. Zhu Y-X, Jia J-H, Yang L, Xia Y-C, Zhang H-L, Jia J-B, et al. Identification of cucumber circular RNAs responsive to salt stress. BMC Plant Biol. 2019;19(1):164.
32. Liu S, Wang Q, Li X, Wang G, Wan Y. Detecting of chloroplast circular RNAs in Arabidopsis thaliana. Plant Signal Behav. 2019;14(8):1621088.
33. Sun X, Wang L, Ding J, Wang Y, Wang J, Zhang X, et al. Integrative analysis of Arabidopsis thaliana transcriptomics reveals intuitive splicing mechanism for circular RNA. FEBS Lett. 2016;590(20):3510–6.
34. Zhao T, Wang L, Li S, Xu M, Guan X, Zhou B. Characterization of conserved circular RNA in polyploid Gossypium species and their ancestors. FEBS Lett. 2017;591(21):3660–9.
35. Sharma D, Sehgal P, Hariprakash J, Sivasubbu S, Scaria V. Methods for annotation and validation of circular RNAs from RNAseq data. In: Computational biology of non-coding RNA. New York: Springer; 2019. p. 55–76.
36. Westholm JO, Miura P, Olson S, Shenker S, Joseph B, Sanfilippo P, et al. Genome-wide analysis of Drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation. Cell Rep. 2014;9(5):1966–80.
37. Ma X-K, Wang M-R, Liu C-X, Dong R, Carmichael GG, Chen L-L, et al. CIRCexplorer3: a CLEAR pipeline for direct comparison of circular and linear RNA expression. Genomics Proteomics Bioinformatics. 2019;17(5):511–21.
38. Gao Y, Wang J, Zhao F. CIRI: an efficient and unbiased algorithm for de novo circular RNA identification. Genome Biol. 2015;16(1):1–16.
39. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, Rybak A, et al. Circular RNAs are a large class of animal RNAs with regulatory potency. Nature. 2013;495(7441):333–8.
40. Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, et al. MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 2010;38(18): e178.
41. Chen L, Yu Y, Zhang X, Liu C, Ye C, Fan L. PcircRNA_finder: a software for circRNA prediction in plants. Bioinformatics. 2016;32(22):3528–9.
42. Zhang P, Liu Y, Chen H, Meng X, Xue J, Chen K, et al. CircPlant: an integrated tool for circRNA detection and functional prediction in plants. Genomics Proteomics Bioinformatics. 2020;18(3):352–8.

43. Tang M, Kui L, Lu G, Chen W. Disease-associated circular RNAs: from biology to computational identification [Internet], vol. 2020. Hindawi: BioMed Research International. 2020 [cited 2021 Feb 14]. p. e6798590. https://www.hindawi.com/journals/bmri/2020/6798590/.
44. Chen L, Wang C, Sun H, Wang J, Liang Y, Wang Y, et al. The bioinformatics toolbox for circRNA discovery and analysis. Brief Bioinform. 2020;22:1706–28.
45. Szabo L, Morey R, Palpant NJ, Wang PL, Afari N, Jiang C, et al. Statistically based splicing detection reveals neural enrichment and tissue-specific induction of circular RNA during human fetal development. Genome Biol. 2015;16(1):126.
46. Hansen TB, Venø MT, Damgaard CK, Kjems J. Comparison of circular RNA prediction tools. Nucleic Acids Res. 2016;44(6):e58.
47. Zeng X, Lin W, Guo M, Zou Q. A comprehensive overview and evaluation of circular RNA detection tools. PLoS Comput Biol. 2017;13(6):e1005420.
48. Zhang P, Meng X, Chen H, Liu Y, Xue J, Zhou Y, et al. PlantCircNet: a database for plant circRNA–miRNA–mRNA regulatory networks. Database [Internet]. 2017 [cited 2021 Feb 14];2017(bax089). https://doi.org/10.1093/database/bax089.
49. Ye J, Wang L, Li S, Zhang Q, Zhang Q, Tang W, et al. AtCircDB: a tissue-specific database for Arabidopsis circular RNAs. Brief Bioinform. 2019;20(1):58–65.
50. Wang K, Wang C, Guo B, Song K, Shi C, Jiang X, et al. CropCircDB: a comprehensive circular RNA resource for crops in response to abiotic stress. Database [Internet]. 2019 [cited 2021 Feb 14];2019(baz053). https://doi.org/10.1093/database/baz053.
51. Meng X, Hu D, Zhang P, Chen Q, Chen M. CircFunBase: a database for functional circular RNAs. Database. 2019;2019:baz003.
52. Zhang J, Hao Z, Yin S, Li G. GreenCircRNA: a database for plant circRNAs that act as miRNA decoys. Database [Internet]. 2020 [cited 2021 Feb 14];2020(baaa039). https://doi.org/10.1093/database/baaa039.
53. Wang H, Wang H, Zhang H, Liu S, Wang Y, Gao Y, et al. The interplay between microRNA and alternative splicing of linear and circular RNAs in eleven plant species. Bioinformatics. 2019;35(17):3119–26.
54. Sun P, Li G. CircCode: a powerful tool for identifying circRNA coding ability. Front Genet. 2019;10:981.
55. Litholdo CG, da Fonseca GC. Circular RNAs and plant stress responses. Circular RNAs. 2018;1087:345–53.
56. Zhao W, Chu S, Jiao Y. Present scenario of circular RNAs (circRNAs) in Plants. Front Plant Sci [Internet]. 2019 [cited 2021 Feb 14];10. https://www.frontiersin.org/articles/10.3389/fpls.2019.00379/full.
57. Wang Y, Xiong Z, Li Q, Sun Y, Jin J, Chen H, et al. Circular RNA profiling of the rice photo-thermosensitive genic male sterile line Wuxiang S reveals circRNA involved in the fertility transition. BMC Plant Biol. 2019;19(1):340.
58. Ghorbani A, Izadpanah K, Peters JR, Dietzgen RG, Mitter N. Detection and profiling of circular RNAs in uninfected and maize Iranian mosaic virus-infected maize. Plant Sci. 2018;274:402–9.
59. Xiang L, Cai C, Cheng J, Wang L, Wu C, Shi Y, et al. Identification of circularRNAs and their targets in Gossypium under Verticillium wilt stress based on RNA-seq. PeerJ. 2018;6:e4500.
60. Wang J, Yang Y, Jin L, Ling X, Liu T, Chen T, et al. Re-analysis of long non-coding RNAs and prediction of circRNAs reveal their novel roles in susceptible tomato following TYLCV infection. BMC Plant Biol. 2018;18(1):104.
61. Wang Y, Yang M, Wei S, Qin F, Zhao H, Suo B. Identification of circular RNAs and their targets in leaves of Triticum aestivum L. under dehydration stress. Front Plant Sci [Internet]. 2017 [cited 2021 Feb 14];7. https://www.frontiersin.org/articles/10.3389/fpls.2016.02024/full.
62. Wang J, Lin J, Wang H, Li X, Yang Q, Li H, et al. Identification and characterization of circRNAs in Pyrus betulifolia Bunge under drought stress. PLoS One. 2018;13(7):e0200692.
63. Zhang P, Fan Y, Sun X, Chen L, Terzaghi W, Bucher E, et al. A large-scale circular RNA profiling reveals universal molecular mechanisms responsive to drought stress in maize and Arabidopsis. Plant J. 2019;98(4):697–713.

64. Tan J, Zhou Z, Niu Y, Sun X, Deng Z. Identification and functional characterization of tomato CircRNAs derived from genes involved in fruit pigment accumulation. Sci Rep. 2017;7 (1):8594.
65. Pan T, Sun X, Liu Y, Li H, Deng G, Lin H, et al. Heat stress alters genome-wide profiles of circular RNAs in Arabidopsis. Plant Mol Biol. 2018;96(3):217–29.
66. Conn VM, Hugouvieux V, Nayak A, Conos SA, Capovilla G, Cildir G, et al. A circRNA from SEPALLATA3 regulates splicing of its cognate mRNA through R-loop formation. Nat Plants. 2017;3(5):1–5.
67. Aitken CE, Lorsch JR. A mechanistic overview of translation initiation in eukaryotes. Nat Struct Mol Biol. 2012;19(6):568.
68. Sonenberg N, Hinnebusch AG. Regulation of translation initiation in eukaryotes: mechanisms and biological targets. Cell. 2009;136(4):731–45.
69. Meyer KD, Patil DP, Zhou J, Zinoviev A, Skabkin MA, Elemento O, et al. 5′ UTR m6A promotes cap-independent translation. Cell. 2015;163(4):999–1010.
70. Legnini I, Di Timoteo G, Rossi F, Morlando M, Briganti F, Sthandier O, et al. Circ-ZNF609 is a circular RNA that can be translated and functions in myogenesis. Mol Cell. 2017;66(1):22–37.
71. Pamudurti NR, Bartok O, Jens M, Ashwal-Fluss R, Stottmeister C, Ruhe L, et al. Translation of circRNAs. Mol Cell. 2017;66(1):9–21.
72. AbouHaidar MG, Venkataraman S, Golshani A, Liu B, Ahmad T. Novel coding, translation, and gene expression of a replicating covalently closed circular RNA of 220 nt. Proc Natl Acad Sci. 2014;111(40):14542–7.
73. Chen C, Sarnow P. Initiation of protein synthesis by the eukaryotic translational apparatus on circular RNAs. Science. 1995;268(5209):415–7.
74. Luo G-Z, MacQueen A, Zheng G, Duan H, Dore LC, Lu Z, et al. Unique features of the m 6 a methylome in Arabidopsis thaliana. Nat Commun. 2014;5(1):1–8.

Application of Metagenomics in Improvement of Rice

23

Pallabi Pati, Gayatri Gouda, and Sushil Kumar Rathore

Abstract

Microorganisms can grow and develop in a wide variety of environmental conditions. However, the majority of them are not cultivable or were never cultivable. Due to metagenomics, now it is possible to study all microorganisms regardless of whether they can be cultured or not by using genomic data obtained directly from an environmental sample to survey microbial diversity and their role in their specific community level, as well as the environment and other communities. So, the microbial screening based on their specific function helps discover novel proteins for many purposes of the scientific and industrial field. Metagenomics also helps provide solutions to many aspects, such as unmasking the microbe's specific function in a specific community. The information regarding inter and intra microbial community interaction and their sequence data analysis identifies the unknown microflora found from environmental samples.

Keywords

Genomic library · 16S rRNA gene · Metagenomic sequence · High throughput sequencing · NGS · Plasmids · Rhizosphere · Sequence alignment · Sequence assembly · Rice

P. Pati (✉)
District Headquarter Hospital, Ganjam, Odisha, India

G. Gouda
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. K. Rathore
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_23

Abbreviations

BAC	Bacterial artificial chromosome
BLAST	Basic Local Alignment Search Tool
CRW	Chinese Rice Wine
ITS2	Internal Transcribed Spacer II
MBN	Microbial biomass nitrogen
PCR-DGGE	Polymerase chain reaction-denaturing gel gradient electrophoresis
QC	Quality control
RFLP	Restriction fragment length polymorphism

23.1 Introduction

Metagenomics deals with the microbial population diversity present in the soil. These soil microbes are very much useful for the rice plant. To study the diversity of microbes in the plant is very much essential to know their structural and functional characteristics. This can be possible by studying the genomic association of microbes with the plant species present in the particular environment where a large, diverse microbial population is adapted [1]. Microorganisms are major elements of soil ecology, biogeochemical mechanisms, and biodiversity. Recent studies have generally suggested that habitat, evolutionary, and balanced systems interact to establish the composition of the microbiome population through spatio-temporal scales in the environment [2]. Microbial ecology is associated with the diversity of microorganisms and their interactions with one another and their ecosystem to produce and retain that diversity. As a result, microbial ecologists have commonly concentrated their attention on two areas of study [3]. The process of studying the microbial diversity present in the soil sample at the DNA level is termed metagenomics. The information on microbial diversity, structure and gene function pattern, and metabolic pathway at the genome level provides an idea about their contribution toward plant growth and development. Microbial function refers to some of the functions of microorganisms in their habitats and how their interactions relate to detected microbial diversity and biogeochemical cycling. Metagenomics has the potential to generate new molecules and novel enzymes with a wider variety of functions and improved properties as compared to enzymes from culturable microorganisms [4].

The microbial population in rice rhizosphere soils has gained a lot of attention due to the fact that numerous microorganisms act in soil functional processes [5]. The rhizosphere is found in the soil present adjacent to the roots of plant species. The rhizosphere microbes are mainly protecting the plant by secreting various organic and inorganic substances, i.e., root exudates that induce resistance from specific damage or stress that occurred in the environment. The microbes that are found at the root of the rice plant helps it from biotic and abiotic stress [6]. Within biotic factors, the plants interact with the microbial communities to attain resistance, called the plant microbiome. Rice microbiome classifies microorganisms linked with rice and its environment. Rice crop production faces many threats, including climate change, population growth, and a growing need for sustainable production. As the plant microbiome plays an important function in crop improvement, various strategies are developed to improve the plant microbiome [7]. Due to the introduction of high-throughput platforms, other molecular techniques, such as polymerase chain reaction-denaturing gel gradient electrophoresis (PCR-DGGE), 16S rRNA gene cloning, and terminal restriction fragment length polymorphism analysis, were used to characterize microbial populations in soil and rice tissues [8].

To develop strategies for sustainable agriculture, an integrated approach to studying the bacterial and fungal communities associated with rice is necessary. Agroecosystems are extremely complex organisms that include the crop plant and various biotic and abiotic factors, such as bacteria and fungi. Among these insects, a significant role is played in the development of crop plants. Thus, when designing microbiome-based techniques for crop productivity enhancement, the relationship between crop plants, microbiome, and insects must be recognized. Metagenomics involves using a combination of genomic technology and bioinformatics techniques to obtain direct access to the genetic information contained in the entire species. Nowadays, the field of metagenomics has made significant advances in microbial ecology, evolution, and diversity. An increase in the number of events brings an increase in the number of methodological skills and experience that can drive potential progress in the area [9].

This chapter reviews the function and method of metagenomic study in rice cultivation using sequencing technology. This technique specifies the detailed information of the rice microbiome in enhancing the potential resistance against stress and its role in crop improvement.

23.2 Metagenome Study

The metagenomic analysis involves direct isolation of bacterial DNA, library construction, and functional analysis. Initially, metagenomic studies were applied to specific experiments designed to detect microbial diversity in the ecosystem, with metagenomic libraries assembled using 16S rDNA amplicons [10]. With the development of NGS, large-scale shotgun metagenomic projects became possible.

The methods for metagenome analysis involve (1) collection of the soil sample, (2) genomic DNA extraction and purification, (3) cloning of metagenomic DNA, (4) library construction and sequencing, (5) screening of metagenomic libraries.

1. *Collection of Soil samples:* For rice, the samples are collected from soil and rhizosphere, where the microbes are present in the roots. During the different stages of growth and development of rice plants, the sample is collected to study the function of microbes. All samples were homogenized and sieved through a 2-mm mesh, then divided into two subsamples: one was partly air-dried for

chemical examination and the other was put in sterile 50-mL containers and kept at 4 °C in the dark for no more than 24 h for microbiological analysis [11].

2. *Genomic DNA extraction:* The extraction of soil DNA is an important stage in these metagenomic approaches and can be classified into two broad strategies. The first and most frequently used technique is direct DNA extraction, which involves the lysis of cells directly within a soil sample. The second technique, indirect DNA extraction, starts with the extraction of cells from soil [12]. As a large quantity of DNA is required for the metagenomics study, the direct extraction method is the best method to extract a high quantity of DNA. The DNA is purified to eliminate the contaminants present in the genomic DNA sample used for sequencing study [13]. DNA molecules are broken into fragments tiny enough to be sequenced. Following that, blunt ends are added to the pieces to facilitate further sorting. Finally, the pieces are ligated to the adaptors. Mechanical or enzymatic fragmentation methods are available, with the later divided up into nebulization, hydrodynamic shearing, and ultrasonication [14].
3. *Cloning of metagenomic DNA:* Cloning large segments of DNA extracted directly from microbes in their natural habitats allows access to soil metagenomic DNA. Many cloning vectors are available, such as plasmids, fosmids, lambda vectors, and bacterial artificial chromosomes (BACs), for cloning of DNA according to the DNA fragments to be cloned. The vector used impacts the size of inserts that can be cloned and the frequency of expressing metagenomic genes.
4. *Library construction:* Large DNA fragments, that is, 25–200 kb isolated from soil samples and cloned into particular vectors, may be used to generate metagenomic libraries. The vector to use is determined by the length of the insert to be cloned. To prevent noisy sequencing data, the free adaptor, adaptor dimers, and any other components must be removed [15].
5. *Sequencing of metagenomic DNA:* The advance of DNA sequencing and bioinformatics analysis enables the discovery of the genetic complexity of the host-associated microbial communities as well as uncultured microbial diversity. High-throughput sequencing technologies allow the generation of large quantities of sequence data in a short period and at a low cost. The ability of this technology to distinguish huge amounts of organisms from diverse ecosystems is one of its most significant applications. While the novel, cloning-independent pyrosequencing technology has tremendous potential for a wide variety of phyla of soil/sediment-dwelling species, the majority of a new study has mainly focused on the biodiversity of prokaryotic populations. Various sequencing technologies have been used in the metagenomic study that includes NGS, Illumina, 454 pyrosequencing.

23.3 High-Throughput Sequencing

The quality, speed, and cost of high-throughput sequencing technology are rapidly improving. As a result, it is increasingly being used to research whole populations of prokaryotes in a variety of niches. High-throughput sequencing, alternatively referred to as next-generation sequencing (NGS), has advanced genomic analysis. NGS technology has gradually advanced in recent years, with costs falling and the amount and variety of sequencing applications expanding exponentially. Without prior preparation, next-generation sequencing enables the analysis and identification of species directly from their environments [16]. In comparison to first-generation sequencing, NGS is capable of simultaneously generating several hundred thousand to millions of sequencing reads. Additionally, sequencing may be produced without some traditional measures, such as vector-based cloning, which minimizes the possibility of DNA contamination from other species. As a result, many next-generation sequencing technologies, including the Roche 454, Illumina,® Applied Biosystems SOLiD sequencer, and Ion Torrent, have been launched. All next-generation sequencing or real-time sequencing technologies (Roche 454, Illumina,® and AB SOLiD) rely on optical sensors that detect luminescent signals produced by base insertion [17, 18]. Now shotgun metagenomic sequencing is a method for analyzing uncultured microbiota. In this method, the DNA is isolated and sheared into small fragments and sequenced separately. This generates DNA sequences called reads which map to several genomic positions for the sample's diverse genomes, including non-microbes [19].

23.4 Metagenomic Data Analysis

Metagenomic data is a set of DNA fragments from various species, which can include microbial, bacterial, or eukaryotic organisms. Many approaches may be used for metagenomics data analysis that includes the collection of raw reads, QC tools, viz., FastQ, AfterQC to detect and eliminate low-quality sequences and contaminants. FastQ Screening of raw reads enables to screen a library of FastQ sequences against a set of sequence databases. After QC analysis, the reads may be arranged into longer contiguous sequences known as contigs, or they can be directly passed to taxonomic classifiers. In taxonomic classification, the direct identification of genetic information and species with close relatives in the database is possible. Metagenomic shotgun sequencing removes primer sampling error and allows the identification of species from all aspects of life, assuming DNA can be isolated from the target population. The trimming method is initiated for adaptor removal and then the masking process is carried out. In this process, the target genes are separated from the unmapped genes present in the genome sequence.

(a) *Marker gene analysis:* Marker gene analysis is one of the simplest and more computationally effective methods for quantifying the taxonomic complexity of a metagenome. This method involves mapping metagenomic reads to a collection of marker genes, finding reads which are marker gene homologs, and taxonomically annotating each metagenomic homolog utilizing sequence or phylogenetic similarities to the marker gene database sequences. The most commonly used marker genes are rRNA or protein-coding genes, which are usually single-copy and ubiquitous in microbial genomes.

(b) *Sequence assembly:* Genome assembly is the process of reconstructing genomes from smaller DNA segments known as reads obtained during a sequencing method [20]. Assembly combines highly correlated metagenomic reads from the same genome into a single contiguous sequence and is important for producing longer sequences that facilitate bioinformatics analysis in comparison to unassembled short metagenomic reads [21]. Mostly, reads are pair-ended or mate-paired, which indicates they are sequenced from the same DNA fragment. The distance between each pair of reads is generally known to overcome ambiguities introduced by repeated sequences during assembly [22]. Metagenomic sequence data assembly into microbial genomes is essential for enhancing our knowledge of microbial ecology and metabolism by illustrating the functional ability of complicated microorganisms [23].

For metagenomics data, two types of assembly methods are available, i.e., reference-based assembly or comparative assembly and de novo assembly.

Reference-based assembly: In reference-based assembly, the reference genome sequence is available for metagenome sequences. The variations between the original genomes of the metagenome sequence and the reference, such as major insertions, deletions, or polymorphisms, may indicate that the assembly is fragmented or that divergent regions are fully removed. The assembly process involves two steps, where all reads are aligned against the reference genome; then, by inferring the alignments, a consensus sequence is formed. This approach is more successful at resolving repeats than de novo assembly and therefore produces better results than de novo approaches, especially at low coverage depths. Long repeats remain a major concern since they result in an ambiguous alignment of reads to the genome, but the use of mate-pair data may partially minimize this problem by assisting in the identification of the appropriate read placement.

De novo assembly: The de novo assembly approach involves the reconstruction of the genome from the read data in which the reference genome is not available as this assembly process requires more computational resources, so new tools are developed for analyzing a large amount of dataset. The most used tool is greedy, De Bruijn graph, and overlap-layout consensus [24–26]. In the greedy method, individual reads are combined into contigs iteratively, starting with the overlapping reads and end with contigs where no more sequence is available. The de Bruijn graph assembly is based on the relation between read-derived substrings of fixed length *k* (*k*-mers) [27]. The *k*-mers are arranged in a graph form, with nodes representing the *k*-1 prefixes and *k*-mers suffixes and edges. Reads are not directly related to this approach; rather, their overlap is inferred by the fact that they share *k*-mers [28]. In 2011, Peng at al developed an algorithm for de novo assembly, namely Meta-IDBA, that divides the de Bruijn graph into independent components of different species. It captures minor variants in the genomes of subspecies within the same species through multiple alignments and represents the genome of a single species using a consensus sequence [29].

(c) *Sequence binning:* A technique known as binning helps to classify each metagenomic sequence.

Binning is the method of grouping DNA sequences comprising of a single genome or genomes from closely related species. Different algorithms have been designed that allow two distinct types of information to present within a target DNA sequence. In general, each sequence is either grouped into a taxonomic group based on comparison to certain reference data or grouped into groups of sequences that reflect taxonomic groups based on similar characteristics, such as GC content. Sequence binning may be implemented by the use of homology-based approaches, such as BLAST, or by the use of composition-based approaches that compare nucleotide frequency patterns. The compositional binning exploits the fact that genomes share a conserved nucleotide structure with specific GC content [30]. However, the unknown DNA fragment can encode a gene, and the sequence similarity to identified genes in a reference database may be used to identify and thereby binning of the sequence. It reduces the scope of the data, allowing for separate post-binning analysis, viz., assembly of each sample of binned reads rather than on the whole population of data. Binning may be performed on assembled or unassembled files, but most algorithms claim that the precision of binning increases with sequence duration. In general, binning algorithms fall into three categories: sequence structure, sequence similarity, and fragment recruitment [9]. Binding algorithms based on sequence similarity and sequence composition markers probably depend on reference genomes or phylogenetic markers from known microorganisms. These algorithms could not be valid in all cases due to the lack of reference genomes and the bias and lack of markers. Unsupervised binning algorithms are an alternative method that can accommodate fragments from unknown organisms [31]. The various tools used for the analysis are mentioned in (Table 23.1).

23.5 Functional Aspect of Metagenomics Study in Rice

Metagenomics study has been applied in rice research in various aspects, and this shows tremendous applications in rice plants. Several studies have been reported for the metagenomic approach and its application toward the development of rice plants by controlling both pathways and mechanisms. In 2018, Kunda et al. reported the diversity of the endophytic bacterial community in rice function using an amplicon metagenomics approach by targeting the 16s rRNA gene. Their analysis found that in the coastal saline condition of West Bengal, the rice species show resistance against salt stress and various environmental stress resulting from the wide diversity of endophytic bacteria [56]. The bacterial diversity associated with rice rhizosphere bacterial communities from a paddy field environment in Kerala was investigated using culture-independent molecular techniques, including 16S rRNA clone library production, RFLP, sequencing, and phylogenetic analysis. Clones with high similarity to database sequences as well as uncultured bacterial sequences were

Table 23.1 Lists of statistical tools used in a metagenomics study

S. No.	Statistical tools	Application	References
1	Metaviz	To analyze the metagenomic data of annotated microbiome	[32]
2	Metastats	For comparing the metagenomic samples	[33]
3	MEGAN	To analyze large metagenomic data	[34]
4	WebMGA	Analysis of complex metagenome data	[35]
5	ShotgunFunctionalizeR	Compares the functional properties of specific genes and whole pathways from the metagenome data	[36]
6	METAGENassist	Analysis of comparative metagenome	[37]
7	MicrobiomeAnalyst	Combines recent advances in statistics and visualization techniques with novel knowledge bases to allow comprehensive analysis of specific microbiome data	[38]
8	MetaSim	Genomics and metagenomics sequencing simulator	[39]
9	VirFinder	Tool for identification of prokaryotic viral sequences, for metagenomic-based studies	[40]
10	MG-RAST	Shotgun metagenome data analysis	[41]
11	PhyloSift	Integrating microbial community DNA sequencing with evolutionary simulation and phylogenetic research	[42]
12	MetaPhyler	Phylogenetic composition identification of metagenome sequence	[43]
13	Megahit	For complex and large metagenomics assembly	[44]
14	Phylopythia	Enables accurate recognition of the majority of sequence fragments in all taxonomic ranks, including those from unknown species	[45]
15	S-GSOM	Used to easily distinguish such species without seeds	[46]
16	Sort-ITEMS	For metagenomic sequences taxonomic estimation	[47]
17	TACAO	For identifying the taxonomic origin of genomic fragments as short as 800 bp	[48]
18	CARMA	Source of organism prediction	[49]
19	Phymm and PhymmBL	To classify reads as short as 100 base pairs with accuracy	[50]
20	CAMERA	A tool for high-performance computational infrastructure for metagenomic data processing	[51]
21	SPHINX	A taxonomic binning algorithm for metagenomic sequences	[52]
22	TETRA	A tool for composition binning	[53]
23	MetaCluster-TA	Annotating metagenomic data taxonomically using assembly-assisted binning	[54]
24	PCAHIER	For short metagenomic fragments binning	[55]

discovered in the rhizosphere bacterial population. The 16S rRNA sequence analysis revealed a high degree of diversity in the rhizosphere bacterial population, with most microbes closely linked to the Proteobacteria. Just a small proportion of the 16S rRNA sequences is extremely similar to those of Acidobacteria, Firmicutes, and Bacteriodetes. Given the complex metabolic properties of rhizosphere-associated microbes and their critical function in plant health, knowledge of their population structure is critical for a proper understanding of their functions and metagenomics [57].

Bhattacharyya et al. identified the bacterial diversity in the low-land rice rhizosphere and studied the composition of carbon and nitrogen in the soil. Rice rhizospheres with elevated CO_2 + temperature (e-CO_2T) exhibited greater structural diversity and functional behaviors associated with nitrogen metabolism, including nitrogen fixation, assimilatory and dissimilatory nitrate depletion, and denitrification, than rice rhizospheres with ambient CO_2 (a-CO_2). Among the three N metabolism pathways, dissimilarity pathways were more prevalent in low-land rice rhizospheres and even more so in the presence of e-CO_2T. As a result, CH_4 emission, microbial biomass nitrogen (MBN), and dehydrogenase activities were 45%, 20%, and 35% higher under e-CO_2T, respectively, than under a-CO_2 [58]. Panneerselvam et al. studied the important microbes present in the rice rhizosphere using different statistical models. They reported that the *Enterobacter* species have the potential to reduce nitrogen stress in rice plants grown naturally by increasing plant nitrogen uptake through the cumulative contribution of nitrogen-fixing and phytohormone synthesis traits of heterotrophic bacterial diazotrophs [59]. Erkel et al. constructed Rice Cluster I (RC-I) in rice for methanogenic archaea to produce methane. Using a metagenomic approach, they studied the methanogens and demonstrated the aerotolerant, H_2/CO_2-dependent lifestyle and previously unknown methanogen enzymatic capacities for carbohydrate metabolism and assimilatory sulfate reduction. These capabilities, along with a special group of antioxidant enzymes and DNA repair mechanisms, as well as oxygen-insensitive enzymes, confer a comparative advantage on RC-I over other methanogens in its environments, illustrating how RC-I methanogens are so prevalent in the rice rhizosphere [60]. A metagenomic method was used by Zecchin et al. to classify novel microorganisms involved in the sulfur cyclin production in rice paddy [61]. Their study in *Nitrospirae* bacterium Nbg-4 found the role of gypsum in rice growth. Nbg-4 encoded the whole dissimilatory sulfate reduction pathway and was expressed in anoxic bulk soil amended with gypsum, as discovered by parallel metaproteomics. The study reported the function of *Nitrospirae* bacterium species in rice development by regulating the sulfur cycle [61].

In 2016, Bora et al. reported the microbial diversity in traditional rice using whole-genome shotgun sequencing approach. They studied the function of microbes in the production of wine from the traditional rice in the region of Assam. Metagenomic analysis showed that the maximum availability of microbes such as *Lactobacillus plantarum, Meyerozyma gulliermondii, Mucor circinelloides,* and *Rhizopus delemar* in the starter rice increases the fermentation rate [62]. Hong et al. conducted metagenome analysis on bacterial 16S rRNA gene and fungal

Internal Transcribed Spacer II (ITS2) to study the role of the microbiome in Chinese Rice Wine (CRW) production using glutinous rice. Their result revealed that microbe metabolisms have an impact on the quality of wine and both the cultivation of favorable microbes and the inhibition of undesirable microbes are essential for the industrial brewery [63]. Aslam et al. studied the effects of traditional and no-tillage practices on bacterial communities in rhizosphere soil were calculated during rice cultivation using a culture-dependent approach for analyzing the bacterial community structure. Consequently, the actinobacterial population was analyzed using metagenomic libraries constructed using actinobacterial- and streptomycete-specific primers [64]. Imchen et al. used a targeted amplicon-based (16S rRNA gene) metagenomic method to characterize the rhizobiome and bulk soil from rice paddy on a spatial and temporal scale. Their study revealed that rhizobiome modulation occurs during the rice plant's development. The number and diversity of plant growth promoting bacteria increased during the growth stages, as determined by 16S rRNA gene affiliation at the genus level [65]. In 2013, Yeh et al. identified the *GH12* cellulolytic gene, *RSC-EG1*, from rice straw composts. Cellulase from *Micromonospora aurantiaca* and *Thermobispora* sp. were found to be significantly identical to the known cellulolytic gene at the amino acid stage. *RSC-EG1* includes a stretch of nearly 86 amino acids and is novel endoglucanase that is stable over a large temperature range and pH range [66].

23.6 Conclusion and Future Perspectives

These metagenomic studies represent a step toward becoming a final commercial product between discovering the interesting active agent and its formulation. Once deficiencies and other problems in the plant population have been corrected, further studies must be conducted to qualify the agent at an industrial level and to ensure the development of an efficient and truly robust product. The acceptability by the relevant regulatory authorities of the novel enzyme or bioactive and its source microorganism must also be considered. After these limitations are overcome, functional metagenomics offers the possibility to develop new innovative products which offer new and useful industrial processes or even improve or pave the way for the current process to be carried out more conveniently by having access to the seemingly infinite diversity of the microbial world. Metagenomics will provide solutions to many unanswerable questions as well as can solve many mysteries. It requires more research to optimize its potentialities.

Conflicts of Interest None

References

1. Jacoby R, Peukert M, Succurro A, Koprivova A, Kopriva S. The role of soil microorganisms in plant mineral nutrition—current knowledge and future directions. Front Plant Sci [Internet]. 2017 [cited 2021 Apr 25];8. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5610682/.
2. Saleem M, Pervaiz ZH, Traw MB. Theories, mechanisms and patterns of microbiome species coexistence in an era of climate change. In: Saleem M, editor. Microbiome community ecology. Fundamentals and application [Internet]. Cham: Springer; 2015 [cited 2021 Apr 25]. p. 13–53. https://doi.org/10.1007/978-3-319-11665-5_2.
3. Xu J. Microbial ecology in the age of genomics and metagenomics: concepts, tools, and recent advances. Mol Ecol. 2006;15:1713–31.
4. Bashir Y, Pradeep Singh S, Kumar Konwar B. Metagenomics: An Application Based Perspective. Chin J Biol. 2014;2014:e146030.
5. Ding L-J, Cui H-L, Nie S-A, Long X-E, Duan G-L, Zhu Y-G. Microbiomes inhabiting rice roots and rhizosphere. FEMS Microbiol Ecol [Internet]. 2019 [cited 2021 Apr 25];95. https://doi.org/10.1093/femsec/fiz040.
6. Turner TR, James EK, Poole PS. The plant microbiome. Genome Biol. 2013;14:209.
7. Compant S, Samad A, Faist H, Sessitsch A. A review on the plant microbiome: ecology, functions, and emerging trends in microbial application. J Adv Res. 2019;19:29–37.
8. Kim H, Lee Y-H. The rice microbiome: a model platform for crop Holobiome. Phytobiomes J. 2019;4:5–18.
9. Thomas T, Gilbert J, Meyer F. Metagenomics - a guide from sampling to data analysis. Microb Inform Exp. 2012;2:3.
10. Handelsman J. Metagenomics: application of genomics to uncultured microorganisms. Microbiol Mol Biol Rev. 2004;68:669–85.
11. Iliev I, Marhova M, Kostadinova S, Gochev V, Tsankova M, Ivanova A, et al. Metagenomic analysis of the microbial community structure in protected wetlands in the Maritza River basin. Biotechnol Biotechnol Equip. 2019;33:1721–32.
12. Martin-Laurent F, Philippot L, Hallet S, Chaussod R, Germon JC, Soulas G, et al. DNA extraction from soils: old Bias for new microbial diversity analysis methods. Appl Environ Microbiol. 2001;67:2354–9.
13. Delmont TO, Robe P, Clark I, Simonet P, Vogel TM. Metagenomic comparison of direct and indirect soil DNA extraction approaches. J Microbiol Methods. 2011;86:397–400.
14. Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. PLoS One [Internet]. 2011 [cited 2021 May 13];6. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3227650/.
15. Head SR, Komori HK, LaMere SA, Whisenant T, Van Nieuwerburgh F, Salomon DR, et al. Library construction for next-generation sequencing: overviews and challenges. BioTech Future Sc. 2014;56:61–77.
16. Mardis ER, Next-Generation DNA. Sequencing methods. Annu Rev Genomics Hum Genet. 2008;9:387–402.
17. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere”. Proc Natl Acad Sci Natl Acad Sci. 2006;103:12115–20.
18. Schadt EE, Turner S, Kasarskis A. A window into third-generation sequencing. Hum Mol Genet. 2010;19:R227–40.
19. Weinstock GM. Genomic approaches to studying the human microbiota. Nature. 2012;489:250–6.
20. Kececioglu JD, Myers EW. Combinatorial algorithms for DNA sequence assembly. Algorithmica. 1995;13:7.
21. Kunin V, Copeland A, Lapidus A, Mavromatis K, Hugenholtz P. A Bioinformatician’s guide to metagenomics. Microbiol Mol Biol Rev. 2008;72:557–78.

22. Treangen TJ, Salzberg SL. Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nat Rev Genet. 2011;13:36–46.
23. Sangwan N, Xia F, Gilbert JA. Recovering complete and draft population genomes from metagenome datasets. Microbiome. 2016;4:8.
24. Iqbal Z, Caccamo M, Turner I, Flicek P, McVean G. De novo assembly and genotyping of variants using colored de Bruijn graphs. Nat Genet. 2012;44:226–32.
25. Lin Y-Y, Hsieh C-H, Chen J-H, Lu X, Kao J-H, Chen P-J, et al. De novo assembly of highly polymorphic metagenomic data using in situ generated reference sequences and a novel BLAST-based assembly pipeline. BMC Bioinform. 2017;18:223.
26. Haider B, Ahn T-H, Bushnell B, Chai J, Copeland A, Pan C. Omega: an overlap-graph de novo assembler for metagenomics. Bioinformatics. 2014;30:2717–22.
27. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008;18:821–9.
28. Compeau PEC, Pevzner PA, Tesler G. How to apply de Bruijn graphs to genome assembly. Nat Biotechnol. 2011;29:987–91.
29. Peng Y, Leung HCM, Yiu SM, Chin FYL. Meta-IDBA: a de novo assembler for metagenomic data. Bioinformatics. 2011;27:i94–101.
30. Sharpton TJ. An introduction to the analysis of shotgun metagenomic data. Front Plant Sci [Internet]. Frontiers; 2014 [cited 2021 Apr 27];5. https://www.frontiersin.org/articles/10.3389/fpls.2014.00209/full#B63.
31. Leung HCM, Yiu SM, Yang B, Peng Y, Wang Y, Liu Z, et al. A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. Bioinformatics. 2011;27:1489–95.
32. Wagner J, Chelaru F, Kancherla J, Paulson JN, Zhang A, Felix V, et al. Metaviz: interactive statistical and visual analysis of metagenomic data. Nucleic Acids Res. 2018;46:2777–87.
33. Paulson JN, Pop M, Bravo HC. Metastats: an improved statistical method for analysis of metagenomic data. Genome Biol. 2011;12:P17.
34. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. Genome Res. 2007;17:377–86.
35. Wu S, Zhu Z, Fu L, Niu B, Li W. WebMGA: a customizable web server for fast metagenomic sequence analysis. BMC Genomics. 2011;12:444.
36. Kristiansson E, Hugenholtz P, Dalevi D. ShotgunFunctionalizeR: an R-package for functional comparison of metagenomes. Bioinformatics. 2009;25:2737–8.
37. Arndt D, Xia J, Liu Y, Zhou Y, Guo AC, Cruz JA, et al. METAGENassist: a comprehensive web server for comparative metagenomics. Nucleic Acids Res. 2012;40:W88–95.
38. Dhariwal A, Chong J, Habib S, King IL, Agellon LB, Xia J. MicrobiomeAnalyst: a web-based tool for comprehensive statistical, visual and meta-analysis of microbiome data. Nucleic Acids Res. 2017;45:W180–8.
39. Richter DC, Ott F, Auch AF, Schmid R, Huson DH. MetaSim—a sequencing simulator for genomics and metagenomics. PLoS One. 2008;3:e3373.
40. Ren J, Ahlgren NA, Lu YY, Fuhrman JA, Sun F. VirFinder: a novel k-mer based tool for identifying viral sequences from assembled metagenomic data. Microbiome. 2017;5:69.
41. Glass EM, Wilkening J, Wilke A, Antonopoulos D, Meyer F. Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harb Protoc. 2010;2010: pdb.prot5368.
42. Darling AE, Jospin G, Lowe E, Iv FAM, Bik HM, Eisen JA. PhyloSift: phylogenetic analysis of genomes and metagenomes. PeerJ. 2014;2:e243.
43. Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics. 2011;12:S4.
44. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics. 2015;31:1674–6.

45. McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I. Accurate phylogenetic classification of variable-length DNA fragments. Nat Methods. 2007;4:63–72.
46. Chan C-KK, Hsu AL, Halgamuge SK, Tang S-L. Binning sequences using very sparse labels within a metagenome. BMC Bioinform. 2008;9:215.
47. Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS. SOrt-ITEMS: sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics. 2009;25:1722–30.
48. Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. TACOA – taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinform. 2009;10:56.
49. Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, et al. Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Res. 2008;36:2230–9.
50. Brady A, Salzberg SL. Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. Nat Methods. 2009;6:673–6.
51. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. CAMERA: a community resource for metagenomics. PLoS Biol. 2007;5:e75.
52. Mohammed MH, Ghosh TS, Singh NK, Mande SS. SPHINX—an algorithm for taxonomic binning of metagenomic sequences. Bioinformatics. 2011;27:22–30.
53. Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO. TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinform. 2004;5:163.
54. Wang Y, Leung HCM, Yiu SM, Chin FYL. MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. BMC Genomics. 2014;15:S12.
55. Zheng H, Wu H. Short prokaryotic dna fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. J Bioinform Comput Biol. 2010;8:995–1011.
56. Kunda P, Dhal PK, Mukherjee A. Endophytic bacterial community of rice (Oryza sativa L.) from coastal saline zone of West Bengal: 16S rRNA gene based metagenomics approach. Meta Gene. 2018;18:79–86.
57. Arjun JK, Haikrishnan K. Metagenomic analysis of bacterial diversity in the rice rhizosphere soil microbiome. Biotechnol Bioinf Bioeng. 2011;1:361–7.
58. Bhattacharyya P, Roy KS, Das M, Ray S, Balachandar D, Karthikeyan S, et al. Elucidation of rice rhizosphere metagenome in relation to methane and nitrogen metabolism under elevated carbon dioxide and temperature using whole genome metagenomic approach. Sci Total Environ. 2016;542:886–98.
59. Panneerselvam P, Senapati A, Sharma L, Nayak AK, Kumar A, Kumar U, et al. Understanding rice growth-promoting potential of Enterobacter spp isolated from long-term organic farming soil in India through a supervised learning approach. Curr Res Microb Sci. 2021;2:100035.
60. Erkel C, Kube M, Reinhardt R, Liesack W. Genome of rice cluster I Archaea–the key methane producers in the rice rhizosphere. Science. 2006;313:370–2.
61. Zecchin S, Mueller RC, Seifert J, Stingl U, Anantharaman K, Bergen M von, et al. Rice paddy nitrospirae carry and express genes related to sulfate respiration: proposal of the new genus "Candidatus Sulfobium". Appl Environ Microbiol [Internet]. 2018 [cited 2021 May 12];84. https://aem.asm.org/content/84/5/e02224-17.
62. Bora SS, Keot J, Das S, Sarma K, Barooah M. Metagenomics analysis of microbial communities associated with a traditional rice wine starter culture (Xaj-pitha) of Assam, India. 3 Biotech. 2016;6:153.

63. Hong X, Chen J, Liu L, Wu H, Tan H, Xie G, et al. Metagenomic sequencing reveals the relationship between microbiota composition and quality of Chinese Rice Wine. Sci Rep. 2016;6:26621.
64. Aslam Z, Yasir M, Yoon HS, Jeon CO, Chung YR. Diversity of the bacterial community in the rice rhizosphere managed under conventional and no-tillage practices. J Microbiol. 2013;51:747–56.
65. Imchen M, Kumavath R, Vaz ABM, Góes-Neto A, Barh D, Ghosh P, et al. 16S rRNA gene amplicon based metagenomic signatures of rhizobiome community in rice field during various growth stages. Front Microbiol [Internet]. Frontiers; 2019 [cited 2021 May 12];10. https://www.frontiersin.org/articles/10.3389/fmicb.2019.02103/full.
66. Yeh Y-F, Chang SC, Kuo H-W, Tong C-G, Yu S-M, Ho T-HD. A metagenomic approach for the identification and cloning of an endoglucanase from rice straw compost. Gene. 2013;519:360–6.

RNA-Induced Gene Silencing

24

Piyali Goswami

Abstract

RNA-mediated gene silencing using small RNAs, which developed as a natural defence mechanism against viruses, has become an important tool in functional genomics. It employs homology-dependent double-stranded RNA binding to the target RNA, which leads to gene knockdown by either transcriptional suppression or mRNA degradation. Studies in plants and *C. elegans* have shown that RNA silencing is quite effective in knocking down gene expression. RNA silencing, if employed successfully, can prove highly beneficial in therapeutics in curing numerous diseases. In this study, we have initially discussed about the types of small RNAs, their mechanism of inhibition and finally discussing the applications of RNAi.

Keywords

RNAi · siRNA · miRNA · DICER · dsRNA · Guide RNA

Abbreviations

Ago-2	Argonaute 2
DGCR8	DiGeorge syndrome critical region 8
dsRNA	Double-stranded RNA
MCNPs	Membrane/core nanoparticles
mRNA	Messenger RNA
miRNA	microRNA

P. Goswami (✉)
Department of Biotechnology, IIT Kharagpur, Kharagpur, West Bengal, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_24

nt	Nucleotide
PAZ	Piwi/Argonaute/Zwille
pi-RNA	Piwi interacting RNA
pre-miRNA	Precursor miRNA
pri-miRNA	Primary miRNA
PTGS	Post-transcriptional gene silencing
rasiRNA	Repeat associated siRNA
RDRP	RNA-dependent RNA polymerase
RISC	RNA-induced silencing complex
RNAi	RNA interference
scnRNA	Small-scan RNA
shRNA	Short hairpin RNA
siRNA	Small interfering RNA
SNALP	Stable nuclei acid-lipid particles
tasiRNA	*Trans*-acting siRNA
TGS	Transcriptional gene silencing
TRBP	Transactivating response RNA binding protein

24.1 Introduction

Gene silencing or gene knockdown is the process of inhibition or suppression of expression of a gene. Unlike gene knockout, gene knockdown is never 100% efficient as it does not lead to complete removal of the gene. But it is a much more preferred technique as it helps to study the role of various essential genes which are crucial to cell survival. Over the years, gene silencing has helped researchers to understand the numerous metabolic pathways in plants and animals [1, 2]. It has several advantages in plants and animals. In animals, it has a lot of therapeutic applications [3]. In plants, it has helped in the enrichment of food quality, increased shelf-life, restored fertility, and improved bacterial and viral resistances of plants [4–8]. Gene silencing can be carried out at the meiotic, transcriptional, and post-transcriptional levels. At the meiotic level, gene silencing mainly occurs by transvection. At the transcriptional level, it occurs through genomic imprinting, paramutation, transposon silencing, transgene silencing, position effect, and RNA-directed DNA methylation. At the post-transcriptional level, it mainly occurs by RNA interference (RNAi) and nonsense-mediated decay. RNA-mediated gene silencing as discussed above is mediated either at the transcriptional level (transcriptional gene silencing/TGS) by suppressing gene transcription by methylation or at the post-transcriptional level (post-transcriptional gene silencing/ PTGS) by RNA interference (RNAi) where dsRNA-mediated target mRNA degradation occurs [9, 10]. In our study, we have restricted our discussion to RNAi-mediated gene silencing.

24.2 History

Gene silencing is a spontaneous process carried out by the cell to control gene expression during the course of development. The efforts to silence the gene began in the early 1990s in various organisms like Petunia, *C. elegans,* which was termed differently for different organisms like co-suppression in plants, quelling in fungi, and RNA interference (RNAi) in nematodes [11, 12]. Andrew Fire and Craig Mellow in 1998 discovered the process of dsRNA-mediated RNA interference which resulted in effective gene silencing for which they received the Nobel Prize in 2006 [10]. Later, this dsRNA-mediated gene silencing approach was used in several organisms to inhibit gene expression and understand the functions of various genes.

24.3 RNA Interference

RNA interference (RNAi) is a naturally occurring process in the cell, wherein a small (19–31 nt long) non-coding dsRNA fragment with high sequence specificity is used to bind the mRNA sequence. The dsRNA binding leads to degradation of the mRNA and preventing its translation [13, 14]. The RNA-induced silencing process has emerged as a self-defence mechanism to protect the cell. It inhibits viral replication and transposon mobilization [15]. RNAi is extensively used in therapeutics nowadays, which will be described in detail in later parts of the chapter. The process of RNAi is conserved in most of the eukaryotes. RNAi can be carried out by several small RNAs like microRNA (miRNA), small interfering RNA (siRNA), which is central to the system. Except these, there are also other small RNAs identified like *trans*-acting siRNA (tasiRNA), small-scan RNA (scnRNA), piwi-interacting RNA (piRNA), and its subspecies repeat-associated small interfering RNA (rasiRNA). The major types of small RNAs involved in the RNAi process are highlighted in Table 24.1.

Table 24.1 The major types of small RNAs involved in the RNAi process

Major types	Description	Length (nt)	Function
siRNA	siRNA are short RNA fragments exogenous in origin produced after processing of long dsRNA fragments by DICER and has a 3′ overhang of 2 nt	19–21	mRNA degradation by cleavage
miRNA	Small endogenous RNA fragments derived from hairpin-structured precursors	21–25	Translation inhibition or methylation of DNA
piRNA	Most abundant of small RNA in animal cells derived from repetitive DNA and transposons. The synthesis of piRNAs are not well understood	25–31	Transposon processing in germ cells

24.4 Components of RNAi

RNAi requires several components to carry out the entire process. Some of the components which are essential to conduct the process are described in detail below.

24.4.1 DICER

Dicer is a ribonuclease that belongs to the RNase III family protein. Dicer processes dsRNA and pre-miRNA into siRNA and miRNA, respectively [16]. It produces short dsRNA fragments of 20–22 bp in size with overhangs of 2 bp in the 3′ end [17, 18]. It is encoded by DICER1 gene and is responsible for the activation of RNAi process. The first crystal structure of DICER was identified in protozoan parasite *Giardia intestinalis* [19]. Human Dicer is made up of four domains namely: amino terminal helicase domain, dual RNAse III motifs, a dsRNA-binding domain, and PAZ (Piwi/Argonaute/Zwille) domains. The PAZ domain is responsible for binding the 2 nt at the 3′ end, and the RNase III catalytic domain initiates the cleavage of the dsRNA strands [19]. The dsRNA-binding domain is also known for binding the dsRNA, and the helicase domain is known for processing of long substrates. DICER varies largely in size in different species due to the presence of different domains though RNAse III and PAZ domain is a common feature in all.

24.4.2 Guide RNA and RNA-Induced Silencing Complex

The RISC is a riboprotein complex consisting of RNA and protein. After the generation of the small siRNA/ miRNA fragments by DICER, it is loaded onto the RNA-induced silencing complex (RISC) by transactivating response RNA binding protein (TRBP). The TRBP has three double-stranded RNA binding domains which bind the siRNA/ miRNA generated by DICER and transfers it to the Ago 2 of the RISC [20]. The duplex RNA is unwound by RISC, and only one of the strands acts as a “guide strand” which binds the Argonaute protein and directs the Argonaute in a homology-dependent manner for endonucleolytic cleavage of the targeted mRNA [21]. The other passenger strand of siRNA/ miRNA is degraded during the course of RISC activation [22].

One of the essential components of the RISC protein complex belongs to the Argonaute family proteins. Human Argonaute protein has eight family members, among which Argonaute 2 (Ago 2) is important as it is involved in targeted mRNA cleavage. Ago 2 is also known as the catalytic centre of RISC [23, 24]. It is a ≈130 kDa basic protein with characteristic features like a central PAZ domain and a C-terminal PIWI domain [25].

24.4.3 RNA-Dependent RNA Polymerase

RNA-Dependent RNA Polymerase (RDRP) is supposed to a play very important role in triggering and enhancing the RNAi signal by increasing the secondary siRNA production. RDRP genes have been identified in plants, fungi, and *C. elegans* but not in *Drosophila* or human genome so far [23, 24, 26, 27]. RDRP is responsible for the production of dsRNA from single-stranded transcripts either by de novo synthesis of the second strand or by using siRNAs as primers. These RNAs are finally the targets for sequence-specific RNA degradation.

24.5 RNAi Mechanism

As discussed above, the RNAi process requires several proteins to conduct the complete process. RNAi can be initiated by dsRNA from both exogenous (virus infection or laboratory manipulation) and endogenous origin (pre-microRNA). In case of exogenous dsRNA, the dsRNA is directly transported to the cytoplasm where it is cleaved to short fragments by DICER, whereas in case of endogenous dsRNA, the primary transcript is first processed to pre-miRNA in the nucleus which is then transported to the cytoplasm. But exogenous dsRNAs like shRNA can be integrated into the genome initially and then can be transferred to the cytoplasm after further processing. In almost all organisms studied so far, siRNAs and miRNAs are responsible for silencing gene expression by RNAi majorly. The detailed mechanism used by siRNA and miRNA is explained in detail below.

24.5.1 siRNA-Mediated Silencing

Small interfering RNA (siRNA) also known as short interfering RNA, or silencing RNA, are small RNA fragments of 21–22 bp in size with a 2 nt overhang at the 3′ end which is responsible for shutting down the gene expression by degrading the mRNA and inhibiting translation. siRNAs generate from either long dsRNA precursors (like complementary RNAs, shRNA) through transgene incorporation, viral infection, active transposons or can be synthesized chemically or biochemically. The in vitro synthesized siRNA/dsRNA is introduced into the cells through various vectors which are described later in this chapter. Once the dsRNA enters into the cell, it is capable of activating DICER. DICER is an RNase III endonuclease which cleaves the long dsRNA precursors into small siRNA fragments of 21–22 bp with 2 nt overhangs at the 3′ end (Fig. 24.1) which is the recognizing feature for RISC [17]. siRNA can also be generated by cleaving shRNA. shRNA is also known as short hairpin RNA. They are artificial RNA molecules that are made up of paired sense and antisense strands making up the stem region and unpaired nucleotides making up the loop region giving it a hairpin structure (Fig. 24.2). shRNA is introduced into the nucleus of the cell either through viral or bacterial vectors where it stably integrates into the host genome, but also has some side effects

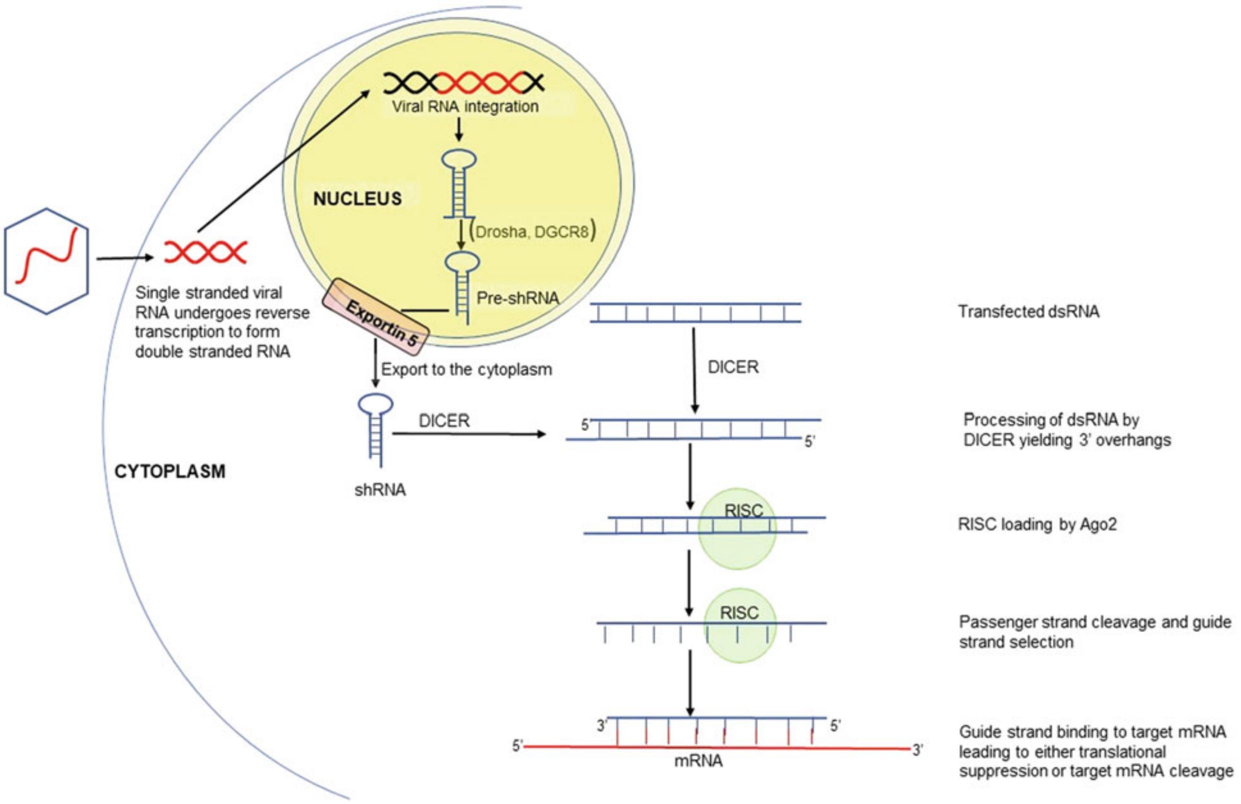

Viral RNA integration
NUCLEUS
(Drosha, DGCR8)
Pre-shRNA
Exportin 5
Single stranded viral RNA undergoes reverse transcription to form double stranded RNA
Export to the cytoplasm
DICER
shRNA
CYTOPLASM
DICER
5'
5'
RISC
RISC
3'
5'
5'
3'
mRNA
Transfected dsRNA
Processing of dsRNA by DICER yielding 3' overhangs
RISC loading by Ago2
Passenger strand cleavage and guide strand selection
Guide strand binding to target mRNA leading to either translational suppression or target mRNA cleavage

Fig. 24.1 Processing and working of siRNA in the cell

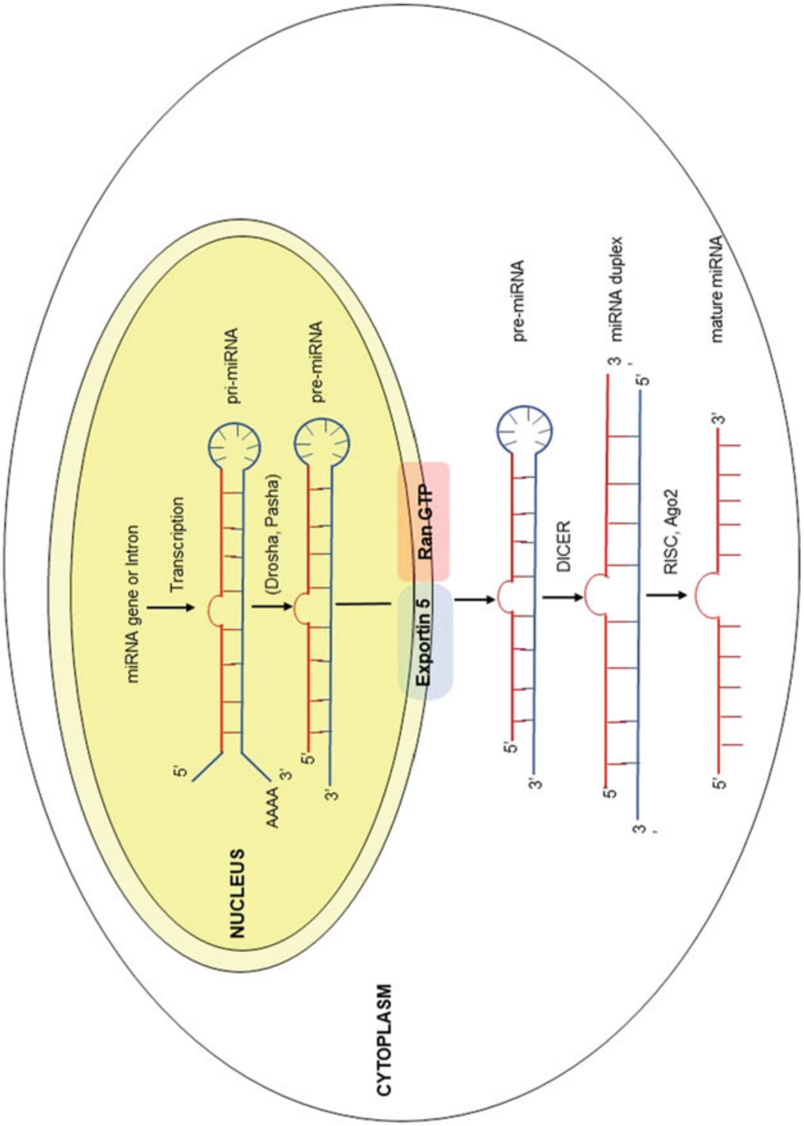


Fig. 24.2 Biogenesis of miRNA

accompanying it as expression vectors are used in the transfer. Once the shRNA integrates into the genome, it is transcribed in the nucleus by either RNA polymerase II or RNA polymerase III [17]. These shRNA precursors are processed by a nucleus RNase III endonuclease, DROSHA, and its dsRNA-binding partner DGCR8 to pre-shRNA which is then transported to the cytoplasm by Exportin-5 where it is further processed by DICER and TRBP/PACT to remove the hairpin and form siRNA. So, both the processes (dsRNA and shRNA processing) converge in this step where both are processed by DICER to form siRNA. The siRNA is then loaded into RISC, which is composed of Ago-2, DICER, and TRBP. The helicase property of DICER helps it to unwind the double-stranded siRNA. Ago-2 then cleaves the passenger (sense) strand by its RNase-H like activity leaving the guide (anti-sense) strand intact, activating the RISC [28]. The guide strand then directs the RISC to the mRNA target for cleavage by Ago-2 [19, 29]. Ago-2 interacts with the guide strand through the MID and PIWI domains at the 5′ end and PAZ domain at the 3′ domain. The guide strand recognizes its target by intermolecular base pairing. The difference in thermodynamic stability serves as the basis of selection of the guide strand and the passenger strand [30, 31]. The guide strand must have low melting temperature and low duplex stability toward the 5′ end, whereas the passenger strand has high melting temperature and high duplex stability, which favours its degradation. If the guide strand has perfect complementarity with the target mRNA, then the mRNA is cleaved off, but if there is not perfect complementarity, then the translation is repressed. Both these cases lead to a reduction in gene expression or gene silencing. The entire process is illustrated in Fig. 24.1.

24.5.2 miRNA-Mediated Silencing

miRNAs are small endogenous dsRNA molecules of 21–22 nucleotides. Most of the miRNA lies in the introns and is derived from the RNA transcripts with a hairpin structure. miRNAs are mostly transcribed by RNA pol II and RNA pol III into a long RNA transcript known as primary miRNA (pri-miRNA). A pri-miRNA may itself have one to six miRNA precursors. DROSHA, an RNase III endonuclease, cleaves the pri-miRNA in the nucleus with the aid of a dsRNA-binding protein DGCR8 (DiGeorge Syndrome Critical Region 8) also known as PASHA into a 60–70 nucleotide hairpin structure known as precursor-miRNA (pre-miRNA) which has a 2-nucleotide overhang at the 3′ end [32]. Exportin 5 then translocates the pre-miRNA from the nucleus into the cytoplasm where it is further processed by DICER to form the mature miRNA [28, 33, 34]. From here, the miRNA shares the common downstream processing machinery as the siRNA. Only one of the strands of the miRNA is loaded into the RISC, making it an active RISC where it interacts with the target mRNA. The degree of complementarity decides whether miRNA will suppress the gene expression by cleaving mRNA or inhibiting translation. In plants, miRNA is able to silence the gene by mRNA degradation as it is completely complimentary to its mRNA target. But in animals, miRNA primarily silences the gene by inhibiting translation as it is not completely complementary, unlike plant cells. It is only able to recognize a short sequence stretch about 6–8 nucleotides known as the "seed region" at the 5′ end of the mRNA [35]. This region has to be perfectly complementary even though the complete mRNA sequence does not match. The biogenesis of miRNA is illustrated in Fig. 24.2.

24.6 Delivery Methods for siRNA

The in vitro synthesized siRNA can be delivered to the cells through a variety of methods which is discussed below.

24.6.1 Viral-Mediated Delivery

Viral vectors are the widely used vectors for the transfer of siRNA due to their efficiency in delivery. Adenovirus, adeno-associated virus, retrovirus, and lentivirus are the most used ones [36–38]. Transfer through lentivirus is the most preferred among the other viruses as it has a comparatively low level of immunogenicity among the others. Viral vectors have the tendency to integrate stably into the genome and helps in long-term gene knockdown. This quality makes it the preferable mode of transfer as compared to the other vectors. But there are also several drawbacks of the viral vectors. Among one of the major drawbacks is biosafety for which other modes of transfer of siRNA into the cells are being considered [39].

24.6.2 Non-viral-Mediated Delivery

Due to the biosafety concerns, the non-viral mode of delivery has gained importance. It can be either transferred through electroporation or through different non-viral vectors using transfection. The non-viral vectors are majorly classified into lipid-based vectors, non-lipid organic-based vectors, and non-lipid inorganic-based vectors. Lipoplexes, lipopolyplexes, stable nuclei acid-lipid particles (SNALPs), and membrane/core nanoparticles (MCNPs) are the widely used lipid vectors [40, 41]. Among the non-lipid organic-based vectors, chitosan, dendrimers, polyethylenimines are profoundly used [42–44]. Gold nanoparticles, superparamagnetic iron oxide nanoparticles, silica-based nanoparticles, semiconductor quantum dots are among the non-lipid inorganic-based vectors mostly used for the transfer of siRNA [45–47]. The non-viral vectors should offer the basic properties like non-toxicity, biocompatibility, biodegradability, stability, and protection to siRNA. Non-viral vectors also have some drawbacks, for example, they are not highly efficient in transferring siRNA as compared to the viral vector.

24.7 Applications of RNAi

Over the years, RNAi has gained a lot of importance in various levels as listed below.

- RNAi has been exploited to understand the function and the behaviour of various genes in experimental biology. It has been used in gene mapping in gene annotation and also to introduce programmed genome arrangements [48].
- RNAi plays a substantial role in the development of multicellular organisms. It has been seen to control the various gene regulation pathways and germline development and stem cell maintenance.
- RNAi approach has been used in the treatment of many viruses like HIV, HPV, hepatitis A, B, and C virus, influenza virus, and many others. They inhibit the viral infection by targeting the virus and the host genes required for viral replication and entry into the cells [49].
- RNAi is also used widely in the treatment of many diseases like cancer, degenerative macular disease, hereditary disorders like Huntington disease, neurodegenerative disorders like Alzheimer's, Parkinson's, and polyglutamine disease [50, 51]. RNAi is seen to be very effective as it directly targets the mutant gene.

24.8 Conclusion

RNAi, which developed as a natural defence mechanism against RNA viruses in plants, is now widely used in functional genomics to understand the function of various genes. It is a highly potent method for knocking down gene expression. In recent years, it has also gained much importance in the therapeutic industry as it has been found effective against several diseases. But there are also drawbacks associated with this process like toxicity, efficacy, and off-target effects. Improvement in the delivery strategies for RNAi can ensure its wide usage in therapeutics on a regular basis.

Conflict of Interest None

References

1. George GM, Bauer R, Blennow A, Kossmann J, Lloyd JR. Virus-induced multiple gene silencing to study redundant metabolic pathways in plants: silencing the starch degradation pathway in Nicotiana benthamiana. Biotechnol J. 2012;7(7):884–90.
2. Hu M, Ni Q, Yang Y, Luo J. RNAi-based gene therapy for blood genetic diseases. RNA Interference, Ibrokhim Y. Abdurakhmonov, IntechOpen. 2016 Apr 6. https://doi.org/10.5772/61644. Available from: https://www.intechopen.com/chapters/49442.
3. Deng Y, Wang CC, Choy KW, Du Q, Chen J, Wang Q, et al. Therapeutic potentials of gene silencing by RNA interference: principles, challenges, and new strategies. Gene. 2014;538(2):217–27.

4. Escobar MA, Civerolo EL, Summerfelt KR, Dandekar AM. RNAi-mediated oncogene silencing confers resistance to crown gall tumorigenesis. Proc Natl Acad Sci. 2001;98(23):13437–42.
5. Niggeweg R, Michael AJ, Martin C. Engineering plants with increased levels of the antioxidant chlorogenic acid. Nat Biotechnol. 2004;22(6):746–54.
6. Xiong A-S, Yao Q-H, Peng R-H, Li X, Han P-L, Fan H-Q. Different effects on ACC oxidase gene silencing triggered by RNA interference in transgenic tomato. Plant Cell Rep. 2005;23 (9):639–46.
7. Nizampatnam NR, Kumar VD. Intron hairpin and transitive RNAi mediated silencing of orfH522 transcripts restores male fertility in transgenic male sterile tobacco plants expressing orfH522. Plant Mol Biol. 2011;76(6):557–73.
8. Obembe OO, Popoola JO, Leelavathi S, Reddy SV. Advances in plant molecular farming. Biotechnol Adv. 2011;29(2):210–22.
9. Fire A, Xu S, Montgomery MK, Kostas SA, Driver SE, Mello CC. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature. 1998;391 (6669):806–11.
10. Montgomery MK, Xu S, Fire A. RNA as a target of double-stranded RNA-mediated genetic interference in Caenorhabditis elegans. Proc Natl Acad Sci. 1998;95(26):15502–7.
11. Napoli C, Lemieux C, Jorgensen R. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. Plant Cell. 1990;2 (4):279–89.
12. Par-1, a gene required for establishing polarity in C. elegans embryos, encodes a putative Ser/Thr kinase that is asymmetrically distributed. Cell. 1995;81(4):611–20.
13. Metzlaff M, O'dell M, Cluster PD, Flavell RB. RNA-mediated RNA degradation and chalcone synthase a silencing in petunia. Cell. 1997;88(6):845–54.
14. Ngo H, Tschudi C, Gull K, Ullu E. Double-stranded RNA induces mRNA degradation in Trypanosoma brucei. Proc Natl Acad Sci. 1998;95(25):14687–92.
15. Waterhouse PM, Wang M-B, Lough T. Gene silencing as an adaptive defence against viruses. Nature. 2001;411(6839):834–42.
16. Lee YS, Nakahara K, Pham JW, Kim K, He Z, Sontheimer EJ, et al. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. Cell. 2004;117 (1):69–81.
17. Vermeulen A, Behlen L, Reynolds A, Wolfson A, Marshall WS, Karpilow JON, et al. The contributions of dsRNA structure to Dicer specificity and efficiency. RNA. 2005;11(5):674–82.
18. Ando Y, Maida Y, Morinaga A, Burroughs AM, Kimura R, Chiba J, et al. Two-step cleavage of hairpin RNA with 5' overhangs by human DICER. BMC Mol Biol. 2011;12(1):1–12.
19. MacRae IJ, Zhou K, Li F, Repic A, Brooks AN, Cande WZ, et al. Structural basis for double-stranded RNA processing by Dicer. Science. 2006;311(5758):195–8.
20. Nicholson AW. Ribonuclease III mechanisms of double-stranded RNA cleavage. Wiley Interdiscip Rev RNA. 2014;5(1):31–48.
21. Chendrimada TP, Gregory RI, Kumaraswamy E, Norman J, Cooch N, Nishikura K, et al. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. Nature. 2005;436(7051):740–4.
22. Jinek M, Doudna JA. A three-dimensional view of the molecular machinery of RNA interference. Nature. 2009;457(7228):405–12.
23. Smardon A, Spoerke JM, Stacey SC, Klein ME, Mackin N, Maine EM. EGO-1 is related to RNA-directed RNA polymerase and functions in germ-line development and RNA interference in C. elegans. Curr Biol. 2000;10(4):169–78.
24. Agrawal N, Dasaradhi PVN, Mohmmed A, Malhotra P, Bhatnagar RK, Mukherjee SK. RNA interference: biology, mechanism, and applications. Microbiol Mol Biol Rev. 2003;67 (4):657–85.
25. Cerutti L, Mian N, Bateman A. Domains in gene silencing and cell differentiation proteins: the novel PAZ domain and redefinition of the Piwi domain. Trends Biochem Sci. 2000;25 (10):481–2.

26. Cogoni C, Macino G. Posttranscriptional gene silencing in Neurospora by a RecQ DNA helicase. Science. 1999;286(5448):2342–4.
27. Dalmay T, Hamilton A, Rudd S, Angell S, Baulcombe DC. An RNA-dependent RNA polymerase gene in Arabidopsis is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. Cell. 2000;101(5):543–53.
28. Lund E, Güttinger S, Calado A, Dahlberg JE, Kutay U. Nuclear export of MicroRNA precursors. Science. 2004;303(5654):95–8.
29. Leuschner PJ, Ameres SL, Kueng S, Martinez J. Cleavage of the siRNA passenger strand during RISC assembly in human cells. EMBO Rep. 2006;7(3):314–20.
30. Schwarz DS, Hutvágner G, Du T, Xu Z, Aronin N, Zamore PD. Asymmetry in the assembly of the RNAi enzyme complex. Cell. 2003;115(2):199–208.
31. Khvorova A, Reynolds A, Jayasena SD. Functional siRNAs and miRNAs exhibit strand bias. Cell. 2003;115(2):209–16.
32. Denli AM, Tops BBJ, Plasterk RHA, Ketting RF, Hannon GJ. Processing of primary microRNAs by the microprocessor complex. Nature. 2004;432(7014):231–5.
33. Bernstein E, Caudy AA, Hammond SM, Hannon GJ. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature. 2001;409(6818):363–6.
34. Yi R, Qin Y, Macara IG, Cullen BR. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev. 2003;17(24):3011–6.
35. Bartel DP. MicroRNAs: target recognition and regulatory functions. Cell. 2009;136(2):215–33.
36. Schiedner G, Morral N, Parks RJ, Wu Y, Koopmans SC, Langston C, et al. Genomic DNA transfer with a high-capacity adenovirus vector results in improved in vivo gene expression and decreased toxicity. Nat Genet. 1998;18(2):180–3.
37. Vigna E, Naldini L. Lentiviral vectors: excellent tools for experimental gene transfer and promising candidates for gene therapy. J Gene Med. 2000;2(5):308–16.
38. Schambach A, Morgan M. Retroviral vectors for Cancer gene therapy. In: Walther W, editor. Current strategies in cancer gene therapy [Internet]. Cham: Springer; 2016 [cited 2021 Feb 10]. p. 17–35. (Recent results in Cancer research). https://doi.org/10.1007/978-3-319-42934-2_2.
39. Raper SE, Chirmule N, Lee FS, Wivel NA, Bagg A, Gao G, et al. Fatal systemic inflammatory response syndrome in a ornithine transcarbamylase deficient patient following adenoviral gene transfer. Mol Genet Metab. 2003;80(1):148–58.
40. Xu Y, Szoka FC. Mechanism of DNA release from cationic liposome/DNA complexes used in cell transfection. Biochemistry. 1996;35(18):5616–23.
41. Hafez IM, Maurer N, Cullis PR. On the mechanism whereby cationic lipids promote intracellular delivery of polynucleic acids. Gene Ther. 2001;8(15):1188–96.
42. Howard KA, Rahbek UL, Liu X, Damgaard CK, Glud SZ, Andersen MØ, et al. RNA interference in vitro and in vivo using a novel chitosan/siRNA nanoparticle system. Mol Ther. 2006;14 (4):476–84.
43. Techaarpornkul S, Wongkupasert S, Opanasopit P, Apirakaramwong A, Nunthanid J, Ruktanonchai U. Chitosan-mediated siRNA delivery in vitro: effect of polymer molecular weight, concentration and salt forms. AAPS PharmSciTech. 2010;11(1):64–72.
44. Helmschrodt C, Höbel S, Schöniger S, Bauer A, Bonicelli J, Gringmuth M, et al. Polyethylenimine nanoparticle-mediated siRNA delivery to reduce α-synuclein expression in a model of Parkinson's disease. Mol Ther Nucl Acids. 2017;9:57–68.
45. Taratula O, Garbuzenko O, Savla R, Andrew Wang Y, He H, Minko T. Multifunctional nanomedicine platform for cancer specific delivery of siRNA by superparamagnetic Iron oxide nanoparticles-dendrimer complexes. Curr Drug Deliv. 2011;8(1):59–69.
46. Huschka R, Barhoumi A, Liu Q, Roth JA, Ji L, Halas NJ. Gene silencing by gold Nanoshell-mediated delivery and laser-triggered release of antisense oligonucleotide and siRNA. ACS Nano. 2012;6(9):7681–91.
47. Niu Y, Popat A, Yu M, Karmakar S, Gu W, Yu C. Recent advances in the rational design of silica-based nanoparticles for gene therapy. Ther Deliv. 2012;3(10):1217–37.

48. Clemens JC, Worby CA, Simonson-Leff N, Muda M, Maehama T, Hemmings BA, et al. Use of double-stranded RNA interference in Drosophila cell lines to dissect signal transduction pathways. PNAS. 2000;97(12):6499–503.
49. McManus MT, Sharp PA. Gene silencing in mammals by small interfering RNAs. Nat Rev Genet. 2002;3(10):737–47.
50. Boudreau RL, Davidson BL. RNAi Therapy for Neurodegenerative Diseases. In: Current topics in developmental biology [Internet]. Academic Press; 2006 [cited 2021 Feb 10]. p. 73–92. https://www.sciencedirect.com/science/article/pii/S0070215306750037.
51. Rao DD, Vorhies JS, Senzer N, Nemunaitis J. siRNA vs. shRNA: similarities and differences. Adv Drug Deliv Rev. 2009;61(9):746–59.

Single-Cell RNA Sequencing Technologies 25

Manoj Kumar Gupta, Gayatri Gouda, Ravindra Donde,
S. Sabarinathan, Piyali Goswami, Goutam Kumar Dash, N. Rajesh,
Pallabi Pati, Sushil Kumar Rathode, Ramakrishna Vadde, and
Lambodar Behera

Abstract

All cellular structures are heterogeneous. Thus, investigating true cell heterogeneity is highly required to further understand cellular connectivity and accountability within a disease or normal conditions. Because of its rapidly decreasing costs, the Next-Generation (NGS) sequence is widely used to analyze various biological data. However, these approaches may fail to provide detailed insight into cells' true heterogeneity. Recently developed single-cell RNA sequencing (scRNA-seq) technology tries to tackle these bulk NGS issues by linking transcriptomic, epigenomic, proteomic, and molecular sequences to a specific cell. Thus, in this chapter, the author addresses the process involved, relative strengths, possible uses, and limitations of scRNA-seq techniques methods. Information obtained revealed that cell isolation methods may be broadly divided

M. K. Gupta · G. Gouda · R. Donde · G. K. Dash · L. Behera (✉)
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

S. Sabarinathan
Crop Improvement Division, ICAR-National Rice Research Institute, Cuttack, Odisha, India

Department of Seed Science and Technology, College of Agriculture, Odisha University of Agriculture and Technology, Bhubaneswar, Odisha, India

P. Goswami
Department of Biotechnology, IIT Kharagpur, Kharagpur, West Bengal, India

N. Rajesh · R. Vadde
Department of Biotechnology and Bioinformatics, Yogi Vemana University, Kadapa, Andhra Pradesh, India

P. Pati
District Headquarter Hospital, Ganjam, Odisha, India

S. K. Rathode
Department of Zoology, Khallikote Autonomous College, Ganjam, Odisha, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_25

into two categories centered on different principles. The first category centered on physical characteristics, while the second category mainly focuses on cell' features and primarily involves affinity approaches. After obtaining raw data, the general approach for analyzing ScRNA-Seq data are pre-processing, batch effect correction, normalization, dimensionality reduction, feature selection, cell type identification, differential expression analysis, and rebuilding cell hierarchy. scRNA-seq technologies are continuously being employed to unmask various biological processes ranging from epigenetic regulation to biomarker identification. However, scRNA-seq technologies do have difficulties like cumbersome activity and high detection costs that restrict technology promotion. Hence, there is an urgent requirement to develop more robust tools so that, in the near future, the technology for single-cell sequencing will be streamlined and is more efficient.

Keywords

Cellular heterogeneity · Machine learning · RNA-seq · Single-cell transcriptomics · single cell

Abbreviation

5mC	5-Methylcytosine
5hmC	5-Hydroxymethylcytosine
AML	Acute myeloid leukemia
CCA	Canonical correlation analysis
cDNAs	Complementary DNAs
DBC	Density-based clustering
DC	Dendritic cells
DE	Differential expression
ESCs	Embryonic stem cell
FACS	Fluorescence-activated cell sorting
GSE	Gene expression
HVG	Highly variable genes
KM	*k*-means
LCM	Laser capture microdissection
MACS	Magnetic activated cell sorting
MCS	Manual cell selection
MNNs	Mutually closest neighbors
NGS	Next-generation sequence
PCA	Principal component analysis
scRNA-seq	Single-cell RNA sequencing
t-SNE	T-distributed stochastic neighbor embedding
UMAP	Uniform manifold approximation and projection
UMIs	Unique molecular identifiers

25.1 Introduction

It is well-documented, both theoretically and experimentally, that almost all cellular structures are heterogeneous [1]. Heterogeneity can occur for numerous reasons and at several levels for improving survival as well as functionality. For instance, both single-celled as well as multicellular species employ population-level survival techniques, like bet-hedging, for achieving a higher survival rate while facing new constraints with a diverse culture [2]. In most cases, genomic, transcriptomic, proteomic, and epigenomic assessments play an essential role in investigating cellular heterogeneity across many respects. However, at one point, the level of variation will not be the same as another. Though cells within an individual have pretty similar genomes, they can produce several distinct types of cells with distinctive expression patterns by various modifications. The genome itself can be precisely reconfigured to create expanded genetic variation within particular cell groups, most prominently T- as well as B-cells, via recombining V(D)J. It has also been well established that differentiation during development enables the cellular specialization needed for intricate multicellular system work. Furthermore, complex epigenomic modifications permit numerous distinct segregation that eventually leads to the continuum of human cell heterogeneity and is also highly required for cancer formation. Thus, investigating true cell heterogeneity is highly required to further understand cellular connectivity and accountability within a disease or normal conditions [3, 4].

Because of its rapidly decreasing costs, the Next-Generation sequence (NGS) is being widely used to analyze a wide variety of biological data [5]. Several, usually tens of hundreds to billions of cells are examined at once in the framework of bulk NGS studies. This, in turn, provides the general image of a particular group of cells. However, these approaches may fail to provide detailed insight into cells' true heterogeneity. Recently developed single-cell RNA sequencing (scRNA-seq) technology tries to tackle these bulk NGS issues by linking transcriptomic, epigenomic, proteomic, and molecular sequences to a specific cell [1]. Two independent research group, namely, Eberwine et al. [6] and Brady et al. [7], pioneered the entire transcriptome sequences at the single-cell level, which extended per cell's complementary DNAs (cDNAs) along with linear amplification through exponential enhancement through PCR or in vitro transcription, respectively. Primarily, the process was proposed for commercial DNA microarray chips and was later modified for scRNA-seq [8]. In 2009, for the first time, scRNA-seq was employed for describing detailed insight into early cell production [9].

Since that research, there's been an increase in enthusiasm to unmask unicellular heterogeneity at high-resolution. Introspectively, evaluating the disparities in gene expression among single cells may also recognize uncommon populations that cannot be distinguished through the bulk cell study. For instance, the capability to find and classify uncommon cells in a population will contribute to a deeper insight into drug resistance and recurrence during cancer treatment [10]. Significant improvements in available laboratory methods and bioinformatics pipelines have also rapidly allowed investigators to deconstruct very different populations of immune cells in both disease and healthy states [11]. Moreover, scRNA-seq is also widely used to study the early development of myoblast differentiation [12] and lymphocyte fate assessment [13]. scRNA-seq, nevertheless, isn't without flaws. One of the greatest difficulties with scRNA-seq is its prices, and while it has declined dramatically over recent times, it remains a major concern during research setup and technological issues, including sensitivity [1]. Thus, in this chapter, the author addresses the process involved, relative strengths, possible uses, and limitations of scRNA-seq techniques methods.

25.2 Single-Cell Isolation

Scientists must isolate or characterize single cells before conducting a single-cell study. The purity (the proportion of the target cells obtained following isolation), effectiveness or productivity (number of cells isolated at a given time), as well as the recovery (the proportion of the target cells retrieved subsequent to isolation relative to the initial target cells number present within the sample) describe the success of cellular isolation. The existing strategies display various benefits on each of the three dimensions [14]. Centered on different principles, current cell isolation methods may be broadly divided into two categories. The first category centered on physical characteristics, like density, electrical adjustments, size and deformability, with techniques such as membrane filtration, centrifuge gradient density, and capture platforms focused on microchips. Single-cell isolation without labeling is, by far, the most desirable physical characteristics. The second category mainly focuses on cell features and primarily involves affinity approaches like strong affinity matrix (fibers, plates, beads) and cell-sorting enabled through fluorescence (Fig. 25.1) [14].

(a) Fluorescence-Activated Cell-Sorting (FACS)

FACS is the most advanced and user-friendly strategy for identifying as well as distinguishing diverse cell types in heterogeneous communities depending on their size, fluorescence, and granularity. FACS allows a multi-parameter qualitative analysis of individual cells [15]. Till isolation, a suspension is rendered, and fluorescent probes specifically target cells. "Fluorophore-conjugated monoclonal antibodies" are the most commonly employed sample that identifies unique target cell's markers. Since cell suspension moves via cytometry, every individual cell is subjected to a laser that permits fluorescence sensors to detect cells on the basis of the characteristics selected. The device charges a droplet (negative or positive) describing a cell of importance and the electrostatic deflection environment allows the aggregation of the charged droplets in the proper collection tubes for subsequent investigation. While FACS has been commonly used to separate highly filtered cell populations, FACS may also be used for sorting single cells [16]. BD cell-sorting systems, like the *BD FACSAria™ III cell sorter*, can isolate single interest cells from millions of cells utilizing nearly 18 surface markers [14]. However, while FACS is commonly used in basic and clinical science, several restricted drawbacks remain. First, FACS needs several cells (>10,000) to be suspended. Thus, the single cells cannot be isolated from a small cell population. Second, the fast movement

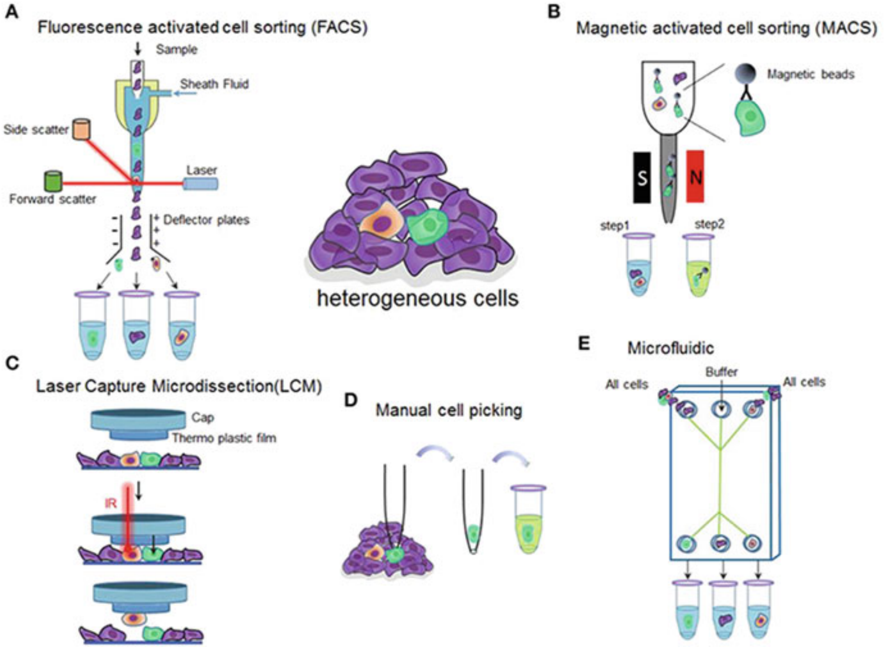


Fig. 25.1 Overview of single-cell isolation approaches. (Adapted from [14])

inside the machine, as well as non-specific fluorescent particles, can damage the vitality of filtered cells, resulting in a loss of isolation. Cells or cell cultures must also be exposed to stimulation procedures as well as processed prior FACS examination in a different setting [14].

(b) Magnetic Activated Cell-Sorting (MACS)

MACS is yet another widely employed method of passive isolation for separating various forms of cells according to their differentiation cluster. MACS has been reported to differentiate unique cell populations having a purity of more than 90% [17]. The MACS comprises enzymes, antibodies, strepavidins, or lectins coupled with magnetic beads for attaching the target cell to particular proteins. If an incoming magnetic field senses a mixed cell population, magnetic beads are triggered, and the labeled cells polarize and other cells are swept away. After the magnetic field has been switched off, the residual cells may be obtained by elution. This approach allows the cells to be isolated by the charge in response to the specific antigens. Color magnetic beads utilize effective isolation methods to attract cells. On the contrary, if species-specific compounds are unavailable, employing negative isolation procedures that use a mixture of antibodies for covering untreated cells is a safe option. In this scenario, unlabeled and labeled cells are retained and discarded, respectively [18]. Even though MACS is reasonably straightforward and cost-effective, the MACS device's obvious drawback is its preliminary magnet separator costs, including maintenance costs. In MACS devices, the isolated cells' final purity also depends upon the affinity and specificity of the antibodies used for target cells. It also relies on capturing non-specific cells. Non-specific contamination can be triggered through adsorption of background cells to the collection system or by their intrusion into the broad abundance of magnetic particles used to label unusual large cells [14]. Another downside of MACS is that it can only employ cell surface molecules for differentiating living cells. Moreover, owing to immunomagnetic strategies that can only segregate cells into negative and positive populations, MACS is far more constrained than FACS. Using MACS, high and low molecule expression cannot be distinguished [14].

(c) Laser Capture Microdissection (LCM)

LCM is an innovative technology that uses a microscope slide for separating pure cell communities from mostly stable tissue samples [19]. It can monitor as well as capture interested cells effectively and reliably by taking full advantage of modern molecular analytical technologies, together with proteomics, microarrays, and PCRs [20]. LCM's underlying idea begins with the visualization of the interested cells via an inverted microscope. Subsequently, it permits a fixed location, a short time span, and a concentrated laser pulse for melting the thin translucent thermoplastic layer onto a cap over the desired cells. The film melts and integrates with the cells concerned. When removing the film, target cells bind strongly to the film, whereas the tissue stays behind. To end, transfer the cells into a buffer tube for numerous downstream analyses [21]. The major benefit of LCM is its pace and its flexibility [22]. LCM offers a rapid as well as an accurate approach toward obtaining pure target cell populations through microscopic visualization from a diverse range of cell/organ preparations [23]. Traditional molecular analysis strategies entail tissue separation, which may lead to intrinsic issues of contamination and reduce the selectivity and reliability of molecular studies [14].

On the contrary, LCM is a no-touch approach that doesn't really kill surrounding tissue after preliminary microdissection. All captured, including the morphology of residual tissue is exceptionally well maintained, and the probability of tissues' degradation decreases [20]. The residual tissue on the diaphragm can also be accessed entirely after the cells have been removed, thereby enabling comparative molecular inspection of neighboring cells. The key criteria for successful LCM are to correctly classify cell subpopulations or individual cells within a complex tissue. However, the main disadvantage is the need for a visual microscopic examination of morphological features to classify cells of interest that, in turn, demand the involvement of a cell classification pathologist, cytologist, or technician [20]. Another major drawback is that there is no coverslip in the microdissected tissue portion. Covering slippage will preclude physical access to the tissue's surface, which is essential during the new microdissection process. The dried portion of tissues has a refractive consistency, which can mask cellular information at higher magnification. LCM also incorporates many technological artifacts, including cell slicing mostly during the processing of tissue parts and RNA/DNA damage from laser-cutting energy [24].

(d) Manual Cell Selection (MCS)
Manual cell selection is another easy, comfortable, & effective way of isolating individual cells. Like LCM, MCS is often made of an inverted microscope coupled with motorized mechanical-level micropipettes. Under the microscope, each isolated cell may be photographed as well as analyzed, allowing impartial separation. Compared to LCM, which always primarily isolates individual cells from fixed tissue parts, micro-manipulation plays a significant role in isolating embryo cells or live culture. MCS can be done effectively using a patch-clamp device in an electrophysiology lab. However, its performance is restricted, and highly trained specialists are needed to work since its effectiveness is restricted when complicated adjustments are identified [14].

(e) Microfluidics
Microfluidics is a powerful and innovative tool for examining the intrinsic cellular structures' complexity, offering reliable fluid control, relatively low intake, miniaturization of instruments, low analytical expense, and simple handling of nanoliter volumes [25]. Cells could be separated through a microfluidic chip via four approaches, namely, "cell-affinity chromatography based microfluidic, physical characteristics of cell-based microfluidic separation, immunomagnetic beads based microfluidic separation, and separation methods based on differences between dielectric properties of various cell types" [14]. Microfluidics can typically be paired with several other isolation approaches, like filtration, sedimentation, or related technologies, such as FACS and MACS. Numerous experiments, as well as applications involving microfluidic systems, have been published in recent years, including single-cell evaluation, cancer study, stem-cell discovery, microbiology, drug discovery, and screening [24]. Microfluidic chips simultaneously have the capability for applications in DNA sequence, protein analysis, cells handling, and cell composition analysis [14]. Besides, the applications of microfluidic technology in researching heterogeneity and variability among single-cell genomes have increased these days, ranging from cancer biology to environmental microbiology as well as neurobiology [14].

25.3 ScRNA-Seq Analysis Approaches

To date, a variety of scRNA-seq approaches for single-cell transcriptomic studies have been proposed. Tang et al. [9] published the first scRNA-seq tool, and subsequently, several more scRNA-seq methods were established. The scRNA-seq approaches may vary in at least one of the following parameters: (1) transcript coverage; (2) cell lysis; (3) strand selectivity; (4) amplification; (5) isolation of cell; (6) reverse transcript; and (7) UMIs (unique molecular markers, molecular tags that can be used for detecting and quantifying the unique transcripts). One distinct dissimilarity among these scRNA-seq approaches is that few can provide

Table 25.1 Widely used scRNA-seq technologies. (Adapted from [27])

Transcript coverage	UMI possibility	Strand specific	Methods	References
Nearly full-length	No	No	Tang method	[9]
Full-length	No	No	Quartz-Seq	[28]
			Smait-seq	[29]
			Smait-seq2	[30]
			SUPeR-seq	[31]
Full-length	Yes	Yes	MATQ-seq	[32]
5′-only	Yes	Yes	STRT-seq and STRT/C1	[33, 34]
3′-only	Yes	Yes	CEL-seq	[35]
			CEL-seq2	[36]
			Chromium	[37]
			Cyto-Seq	[38]
			DroNC-sea	[39]
			Drop-seq	[40]
			InDrop	[41]
			MARS-seq	[42]
			sci-RNA-seq	[43]
			Seq-Well	[44]
			SPLiT-seq	[26]
			Quartz-Seq2	[45]

either full-length or almost full-length transcript sequence data (e.g., MATQ-seq, Smart-seq2, and SUPeR-seq), while others can only collect and sequence the 5′-end (e.g., STRT-seq) or 3′-end (e.g., SPLiT-seq and Drop-seq) [26] of the transcripts [27]. Earlier studies have reported that Smart-seq2 identifies more expressed genes in comparison to other technologies, like CEL-seq2, MARS-seq, Smart-seq, and Drop-seq. Sheng et al. (2019) reported a full-length MATQ-seq transcript sequencing method that can identify low-abundance genes more precisely, and it outperforms Smart-seq2 (Table 25.1).

The full-length scRNA-seq techniques give unique advantages over 3′ end or 5′ end counting approaches, especially during detection of allelic expression and RNA editing due to their dominance in the transcript. In addition, the full-length scRNA-seq approaches may be more potent than just a 3′ or 5′ sequencing system toward identifying any low-expressed genes/transcripts [46]. Importantly, droplet-based technology could typically produce lager cellular outputs and lower sequence cost per cell relative to the full-scRNA-seq script, e.g., Drop-sq [40], Chromium [37], and InDrop [41]. Thus, Drop-sq protocols are ideal for producing massive cell numbers to classify specific tissue or tumor sample cell subpopulations. It is pertinent to note that some ScRNA-seq technologies, e.g., MATQ-seq [32] & SUPeR-seq [31], capture polyA+ and polyA-RNAs [32]. These procedures are incredibly helpful for sequencing circular RNAs (circRNAs) and long RNAs (lncRNAs). A variety of experiments have revealed that lncRNAs and circRNAs play a central role in a wide range of biological cells processes and may serve as essential biomarkers for various diseases, including cancer, which in turn aid scRNA-seq approaches to investigate the mechanisms of gene expression more precisely [27].

The scRNA-seq protocols experience greater technological variations relative to conventional Bulk RNA-seq technology. Spike-ins, like "External RNA Control Consortium controls" [47] and UMIs, have been commonly used in the subsequent scRNA-seq analysis for measuring technological differences across different cells. The RNA spike-ins are RNA transcripts (having acknowledged sequences & quantities) employed for measuring RNA hybridization assay that could approximate absolute molecular amounts. Interestingly, the underlying protocol variations, e.g., ERCC and UMIs, do not adhere to all scRNA-seq technologies. In techniques like SUPeR-seq and Smart-seq2, Spike-ins are used but are not consistent with InDrop methods [40]. As a result, users can select an appropriate scRNA-seq approach based on technological characteristics and benefit, sequencing the number of cells and costs.

25.4 Computational Approaches for Analyzing scRNA-seq Data

While several laboratories are becoming more open to experimental methods for scRNA-seq, computational pipelines that handle raw data files still remain limited. Some enterprises offer software applications like 10× genomics & fluidigm, but that's still in its development phase, and gold-standard tools are yet to be invented. In this chapter, we will address the general computational pipelines that are commonly employed for evaluating scRNA-seq data [8]. After obtaining raw data, ScRNA-Seq analysis approaches mainly comprise pre-processing, batch effect correction, normalization, dimensionality reduction, feature selection, cell type identification, differential expression analysis, rebuilding cell hierarchy, and compositional analysis (Fig. 25.2).

25.4.1 Quality Control

scRNA-seq is a lossy technology and what induces various types of failure is not well known. Actually, this implies the first step toward quality management after the acquisition of readings from a scRNA-seq experiment. Reads are analyzed similarly to data obtained from the RNA-seq experiment, which subsequently allows quantification of expression. Hence, evaluation of both raw data (which can be carried out with bulk RNA-seq tools, like Kraken [49] or FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)) and the aligned result is very important. The cell-by-cell quality management is imperative in scRNA-seq to ensure that low-quality cells are excluded from subsequent studies. Many indicators may be used to determine cell efficiency, such as the amount of gene or reads observed, the ratio of mitochondrial-genes mapping reads (which may indicate the leaks of apoptosis cells

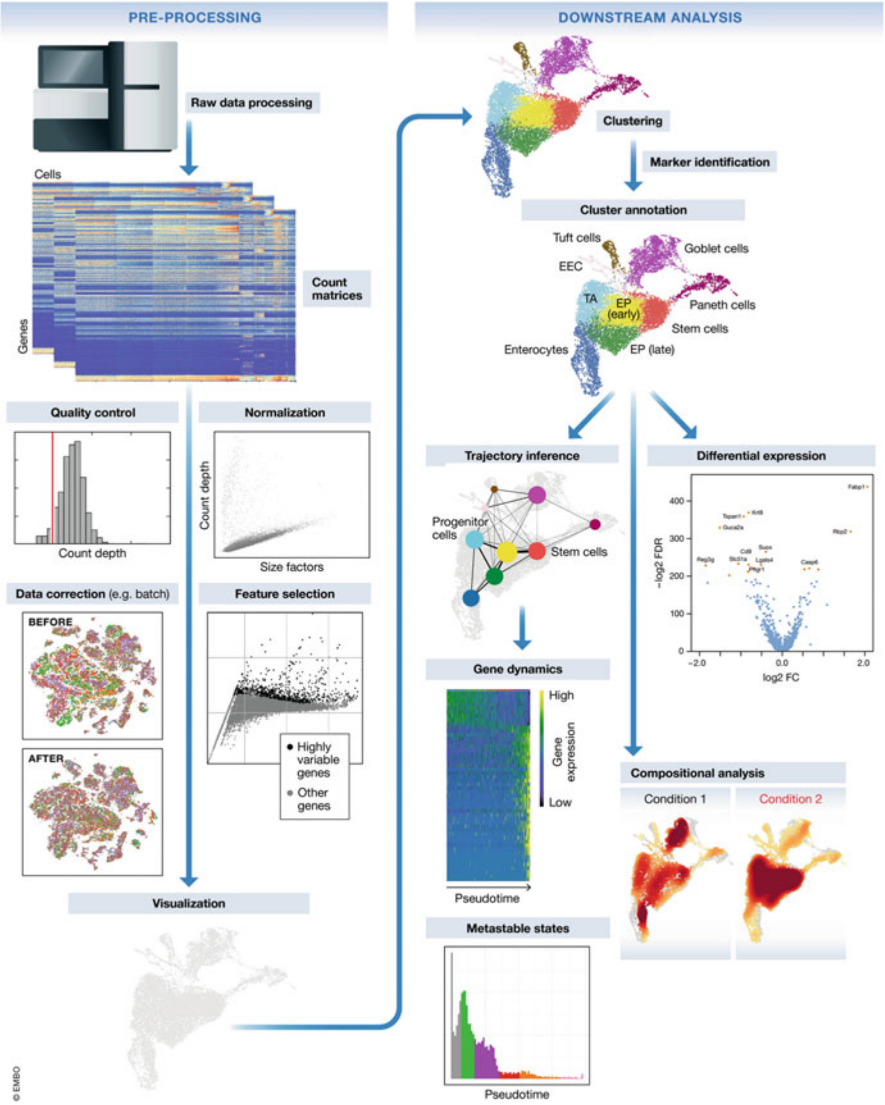


Fig. 25.2 The general computational pipelines that are commonly employed for evaluating scRNA-seq data. (Adapted from [48])

or cytoplasmic RNA), or the ratio of reads to externally spiken RNA molecules, where used in experiments [50].

25.4.2 Batch Effect Correction

Single-cell data are usually derived from several experiments that may experience discrepancies in capture-time, handling, reagent lots, tools, and sometimes even platforms. These disparities contribute to a wide range of variations or batch effects in data, which may confound important information during data integration. Successful elimination of batch results is therefore necessary. Batch results can be incredibly non-linear, making it hard to appropriately coordinate various datasets while retaining significant biological changes [51]. To overcome these problems, the scRNA-seq data is generally subjected to tools designed for microarray data's batch correction like Limma [52] and ComBat [53]. However, single-cell specimens experience "drop-out" episodes because of stochastics gene expression or RNA capture loss or amplification at the time of sequencing [54]. This has contributed to attempts to establish novel workflows and tools for processing data of this type [55].

Pioneered by Haghverdi et al. [55], the common and successful method defines cell mapping among datasets and afterward recreates data in a shared space. This methodology first selects mutually closest neighbors (MNNs) for connecting two datasets. In the corresponding array of paired cells (or MNNs), the translation vector is determined to match the datasets in a shared area. The benefit of this method is that it achieves a structured gene expression matrix, which can be used for the downstream study. However, the processing time and memory are computationally taxing because of the need to measure the neighbors' list in a complex gene expression space. The developers then implemented fastMNN (https://marionilab.github.io/FurtherMNN2018/), which is used in feature space computed with the principal component analysis (PCA) [56] which results in substantial improvements in performance and reliability. Two other approaches, Scanorama [57] and BBKNN [58], also often look for MNNs in dimensional-diminished spaces and use them similarly to guide batch integration.

In 2017, the Satija lab developed the Seurat MultiCCA system from the famous Seurat package [59]. Employing "canonical correlation analysis" (CCA) [60], it decreases data dimensionality and captures the most associated data attributes to match data batches. A novel variant, namely, Seurat Integration (Seurat 3) [61], first uses CCA for projecting data into a feature space for identifying dataset similarities. MNNs are then measured within the CCA subspace and act as "anchors" for data correction. An alternative tool, namely, Harmony [62], employs PCA for minimizing dimensionality. Harmony eliminates the batch results in the PCA space iteratively. It clusters identical cells from separate batches during each iteration, thereby optimizing batches' variability inside the individual cluster, and subsequently determines a correction factor for each cell. This method is fast and also can predict the important biological relation across datasets efficiently.

LIGER is a recently emerging approach that fixes a suspected limitation in other approaches, which is expected to be attributed to technological anomalies and not biological sources to eliminate inconsistencies between the datasets [63]. To achieve the low-dimensional representation of the input data, LIGER uses inclusive non-negative matrix factorization. The illustration consists of two components: a collection of batch factors and a set of factors exchanged. The clustering method then follows a quest for common clusters utilizing a common neighborhood factor diagram for linking cells having identical neighborhoods. Subsequently, the factor-loading quantiles are normalized with the defined clusters to fit a selected dataset (generally set with the maximum cell number), resulting in batches correction. In recent years, researchers have also taken advantage of deep neural networks to use neural networks for solving batch correction. Shaham et al. [64], for instance, developed batch correction residual neural networks for minimizing the mean difference among source and target batch distributions. Lotfollahi et al. [65] have established a scGen, in which a VAE model is trained on the reference dataset before the real data is corrected [66]. Subsequently, reasonable data normalization is required, based on the research's' objective. Several normalization methods have been designed, many of which respond to variations in the depth of sequencing and/or use a spike in and/or Unique molecular identifiers (UMIs).

25.4.3 Normalization

The key purpose of normalization is to minimize the impact of technological influences on the underlying molecular numbers and maintain the original biological heterogeneity [67]. Generally speaking, a gene's normalized expression level should not correspond with a cell's total sequencing depth. Additionally, variation within the normalized gene (across cells) should predominantly indicate biological variability, regardless of sequencing depth or gene abundance. For example, post-normalization, housekeeping genes should experience low variance, whereas genes with greater variance should be expressed differently across cell types [67]. Because of its significance, to date, several different methods for normalizing scRNA-seq data have been proposed [68, 69]. Broadly these approaches can be divided into two groups. The first group tries to predict "scale influences" for each cell, like bulk RNA-seq analysis. For instance, BASiCS employs spike-ins for infusing cell-specific normalizing constants for differentiating technological noise generated from cell-to-cell inconsistency. Scran tool collects cells with comparable library sizes and employs cumulative values for estimating pool-based size factors determined by cell size factors [70]. These techniques, by doing a uniform scaling per cell, presume that the underlying RNA content is unchanged for all cells throughout the dataset and that a single threshold value can be extended to all genes.

Other normalization strategies model molecular counts through probabilistic techniques. Preliminary techniques mainly focus on read-level information and estimate each cell measurement through two components, namely, a mixture of the drop-out (Poisson) and "amplification" (negative binomial, NB) components [54]. For modern UMI-based measurement, modeling approaches have primarily been centered on using the NB [71] component, with a likely additional Zero Inflation Model (ZINB) parameter. The "ZINB noise model" [69] is also used by DCA (https://github.com/theislab/dca) and ScVI (https://github.com/YosefLab/

scvi-tools) either for normalizing and reducing dimensionality through Bayesian hierarchical modeling techniques or to denoise an autoencoder. The revolutionary methods perform beyond pre-processing and normalizations and focus on the detailed evaluation of per-gene error models.

Recently, Hafemeister and Satija hypothesized that Pearson residuals from "regularized negative binomial regression," where the depth of cells sequences is used in a generalized linear model as a covariate, this in turn, effectively eliminates the effect of technological features from subsequent assessments while retaining biological diversity [67]. Importantly, they demonstrate that an unrestrained negative binomial model will overpower and conquer the scRNA-seq data through pooling genes of identical abundances to achieve stable parameter estimations [67]. Additionally, the heuristic steps, including log transformation or pseudo count, are not needed during downstream analytical tasks like dimension reduction, differential expression (DE), and variable gene selection [67].

25.4.4 Dimensionality Reduction and Feature Selection

scRNA-seq data are of high-dimensionality and can contain a huge number of cells and thousands of genes. Reduction of dimensionality and feature selection are two main approaches for handling high-dimensional details [72]. Dimensionality reduction approaches typically map the data to a lower-dimensional space by minimizing some primary data characteristics. PCA is a linear algorithm for dimensional reduction, presuming that the data is disseminated roughly normal. "T-distributed stochastic neighbor embedding" (t-SNE) is a non-linear technique primarily developed to visualize high-dimensional data. In multiple scRNA-seq experiments, both t-SNE and PCA have been commonly used for minimizing data measurements and visualize discriminated cells in various sub-populations [26, 73]. It is pertinent to the dynamic nature of sRNA-seq data, which cannot be efficiently expressed by PCA, and t-SNE has sluggish computational limitations and diverse embeddings to process the same dataset numerous times [27]. scvis [74] and UMAP [75] were recently established to reduce the scRNA-seq data dimensions. Becht et al. have demonstrated that UMAP seems to have the fastest speeds, best reproductivity, and more meaningful cell cluster organization in comparison to other reduction methods [76].

The feature selection eliminates the inaccurate genes and recognizes the most important features for reducing the dimensions' number for downstream study. Lowering the number of the genes through feature selection will accelerate large-scale scRNA-seq calculations [72]. DE in bulk RNA-seq experimentations is a commonly used technique of feature selection. However, it is problematic to use in scRNA-seq data due to the homogeneous and/or pre-determined subpopulation information, which is necessary to call scRNA-seq data for differential expression, that is sometimes not accessible [27]. Unsupervised feature selection approaches, particularly for scRNA-seq data analysis, may be divided into (1) drop-out-based; (2) highly variable genes (HVG), and (3) spike-in-based algorithms [72]. HVG approaches presume that genes with strongly variable cell expression are caused by biological effects instead of technical noise. The HVG methods employ algorithms proposed via Seurat's FindVariableGenes [77] and Brennecke [78] algorithms. Spike-in-based methods distinguish genes with significantly higher variances from spike-ins with similar expression levels. For example, BASiCS and scLVM [79] share a common concept of HVG [72]. In another study, Andrews & Hemberg demonstrated that their M3Drop method surpasses current filtering methods dependent on variance [72].

25.4.5 Clustering and Clustering Annotation

One of the scRNASeq's common uses is to classify and describe cell populations. Scientifically speaking, cell groups mostly comprise distinct cell classes, e.g., neurons and glia in a brain study, but they may often lead to different cell-type states, e.g., inactivated and activated T-cells [72]. From a statistical standpoint, ScRNA-Seq's cell populations are unsupervised. To date, the topic has been extensively studied within machine learning research, and several excellently established techniques have been developed for scRNASeq data. The clustering algorithms can primarily be categorized into (1) *k*-means (KM); (2) graph-based clustering, (3) Density-based clustering (DBC), and (4) hierarchical clustering [27].

KM is a widely employed clustering algorithm at the time of single-cell analysis. It is a straightforward approach that allocates cells to the nearest cluster's center, and subsequently re-calculates the cluster centroids. Nevertheless, KM involves pre-determination of the cluster's number and employs stochastic starting points for every cluster. Hence, it demands that the robustness of these parameters be tested several times. These multiple outcomes can then be integrated by measuring a consensus, for instance, like SC3 does [80]. The major limitation of KM is that the system presumes a pre-determined quantity of equally sized rounded clusters. When these presumptions are breached, KM will recognize several neighboring clusters anywhere along the differentiation route and combine uncommon cells having a much more dominant cell type. Tools that combine k mean with outlier identification strategies, for instance, RaceID [81], may be used to classify rare cell populations. RaceID, however, fails miserably when data may not include rare cell populations [81].

The clustering hierarchy is another prominent clustering tool that is widely used for classifying cell populations [72]. Though there are numerous types of hierarchical clustering, the most widely used during cluster analysis are Ward's [82] and "complete." However, the *k*-mean is faster in comparison to hierarchical clustering. Nevertheless, hierarchical clustering has the benefit of evaluating the connections between the numerous small clusters, because the outcome can be depicted as a dendrogram. This dendrogram is then "broken" to establish several numbers of groups at varying heights. CIDR [83], PcaReduce [84], and SINCERA [85] are the most widely used approaches for hierarchical clustering of single-cell RNASeq results. However, these approaches fail to classify clusters that might have the same type of cell [72].

DBC defines clusters as adjoining areas with a higher cell density. In comparison to hierarchical clustering or KM, DBC does not presume clusters of a specific type or scale. Density-based approaches, therefore, also presume that all clusters are similarly dense, that is, homogeneous cell populations. Furthermore, the algorithm requires to be granted this density through one or more criteria. The density parameter setting is similar to selecting the number of clusters for KM or deciding where to cut the tree for hierarchy. Because DBC involves several samples, it works well on cytometry experiments, droplet-based datasets, and large RT-qPCR experiments, which include data for tens of millions to billions of cells [40, 72, 86].

Graph-based clustering, also known as "community detection," is an enhancement of DBC and primarily designed for data defined as a graph, that is, a collection of "edges"-connected cells. Because graphs can effectively be a complicated non-linear system with limited assumptions, it is possible to classify cell populations with varying sizes, densities, and types [87]. The graphic approaches have also been used to scale them thousands to hundreds of cells [88, 89]. The graph's density can be calculated as the number of edges joining a group of cells and comparing it with a null hypothesis, such as a totally random graph or a random graph regulated with metric called modularity. The most common modular methods are the Louvain algorithm [87] employed within the PhenoGraph [90] and Seurat version 1.4. Another density measurement method employed through SNN-Cliq utilize overlaps among each cell's *k*-nearest neighbors [91].

The key disadvantage to graph-based approaches is that scRNASeq has no underlying graph form. Thus, these methods' efficiency depends on how efficiently the scRNASeq data is transformed into a graph illustration. Typically, scRNASeq data is transformed into a graph via describing cells as edge-connected nodes to their *k* nearest neighbors (kNN) [77, 92]. This interpretation implies equivalent to cell populations. Nevertheless, owing to the dimensionality curse, defining *k*-nearest-neighbors could not be a robust technique [93]. Hence, feature selection and/or few dimensionality reduction are often needed before defining kNN graphs for avoiding biased clustering algorithms [72].

25.4.6 Differential Expression Analysis

A significant feature of the bulk RNA series is DE analysis. Several resources are available, and DESeq2 & edgeR are the most widely used. As scRNA-seq expression data is zero-inflated, single-cell data is somewhat distinct from traditional bulk RNA sequencing. In bulk RNAseq, DE is used primarily for evaluating a few replicates of two or more biological conditions [94]. In scRNAseq, DE analysis is routinely used to recognize genes in scRNA-seq experiments that can be distinguished among cell subpopulations, research environments, and between case-control categories. SCDE [54], MAST [95], and ZingeR [96] are frequently used for DE methods in scRNA-seq. Although various DE approaches allow different modeling assumptions for capturing different facets of the scRNA-seq data [97], almost all tools analyze one gene at a time.

The study of a gene could contribute to possible power loss, because this method does not use reliable DE data that would otherwise be used to boost the DE study's power. Various DE detecting algorithms in scRNA-seq methods may appear to favor different DE genes in real data applications because of their low statistical capacity, resulting in weak output and incoherency of findings across multiple techniques. It has been well established in several other forms of association analyses such as genome-wide association studies that Bayesian techniques, which model several of the predictor variables, also together with a simplified composite probability technique, where information is collected via several predictors, each handled separately, significantly increase power against univariate approaches [98].

Gene expression (GSE) analysis is often a standard task to integrate DE information at the genes set or pathways level. GSE analysis can promote the rigorous biological interpretation of DE findings by combining gene-level data. To date, several various GSE analytical methods were developed, but almost all are developed in the RNAseq research environment [99]. These current GSE techniques involve over-representation analytical techniques like the Fisher exact test [100] and DAVID [101]; testing self-contained methods, e.g., *t*-test [102], and competitive test methods such as GSEA [103], PAGE [104], and CAMERA [105]. Although the GSE methods are plentiful, their efficiency in scRNA-seq research remains uncertain. In fact, no comprehensive studies have been carried out to date to determine the efficacy of current GSE approaches in the scRNA-seq context. Besides that, and more critically possibly, almost all current GSE approaches consider GSE research as a separate analytical phase after DE research.

However, GSE analysis, as well as DE analysis, are statistically connected. While DE findings are definitely invaluable in identifying enriched gene sets to conduct GSE analyses, enriched or non-enriched gene sets provide useful knowledge that may provide input on DE analysis for improving their statistical strength. The combination of DE and GSE analysis can significantly enhance both the capacity and viability of the scRNA-seq analyses [106]. Recently, Ma et al. [106] created an integrative and versatile computational tool, namely, iDEA, for joint DE and GSE analysis via a Bayesian hierarchical system. By combining DE and GSE analyses, iDEA enhances the strength, consistency, and accuracy of DE research. Importantly, iDEA only requires DE summary statistics as inputs for productive data modeling via complimenting and matching different current DE approaches. iDEA's power gain helps one to find several pathways that could not be found via other existing approaches [106].

25.4.7 Trajectory Inferences

A key challenge during the study of developmental biology is understanding the series of fate decisions that contribute to every mature cell type in tissue and organism [107]. The hierarchy may be traced employing lineage tracing, in which a tracer molecule or DNA change is incorporated in a community of initial cells and tracked over time, enabling the recognition of the progeny of the cells [108]. Recently, developments in DNA sequences have rendered that thousands of lineage trace assays can be paralleled in a single experiment by labeling cells with particular DNA barcodes. This can be traced through lineage tracing, in which the tracer or DNA alteration in an early cell population is inserted and over time tracked, enabling for identifying the progeny of the cells [109, 110].

Two novel methods are widely used for performing lineage tracing. While "Prospective" lineage tracing attempts to assess the destiny of a group of cells that were labeled at an early stage by observing them at a specific point in time, "Retrospective" lineage reconstruction attempts to recreate the lineage associations among cells at a single time point as a means to conclude the past division from branching incidents that they have undergone [111]. This method has its origins in the practice of inferring phylogeny among organisms on the basis of their common and unique features, like shared anatomical characteristics or gene sequence alleles missing from an outgroup [112].

However, the need for accumulating variations within barcodes across a large developmental window limits this phylogenetic approach. It is oblivious to destiny options after diversification of the barcode has ceased. To date, several experimental techniques have now been developed for continuous barcode cells [113]. However, these approaches need optimization to enable standardized long-term barcoding rates and evaluate tissues with variable division rates [114]. Since most of the current approaches only mark cells inside a narrow-time window [110, 115, 116]; it may help to establish lineage reconstruction mechanisms outside barcoding. In a restricted situation, one might wonder if it is feasible to create a retrospective lineage association where clonal barcoding happens only once in a single cell population [107].

Considering this, recently, Weinreb and Klein explored whether the stochasticity of cell fate selection during development may be used to infer lineage associations at a single period after bar code [107]. Result obtained revealed that the approach is most appropriate for experimental data obtained from widely used barcoding approaches [107]. However, these findings are based on a variety of biological hypotheses that do not necessarily adhere. For instance, the branching model eliminates cell-cell interactions, which could not be represented by a medium area. Though numerous mechanisms are well captured by those assumptions, some phenomena are not considered, like strict asymmetric divisions [117]. The model often lacks the cell death or normal replication information that might have generated because of auto-renewal of the cell state, believing that bar codes are all accumulated in one cell form at a single stage of growth, which only applies roughly for specific experimental strategies [110, 116] and not for others [113, 118].

25.4.8 Compositional Analysis

At the cell stage, clustered data may be studied with respect to their compositional form. The study of compositional data revolves around the cell proportions that fall into each cell-identity cluster. In response to illness, these proportions will change. Salmonella infection, for example, has been shown to enhance the capacity of enterocytes in the epithelium of the mouse intestine [119]. Thus, analyzing compositional modifications to single-cell data demands adequate cell numbers in order to determine robustly the cell-identity cluster proportions and sufficiently sample numbers in order to analyze predicted cell-identity cluster history variations. As suitable databases have been accessible only recently, such methods are still to be created. Earlier study in mouse, cell identity counts were modeled with a Poisson procedure, including the condition as a covariate and the total number of cells as offset detected. In this scenario, a predictive test can be conducted via the regression coefficient to determine whether the occurrence of a certain cell identity has changed substantially [48]. Measures on other cell identities in the same dataset are nevertheless not independent. When the proportion of a cell-identity cluster increases, the proportion of all other clusters would also have shifted. Therefore, it is difficult to determine if the general composition has modified dramatically using this model. In the absence of special tools, visual comparison of compositional data can notify changes in composition across samples. Future advances in this area are likely to take advantage of mass cytometry or microbiome literature in which greater priority has been given to compositional data analysis [48].

25.5 Applications of scRNA-seq Technologies

scRNA-seq technologies are continuously being employed to unmask various biological processes ranging from epigenetic regulation to biomarker identification.

25.5.1 Epigenetic Modification

Epigenetic changes are characterized as transcriptional repression or activation generated via associations produced in bulk cell populations. Nevertheless, the research has shown this hypothesis's naivety and the extreme complexity during epigenetic regulation [120]. For instance, 5-Methylcytosine (5mC) is largely believed to be a transcriptionally repressive mark, as the promoter's methylation is negatively linked to gene expression. In certain instances, however, DNA methylation has been positively associated with transcript, indicating that the genome background may also influence the biological result [121]. Furthermore, global DNA hypomethylation shown by naive embryonic stem cell(ESCs) does not require a widespread transcriptional initiation, which indicates that the intensity of regulatory relationships among DNA methylation and transcription can differ on the basis of the stage of the production and cellular background [120]. Hence, the usage of single-cell methods can enhance our interpretation of DNA changes as epigenetic regulatory markers (Table 25.2). For such experiments, the latest advancements of integrated single-cell approaches (e.g., scM&T-seq) would be great. Moreover, very low levels of 5-Hydroxymethylcytosine (5hmC) determined in mass cell samples (e.g. <5% of CpG sites in primed ESCs) suggest that this alteration is present in just a few cells with some unique trace in cytosine. Therefore, simultaneous 5hmC profiling and transcription would profoundly influence our interpretation of this epigenetic mark. In the future, several epigenetic traits (e.g., DNA methylation and accessibility to chromatin) can also be tested with gene expression in the same human cell, which will contribute to more refinements in our perspective of the epigenomic effect on the transcriptome [120] (Fig. 25.3).

25.5.2 Clinical Studies

Earlier researches have also shown that genetic variants may contribute to cellular heterogeneity within the tumor tissue. However, conventional sequencing strategies, like RNA-seq, are only incapable of capturing signal at cellular level. These limitations can be adequately overcome through scRNA-seq technologies [131]. Zhang et al. (2018) analyzed the T-cell immunoreceptor of colorectal cancer through scRNA-seq technology and reported a potential state transition between T-cell populations and tissue-distributed subpopulations [132]. Bian et al. employed scRNA-seq technologies and reported association between metastasis of the human colorectal cancer, gene expression alteration, a genomic variance of copy number, and irregular DNA methylation [133]. In another single-cell study, T-cell immune map of lung and liver cancer microenvironment were plotted by Chinese researchers [134, 135]. They provided detailed insight of subgroups, tumor heterogeneity, the features of the tissue distribution, and the pattern of gene expression. This research provides detailed insight into the liver and lung cancer's immune microenvironments and, in the near future, may be useful for discovering successful biomarkers, novel tumor immunotherapy, and drug targets [134, 135]. Ledergor et al. [136] reported that the high variability within myeloma plasma cells can be influenced by variations in the gene expression pattern and the samples' chromosome layout.

scRNA-seq technologies can also identify trait/disease specific immune cells, how these immune cells differentiate into various classes of immune cells and their associations. This, in turn, explains the diverse immune system and also help us in detecting a new biomarker [131]. For instance, Crinier et al. [137] through scRNA-seq sequencing reported about distinguishing characteristics that separate spleen & blood NK cell in humans and mice. In another study, Villani et al. identified a new subclass of Dendritic cells (DC) with plasmacytoid but can also powerfully activate T cells [138]. Recently, Wilk et al. employed scRNA-seq technologies for profiling of peripheral blood mononuclear cells from seven COVID-19-hospitalized patients [139]. They identified COVID-19-associated peripheral immune cell phenotype, including the downregulation of HLA class II and the development of neutrophil

Table 25.2 Current and emerging single-cell epigenetics techniques (Adapted from [120])

Technique	Epigenomic feature	Method	Approach	Single cell
Cytosine modification	5mC	Aba-seq	5hmC specific restriction enzyme	Possible
		BS-seq	Bisulfite converts only C but not 5hmC (or 5mC) into U thus only methylated sites are sequenced as "C"	Yes [122–124]
		hMeDIP-seq	Sequencing preceded via 5hmC DNA immunoprecipitation	Not currently possible
		MeDIP-seq	Sequencing preceded via 5mC DNA immunoprecipitation	Not currently possible
		Methyl-seq	Sequencing preceded via 5mC specific restriction enzyme	Possible
		oxBS-seq	5hmC is oxidized to 5caC so that only 5mC survives bisulfite alteration. Readout is pure 5mC as well as subtraction from BS-seq regulates 5hmC	Not possible for measuring 5hmC due to the need for subtraction
		TAB-seq	Maps 5hmC through enzymatic oxidation before bisulfite treatment: only 5hmC survives conversion	Possible
Protein–DNA interaction	Histone modification	ChIP-seq	DNA immunoprecipitation attached with a specific histone variant or transcription factor	Yes [125]
	Transcription factor binding	DamID	Fusion of Dam protein and transcription factor gene transfect cells that methylates adenine residues within proximity toward the binding site. 6 mA specific restriction digest is employed for mapping.	Yes for nuclear lamina interactions [126]
Chromatin structure	DNA accessibility	ATAC-seq	Tn5 transposase enzyme fragments and bind adapter with open chromatin	Yes [127, 128]
		DNase-seq	DNaseI digestion of open chromatin into small fragments suitable for library preparation as well as sequencing	Yes [129]

(continued)

Table 25.2 (continued)

Technique	Epigenomic feature	Method	Approach	Single cell
		FAIRE-seq	Chromatin is crosslinked, sonicated, and subsequently purified through phenol–chloroform extraction. The aqueous layer comprises of only DNA that are not associated with protein	Not currently possible
	Nucleosome positioning	MNase-seq	Microcococal nuclease digestion of chromatin and sequencing of the product which are regions protected by nucleosomes	Possible
		NOME-seq	GpC methylation of DNA not protected by nucleosomes followed by BS-seq	Possible
Three-dimensional organization	Chromosome conformation	HiC	DNA is crosslinked, then restriction digested to fragment before ligation and reversal of the crosslinks. Resulting fragments are hybrids from separate genomic locations that were in close proximity in three-dimensional space. Paired-end sequencing is used to link the two regions	Yes [130]

C cytosine, *5caC* 5-carboxylcytosine, *5hmC* hydroxymethylcytosine, *5mC* methylcytosine, *U* uracil

community, that may lead to plasmablasts in acute respiratory failure patients requiring mechanical ventilation [139].

25.5.3 Development and Regeneration

scRNA-seq technologies can also be used for sequencing and quantifying the complete genome of both embryonic and germ cells at a single-cell level. This in turn will explain the germ cells' origin and also help us to detect, diagnose, and treat reproductive and genetic disorders [131]. Recently, Chen et al. [140] employed scRNA-seq technologies to unmask the complex mechanism related with spermatogenesis in mice. Result obtained revealed the unique patterns of alternate splicing and central regulators that are associated with the various growth stages of male germ cells. Vento-Tormo et al. [141] conducted the early pregnancy's placental cell transcriptome analysis using scRNA-seq technology. The cellular compositional

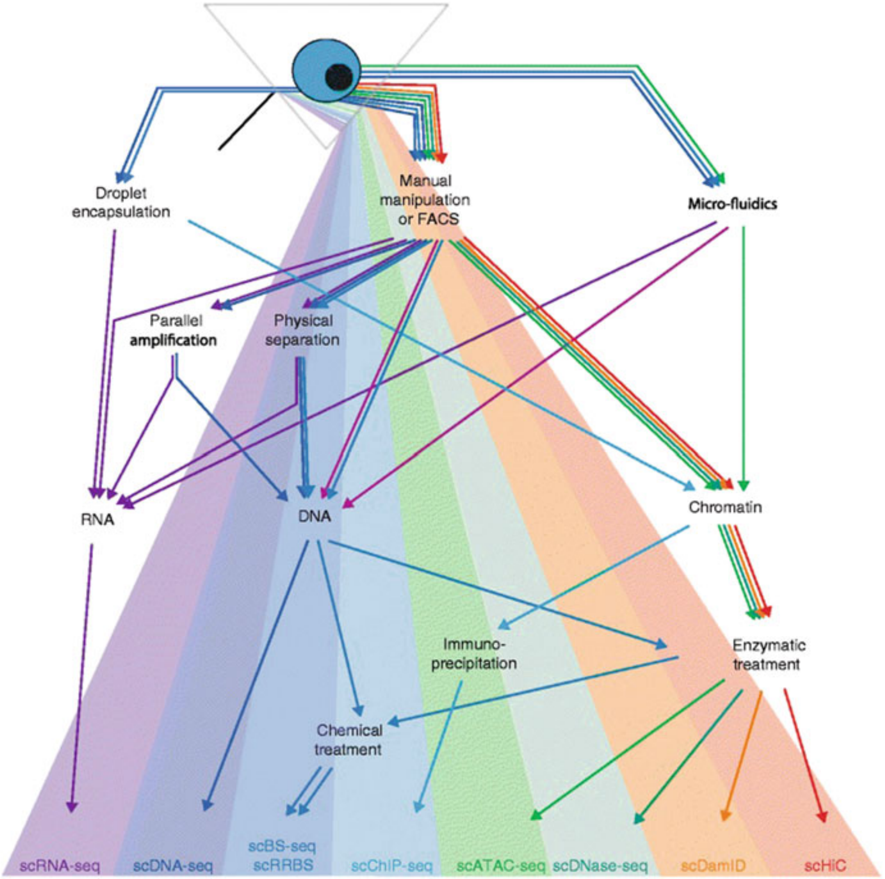


Fig. 25.3 The diagram outlines epigenomics and the spectrum of scRNA-seq technologies. It is pertinent to note that single-cell bisulfite sequencing (scBS-seq) conversion is not consistent with the simultaneous amplification of RNA or DNA, because DNA methylation is not preserved during in vitro amplification. In single cell epigenomic methods, chemical therapies are used of DNA (conversion of bisulfite), in the analysis of DNA improvements (scBS-seq and scRRBS), immuno-precipitation or enzyme digesting (for example, DNaseI), histone modifications (scChIP-seq, chromatin conformation (scDamID, scHiC), and accessibility of DNA (scATAC-seq, scDNase-seq). (Adapted from [120])

analysis of the human decidua reveals sub-sets of perivascular and stroma cells in different decidual layers. They also identified three major subsets of decidual natural killer cells that have distinct chemokine and immunomodulatory profiles [141]. This study has also established regulatory processes that can minimize unhealthy mothers' immune response. Hence, these findings are helpful toward understanding the initial pregnancy and for detecting and treating pregnancy-associated diseases [141].

scRNA-seq technologies has also provided us with an opportunity to study how the cell-type specificity and temporal variation influence the regenerative response to injury in both plants and animals at incredible cell resolution [142]. By combining scRNA-seq technologies and inducible Cre/loxP-mediated lineage tracing techniques, earlier researcher have demonstrated that axolotl limb regeneration is associated with fibroblast dedifferentiation instead of pre-existing stem cells that are not present during blastema formation [143]. A similar combinatorial method was used by another group of researcher to understand how the dissection of the root tip could cause root regeneration, leading to a hypothesis that root tip regeneration primarily depends on cell's types and phases [144, 145]. A recent single-cell study of *Arabidopsis* roots has also revealed that several cell types could quickly reassemble stem cells by recreating embryogenesis patterns [145], endorsing the notion of a centralized stem cell management system [146]. Thus, single-cell genomic research offers unique opportunity to detect cells that have key roles in tissue growth, regeneration, repair, and disease formation.

25.6 Conclusion and Future Perspective

In conclusion, scRNA-seq is revolutionizing our basic view of biology. This approach has introduced new frontiers for studying that goes well beyond descriptive cell-state studies. scRNA-seq technologies are widely employed to unmask molecular mechanisms associated with oncology, neurology, microbiology, immunology, reproduction, urinary & digestive systems, and plant biology. scRNA-seq technologies, however, do have difficulties like cumbersome activity and high detection costs that restrict technology promotion [131]. Though there are several protocols for scRNA-seq study, practically all require poly-A selection, thereby restricting the opportunity to analyze non-polyadenylated transcripts, like pre-mRNAs, histone mRNAs, small nucleolar RNAs, and long non-coding RNAs, which may play different regulatory characters in cancer [147, 148]. Even in poly-A allowed scRNA-seq protocols, droplet-based protocols, which limit 3′ or 5′, are intrinsically more restrictive than full-transcript single-cell RNA-seq protocols. Moreover, certain cell types (e.g., neurons, epithelial cells, and neutrophils) may not be consistent with all scRNA-seq protocols dissociation, encapsulation, or other processing measures [148]. Earlier researchers have suggested that the exact shortcomings and weaknesses can be overcome by unified studies of the same cancer samples (fresh vs. frozen and whole-cell vs. nuclei) [148]. Soon, it is hoped that technology for single-cell sequencing will be streamlined and more efficient. Additionally, the detection cost will be too minimized so that innovations can be extended to basic research and play a key role in clinical diagnosis and care. Coupled with gene-editing technology, including modeling target-based regulatory networks, single-cell sequencing may also accelerate crop improvement [149]. Additionally, these approaches will also play a key function during clinical diagnosis and treatment, which will enhance the drug discovery process.

Conflict of Interest None

Additional Information Figure 25.1 [14], Fig. 25.2 [48] and Fig. 25.3 [120] are reused under the terms of the Creative Commons Attribution License CC BY 4.0.

References

1. Goldman SL, MacKay M, Afshinnekoo E, Melnick AM, Wu S, Mason CE. The impact of heterogeneity on single-cell sequencing. Front Genet [Internet]. 2019 [cited 2020 Oct 6];10. https://www.frontiersin.org/articles/10.3389/fgene.2019.00008/full.
2. Grimbergen AJ, Siebring J, Solopova A, Kuipers OP. Microbial bet-hedging: the power of being different. Curr Opin Microbiol. 2015;25:67–72.
3. Li S, Garrett-Bakelman F, Perl AE, Luger SM, Zhang C, To BL, et al. Dynamic evolution of clonal epialleles revealed by methclone. Genome Biol. 2014;15(9):472.
4. Li S, Garrett-Bakelman FE, Chung SS, Sanders MA, Hricik T, Rapaport F, et al. Distinct evolution and dynamics of epigenetic and genetic heterogeneity in acute myeloid leukemia. Nat Med. 2016;22(7):792–9.
5. Mason CE, Porter SG, Smith TM. Characterizing multi-omic data in systems biology. Adv Exp Med Biol. 2014;799:15–38.
6. Eberwine J, Yeh H, Miyashiro K, Cao Y, Nair S, Finnell R, et al. Analysis of gene expression in single live neurons. Proc Natl Acad Sci U S A. 1992;89(7):3010–4.
7. Brady G, Barbara M, Iscove NN. Representative in vitro cDNA amplification from individual hemopoietic cells and colonies. Methods Mol Cell Biol. 1990;2(1):17–25.
8. Hwang B, Lee JH, Bang D. Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med. 2018;50(8):1–14.
9. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods. 2009;6(5):377–82.
10. Shaffer SM, Dunagin MC, Torborg SR, Torre EA, Emert B, Krepler C, et al. Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance. Nature. 2017;546(7658):431–5.
11. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, et al. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013;498(7453):236–40.
12. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. 2014;32(4):381–6.
13. Stubbington MJT, Lönnberg T, Proserpio V, Clare S, Speak AO, Dougan G, et al. T cell fate and clonality inference from single-cell transcriptomes. Nat Methods. 2016;13(4):329–32.
14. Hu P, Zhang W, Xin H, Deng G. Single cell isolation and analysis. Front Cell Dev Biol [Internet]. 2016 [cited 2020 Oct 7];4. https://www.frontiersin.org/articles/10.3389/fcell.2016.00116/full.
15. Gross A, Schoendube J, Zimmermann S, Steeb M, Zengerle R, Koltay P. Technologies for single-cell isolation. Int J Mol Sci. 2015;16(8):16897–919.
16. Schulz KR, Danna EA, Krutzik PO, Nolan GP. Single-cell Phospho-protein analysis by flow cytometry. Curr Protoc Immunol. 2007;78(1):8.17.1–8.17.20.
17. Miltenyi S, Müller W, Weichel W, Radbruch A. High gradient magnetic cell separation with MACS. Cytometry. 1990;11(2):231–8.
18. Grützkau A, Radbruch A. Small but mighty: how the MACS®-technology based on nanosized superparamagnetic particles has helped to analyze the immune system within the last 20 years. Cytometry A. 2010;77A(7):643–7.
19. Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, et al. Laser capture microdissection. Science. 1996;274(5289):998–1001.

20. Esposito G. Complementary techniques. In: Mocellin S, editor. Microarray technology and cancer gene profiling [Internet]. New York: Springer; 2007 [cited 2020 Oct 8]. p. 54–65. (Advances in experimental medicine and biology). https://doi.org/10.1007/978-0-387-39978-2_6.
21. Kummari E, Guo-Ross SX, Eells JB. Laser capture microdissection - a demonstration of the isolation of individual dopamine neurons and the entire ventral tegmental area. J Vis Exp. 2015;(96):e52336.
22. Fend F, Raffeld M. Laser capture microdissection in pathology. J Clin Pathol. 2000;53(9):666–72.
23. Bonner RF, Buck ME, Cole K, Pohida T, Chuaqui R, Goldstein S, et al. Laser capture microdissection: molecular analysis of tissue. Science. 1997;278(5342):1481.
24. Arora A, Simone G, Salieb-Beugelaar GB, Kim JT, Manz A. Latest developments in micro total analysis systems. Anal Chem. 2010;82(12):4830–47.
25. Whitesides GM. The origins and the future of microfluidics. Nature. 2006;442(7101):368–73.
26. Rosenberg AB, Roco CM, Muscat RA, Kuchina A, Sample P, Yao Z, et al. Single-cell profiling of the developing mouse brain and spinal cord with split-pool barcoding. Science. 2018;360(6385):176–82.
27. Chen G, Ning B, Shi T. Single-cell RNA-Seq technologies and related computational data analysis. Front Genet. 2019;10:317.
28. Sasagawa Y, Nikaido I, Hayashi T, Danno H, Uno KD, Imai T, et al. Quartz-Seq: a highly reproducible and sensitive single-cell RNA sequencing method, reveals non-genetic gene-expression heterogeneity. Genome Biol. 2013;14(4):R31.
29. Ramsköld D, Luo S, Wang Y-C, Li R, Deng Q, Faridani OR, et al. Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells. Nat Biotechnol. 2012;30(8):777–82.
30. Picelli S, Björklund ÅK, Faridani OR, Sagasser S, Winberg G, Sandberg R. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. Nat Methods. 2013;10(11):1096–8.
31. Fan X, Zhang X, Wu X, Guo H, Hu Y, Tang F, et al. Single-cell RNA-seq transcriptome analysis of linear and circular RNAs in mouse preimplantation embryos. Genome Biol. 2015;16(1):148.
32. Sheng K, Cao W, Niu Y, Deng Q, Zong C. Effective detection of variation in single-cell transcriptomes using MATQ-seq. Nat Methods. 2017;14(3):267–70.
33. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Characterization of the single-cell transcriptional landscape by highly multiplex RNA-seq. Genome Res. 2011;21(7):1160–7.
34. Islam S, Kjällquist U, Moliner A, Zajac P, Fan J-B, Lönnerberg P, et al. Highly multiplexed and strand-specific single-cell RNA 5′ end sequencing. Nat Protoc. 2012;7(5):813–28.
35. Hashimshony T, Wagner F, Sher N, Yanai I. CEL-Seq: single-cell RNA-Seq by multiplexed linear amplification. Cell Rep. 2012;2(3):666–73.
36. Hashimshony T, Senderovich N, Avital G, Klochendler A, de Leeuw Y, Anavy L, et al. CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. 2016;17:77.
37. Zheng GXY, Terry JM, Belgrader P, Ryvkin P, Bent ZW, Wilson R, et al. Massively parallel digital transcriptional profiling of single cells. Nat Commun. 2017;8(1):14049.
38. Fan HC, Fu GK, Fodor SPA. Expression profiling. Combinatorial labeling of single cells for gene expression cytometry. Science. 2015;347(6222):1258367.
39. Habib N, Avraham-Davidi I, Basu A, Burks T, Shekhar K, Hofree M, et al. Massively parallel single-nucleus RNA-seq with DroNc-seq. Nat Methods. 2017;14(10):955–8.
40. Macosko EZ, Basu A, Satija R, Nemesh J, Shekhar K, Goldman M, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell. 2015;161(5):1202–14.
41. Klein AM, Mazutis L, Akartuna I, Tallapragada N, Veres A, Li V, et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell. 2015;161(5):1187–201.

42. Jaitin DA, Kenigsberg E, Keren-Shaul H, Elefant N, Paul F, Zaretsky I, et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. Science. 2014;343(6172):776–9.
43. Cao J, Packer JS, Ramani V, Cusanovich DA, Huynh C, Daza R, et al. Comprehensive single-cell transcriptional profiling of a multicellular organism. Science. 2017;357(6352):661–7.
44. Gierahn TM, Wadsworth MH, Hughes TK, Bryson BD, Butler A, Satija R, et al. Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput. Nat Methods. 2017;14(4):395–8.
45. Sasagawa Y, Danno H, Takada H, Ebisawa M, Tanaka K, Hayashi T, et al. Quartz-Seq2: a high-throughput single-cell RNA-sequencing method that effectively uses limited sequence reads. Genome Biol [Internet]. 2018 [cited 2020 Dec 8];19. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5845169/.
46. Ziegenhain C, Vieth B, Parekh S, Reinius B, Guillaumet-Adkins A, Smets M, et al. Comparative analysis of single-cell RNA sequencing methods. Mol Cell. 2017;65(4):631–643.e4.
47. External RNA Controls Consortium. Proposed methods for testing and selecting the ERCC external RNA controls. BMC Genomics. 2005;6(1):150.
48. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol [Internet]. 2019 [cited 2019 Nov 15];15(6). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6582955/.
49. Davis MPA, van Dongen S, Abreu-Goodger C, Bartonicek N, Enright AJ. Kraken: a set of tools for quality control and analysis of high-throughput sequence data. Methods (San Diego, Calif). 2013;63(1):41–9.
50. Ilicic T, Kim JK, Kolodziejczyk AA, Bagger FO, McCarthy DJ, Marioni JC, et al. Classification of low quality cells from single-cell RNA-seq data. Genome Biol. 2016;17:29.
51. Tran HTN, Ang KS, Chevrier M, Zhang X, Lee NYS, Goh M, et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. Genome Biol. 2020;21(1):12.
52. Smyth GK, Speed T. Normalization of cDNA microarray data. Methods. 2003;31(4):265–73.
53. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007;8(1):118–27.
54. Kharchenko PV, Silberstein L, Scadden DT. Bayesian approach to single-cell differential expression analysis. Nat Methods. 2014;11(7):740–2.
55. Haghverdi L, Lun ATL, Morgan MD, Marioni JC. Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. Nat Biotechnol. 2018;36(5):421–7.
56. Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans R Soc A Math Phys Eng Sci. 2016;374(2065):20150202.
57. Hie B, Bryson B, Berger B. Efficient integration of heterogeneous single-cell transcriptomes using Scanorama. Nat Biotechnol. 2019;37(6):685–91.
58. Polański K, Young MD, Miao Z, Meyer KB, Teichmann SA, Park J-E. BBKNN: fast batch alignment of single cell transcriptomes. Bioinformatics. 2020;36(3):964–5.
59. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol. 2018;36(5):411–20.
60. Hardoon DR, Szedmak S, Shawe-Taylor J. Canonical correlation analysis: an overview with application to learning methods. Neural Comput. 2004;16(12):2639–64.
61. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM, et al. Comprehensive integration of single-cell data. Cell. 2019;177(7):1888–1902.e21.
62. Korsunsky I, Millard N, Fan J, Slowikowski K, Zhang F, Wei K, et al. Fast, sensitive and accurate integration of single-cell data with harmony. Nat Methods. 2019;16(12):1289–96.
63. Welch J, Kozareva V, Ferreira A, Vanderburg C, Martin C, Macosko E. Integrative inference of brain cell similarities and differences from single-cell genomics. bioRxiv. 2018;2:459891.
64. Shaham U, Stanton KP, Zhao J, Li H, Raddassi K, Montgomery R, et al. Removal of batch effects using distribution-matching residual networks. Bioinformatics. 2017;33(16):2539–46.

65. Lotfollahi M, Wolf FA, Theis FJ. Generative modeling and latent space arithmetics predict single-cell perturbation response across cell types, studies and species. bioRxiv. 2018;478503.
66. Goodfellow IJ, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, et al. Generative adversarial networks. arXiv:14062661 [cs, stat] [Internet]. 2014 [cited 2020 Oct 9]. http://arxiv.org/abs/1406.2661.
67. Hafemeister C, Satija R. Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. 2019;20(1):296.
68. Vallejos CA, Marioni JC, Richardson S. BASiCS: Bayesian analysis of single-cell sequencing data. PLoS Comput Biol. 2015;11(6):e1004333.
69. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053–8.
70. Lun ATL, Bach K, Marioni JC. Pooling across cells to normalize single-cell RNA sequencing data with many zero counts. Genome Biol. 2016;17(1):75.
71. Grün D, Kester L, van Oudenaarden A. Validation of noise models for single-cell transcriptomics. Nat Methods. 2014;11(6):637–40.
72. Andrews TS, Hemberg M. Identifying cell populations with scRNASeq. Mol Asp Med. 2018;59:114–22.
73. Chen G, Schell JP, Benitez JA, Petropoulos S, Yilmaz M, Reinius B, et al. Single-cell analyses of X chromosome inactivation dynamics and pluripotency during differentiation. Genome Res. 2016;26(10):1342–54.
74. Ding J, Condon A, Shah SP. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. Nat Commun. 2018;9(1):2002.
75. McInnes L, Healy J, Melville J. UMAP: uniform manifold approximation and projection for dimension reduction. arXiv:180203426 [cs, stat] [Internet]. 2018 [cited 2019 Dec 31]. http://arxiv.org/abs/1802.03426.
76. Becht E, McInnes L, Healy J, Dutertre C-A, Kwok IWH, Ng LG, et al. Dimensionality reduction for visualizing single-cell data using UMAP. Nat Biotechnol. 2019;37(1):38–44.
77. Satija R, Farrell JA, Gennert D, Schier AF, Regev A. Spatial reconstruction of single-cell gene expression data. Nat Biotechnol. 2015 May;33(5):495–502.
78. Brennecke P, Anders S, Kim JK, Kołodziejczyk AA, Zhang X, Proserpio V, et al. Accounting for technical noise in single-cell RNA-seq experiments. Nat Methods. 2013;10(11):1093–5.
79. Buettner F, Natarajan KN, Casale FP, Proserpio V, Scialdone A, Theis FJ, et al. Computational analysis of cell-to-cell heterogeneity in single-cell RNA-sequencing data reveals hidden subpopulations of cells. Nat Biotechnol. 2015;33(2):155–60.
80. Kiselev VY, Kirschner K, Schaub MT, Andrews T, Yiu A, Chandra T, et al. SC3: consensus clustering of single-cell RNA-seq data. Nat Methods. 2017;14(5):483–6.
81. Grün D, Lyubimova A, Kester L, Wiebrands K, Basak O, Sasaki N, et al. Single-cell messenger RNA sequencing reveals rare intestinal cell types. Nature. 2015;525(7568):251–5.
82. Ward JH Jr. Hierarchical grouping to optimize an objective function. J Am Stat Assoc. 1963;58(301):236–44.
83. Lin J-T, Lee W-H, Lin P-H, Haga SW, Chen Y-R, Kranti A. A new electron bridge channel 1T-DRAM employing underlap region charge storage. IEEE J Electron Devices Soc. 2017;5(1):59–63.
84. Žurauskienė J, Yau C. pcaReduce: hierarchical clustering of single cell transcriptional profiles. BMC Bioinform. 2016;17(1):140.
85. Guo M, Wang H, Potter SS, Whitsett JA, Xu Y. SINCERA: a pipeline for single-cell RNA-Seq profiling analysis. PLoS Comput Biol. 2015;11(11):e1004575.
86. Campbell JN, Macosko EZ, Fenselau H, Pers TH, Lyubetskaya A, Tenen D, et al. A molecular census of arcuate hypothalamus and median eminence cell types. Nat Neurosci. 2017;20(3):484–96.
87. Lancichinetti A, Fortunato S. Community detection algorithms: a comparative analysis. Phys Rev E. 2009;80(5):056117.

88. Danon L, Díaz-Guilera A, Duch J, Arenas A. Comparing community structure identification. J Stat Mech. 2005;2005(09):P09008.
89. Schaeffer SE. Graph clustering. Comput Sci Rev. 2007;1(1):27–64.
90. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-driven phenotypic dissection of AML reveals progenitor-like cells that correlate with prognosis. Cell. 2015;162(1):184–97.
91. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics. 2015;31(12):1974–80.
92. Ding J, Shah S, Condon A. densityCut: an efficient and versatile topological approach for automatic clustering of biological data. Bioinformatics. 2016;32(17):2567–76.
93. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? In: Beeri C, Buneman P, editors. Database theory — ICDT'99. Berlin: Springer; 1999. p. 217–35. (Lecture Notes in Computer Science).
94. Alessandrì L, Arigoni M, Calogero R. Differential expression analysis in single-cell transcriptomics. Meth Mol Biol (Clifton, NJ). 2019;1979:425–32.
95. Finak G, McDavid A, Yajima M, Deng J, Gersuk V, Shalek AK, et al. MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. Genome Biol. 2015;16(1):278.
96. Van den Berge K, Perraudeau F, Soneson C, Love MI, Risso D, Vert J-P, et al. Observation weights unlock bulk RNA-seq tools for zero inflation and single-cell applications. Genome Biol. 2018;19(1):24.
97. Soneson C, Robinson MD. Bias, robustness and scalability in single-cell differential expression analysis. Nat Methods. 2018;15(4):255–61.
98. Zhou X, Carbonetto P, Stephens M. Polygenic modeling with Bayesian sparse linear mixed models. PLoS Genet. 2013;9(2):e1003264.
99. Khatri P, Sirota M, Butte AJ. Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.
100. Camp JG, Badsha F, Florio M, Kanton S, Gerber T, Wilsch-Bräuninger M, et al. Human cerebral organoids recapitulate gene expression programs of fetal neocortex development. PNAS. 2015;112(51):15672–7.
101. Huang DW, Sherman BT, Tan Q, Kir J, Liu D, Bryant D, et al. DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucl Acids Res. 2007;35(Web Server issue):W169–75.
102. Oron AP, Jiang Z, Gentleman R. Gene set enrichment analysis using linear models and diagnostics. Bioinformatics. 2008;24(22):2586–91.
103. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50.
104. Kim S-Y, Volsky DJ. PAGE: parametric analysis of gene set enrichment. BMC Bioinform. 2005;6(1):144.
105. Wu D, Smyth GK. Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. 2012;40(17):e133.
106. Ma Y, Sun S, Shang X, Keller ET, Chen M, Zhou X. Integrative differential expression and gene set enrichment analysis using summary statistics for scRNA-seq studies. Nat Commun. 2020;11(1):1585.
107. Weinreb C, Klein AM. Lineage reconstruction from clonal correlations. PNAS. 2020;117 (29):17041–8.
108. Jensen P, Dymecki SM. Essentials of recombinase-based genetic fate mapping in mice. In: Lewandoski M, editor. Mouse molecular embryology: methods and protocols [Internet]. Boston: Springer; 2014 [cited 2020 Oct 9]. p. 437–54. (Methods in molecular biology). https://doi.org/10.1007/978-1-60327-292-6_26.

109. Lu R, Neff NF, Quake SR, Weissman IL. Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding. Nat Biotechnol. 2011;29(10):928–33.
110. Biddy BA, Kong W, Kamimoto K, Guo C, Waye SE, Sun T, et al. Single-cell mapping of lineage and identity in direct reprogramming. Nature. 2018;564(7735):219–24.
111. Woodworth MB, Girskis KM, Walsh CA. Building a lineage from single cells: genetic techniques for cell lineage tracking. Nat Rev Genet. 2017;18(4):230–44.
112. Yang Z, Rannala B. Molecular phylogenetics: principles and practice. Nat Rev Genet. 2012;13 (5):303–14.
113. Kalhor R, Kalhor K, Mejia L, Leeper K, Graveline A, Mali P, et al. Developmental barcoding of whole mouse via homing CRISPR. Science [Internet]. 2018 [cited 2020 Oct 9];361(6405). https://science.sciencemag.org/content/361/6405/eaat9804.
114. Wagner DE, Klein AM. Lineage tracing meets single-cell omics: opportunities and challenges. Nat Rev Genet. 2020;31:1–18.
115. Pei W, Feyerabend TB, Rössler J, Wang X, Postrach D, Busch K, et al. Polylox barcoding reveals haematopoietic stem cell fates realized in vivo. Nature. 2017;548(7668):456–60.
116. Weinreb C, Rodriguez-Fraticelli A, Camargo FD, Klein AM. Lineage tracing on transcriptional landscapes links state to fate during differentiation. Science [Internet]. 2020 [cited 2020 Oct 9];367(6479). https://science.sciencemag.org/content/367/6479/eaaw3381.
117. Knoblich JA. Mechanisms of asymmetric stem cell division. Cell. 2008;132(4):583–97.
118. Rodriguez-Fraticelli AE, Wolock SL, Weinreb CS, Panero R, Patel SH, Jankovic M, et al. Clonal analysis of lineage fate in native haematopoiesis. Nature. 2018;553(7687):212–6.
119. Haber AL, Biton M, Rogel N, Herbst RH, Shekhar K, Smillie C, et al. A single-cell survey of the small intestinal epithelium. Nature. 2017;551(7680):333–9.
120. Clark SJ, Lee HJ, Smallwood SA, Kelsey G, Reik W. Single-cell epigenomics: powerful new methods for understanding gene regulation and cell identity. Genome Biol. 2016;17(1):72.
121. Schübeler D. Function and information content of DNA methylation. Nature. 2015;517 (7534):321–6.
122. Guo H, Zhu P, Wu X, Li X, Wen L, Tang F. Single-cell methylome landscapes of mouse embryonic stem cells and early embryos analyzed using reduced representation bisulfite sequencing. Genome Res. 2013;23(12):2126–35.
123. Smallwood SA, Lee HJ, Angermueller C, Krueger F, Saadeh H, Peat J, et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat Methods. 2014;11(8):817–20.
124. Farlik M, Sheffield NC, Nuzzo A, Datlinger P, Schönegger A, Klughammer J, et al. Single-cell DNA methylome sequencing and bioinformatic inference of epigenomic cell-state dynamics. Cell Rep. 2015;10(8):1386–97.
125. Rotem A, Ram O, Shoresh N, Sperling RA, Goren A, Weitz DA, et al. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. Nat Biotechnol. 2015;33(11):1165–72.
126. Kind J, Pagie L, de Vries SS, Nahidiazar L, Dey SS, Bienko M, et al. Genome-wide maps of nuclear lamina interactions in single human cells. Cell. 2015;163(1):134–47.
127. Cusanovich DA, Daza R, Adey A, Pliner HA, Christiansen L, Gunderson KL, et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. Science. 2015;348(6237):910–4.
128. Buenrostro JD, Wu B, Litzenburger UM, Ruff D, Gonzales ML, Snyder MP, et al. Single-cell chromatin accessibility reveals principles of regulatory variation. Nature. 2015;523 (7561):486–90.
129. Jin W, Tang Q, Wan M, Cui K, Zhang Y, Ren G, et al. Genome-wide detection of DNase I hypersensitive sites in single cells and FFPE tissue samples. Nature. 2015;528(7580):142–6.
130. Nagano T, Lubling Y, Stevens TJ, Schoenfelder S, Yaffe E, Dean W, et al. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. Nature. 2013;502(7469):59–64.
131. Tang X, Huang Y, Lei J, Luo H, Zhu X. The single-cell sequencing: new developments and medical applications. Cell Biosci. 2019;9(1):53.

132. Zhang L, Yu X, Zheng L, Zhang Y, Li Y, Fang Q, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. Nature. 2018;564(7735):268–72.
133. Bian S, Hou Y, Zhou X, Li X, Yong J, Wang Y, et al. Single-cell multiomics sequencing and analyses of human colorectal cancer. Science. 2018;362(6418):1060–3.
134. Zheng C, Zheng L, Yoo J-K, Guo H, Zhang Y, Guo X, et al. Landscape of infiltrating T cells in liver cancer revealed by single-cell sequencing. Cell. 2017;169(7):1342–1356.e16.
135. Guo X, Zhang Y, Zheng L, Zheng C, Song J, Zhang Q, et al. Global characterization of T cells in non-small-cell lung cancer by single-cell sequencing. Nat Med. 2018;24(7):978–85.
136. Ledergor G, Weiner A, Zada M, Wang S-Y, Cohen YC, Gatt ME, et al. Single cell dissection of plasma cell heterogeneity in symptomatic and asymptomatic myeloma. Nat Med. 2018;24 (12):1867–76.
137. Crinier A, Milpied P, Escalière B, Piperoglou C, Galluso J, Balsamo A, et al. High-dimensional single-cell analysis identifies organ-specific signatures and conserved NK cell subsets in humans and mice. Immunity. 2018;49(5):971–986.e5.
138. Villani A-C, Satija R, Reynolds G, Sarkizova S, Shekhar K, Fletcher J, et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. Science (New York, NY). 2017;356(6335):eaah4573.
139. Wilk AJ, Rustagi A, Zhao NQ, Roque J, Martínez-Colón GJ, McKechnie JL, et al. A single-cell atlas of the peripheral immune response in patients with severe COVID-19. Nat Med. 2020;26(7):1070–6.
140. Chen Y, Zheng Y, Gao Y, Lin Z, Yang S, Wang T, et al. Single-cell RNA-seq uncovers dynamic processes and critical regulators in mouse spermatogenesis. Cell Res. 2018;28 (9):879–96.
141. Vento-Tormo R, Efremova M, Botting RA, Turco MY, Vento-Tormo M, Meyer KB, et al. Single-cell reconstruction of the early maternal–fetal interface in humans. Nature. 2018;563 (7731):347–53.
142. Mironova V, Xu J. A single-cell view of tissue regeneration in plants. Curr Opin Plant Biol. 2019;52:149–54.
143. Gerber T, Murawala P, Knapp D, Masselink W, Schuez M, Hermann S, et al. Single-cell analysis uncovers convergence of cell identities during axolotl limb regeneration. Science [Internet]. 2018 [cited 2020 Oct 12];362(6413). https://science.sciencemag.org/content/362/6413/eaaq0681.
144. Sena G, Wang X, Liu H-Y, Hofhuis H, Birnbaum KD. Organ regeneration does not require a functional stem cell niche in plants. Nature. 2009;457(7233):1150–3.
145. Efroni I, Mello A, Nawy T, Ip P-L, Rahni R, DelRose N, et al. Root regeneration triggers an embryo-like sequence guided by hormonal interactions. Cell. 2016;165(7):1721–33.
146. Rahni R, Efroni I, Birnbaum KD. A case for distributed control of local stem cell behavior in plants. Dev Cell. 2016;38(6):635–42.
147. Cieślik M, Chinnaiyan AM. Cancer transcriptome profiling at the juncture of clinical translation. Nat Rev Genet. 2018;19(2):93–109.
148. Fan J, Slowikowski K, Zhang F. Single-cell transcriptomics in cancer: computational challenges and opportunities. Exp Mol Med. 2020;15:1–14.
149. Yuan Y, Lee H, Hu H, Scheben A, Edwards D. Single-cell genomic analysis in plants. Genes (Basel) [Internet]. 2018 [cited 2020 Oct 9];9(1). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5793201/.

Recent Advancement in NGS Technologies 26

Abhilek Kumar Nautiyal, Vishal Ahuja, Siddheshwar Kshirsagar, and Diptarka Dasgupta

Abstract

Rice is one of the most treasured grains across the globe for feeding human population. Rice genomics has taken giant strides in the direct use of fundamental scientific advances in the molecular mechanisms for agronomic traits and the use of various germplasm tools. In addition to multiple high-performance genome sequencing initiatives, Next-Generation Sequencing (NGS) is a crucial technique for the invention of domestication genes in crop flora and their wild relatives. This understanding will assist in speeding up the domestication of the latest plant species as seeds. Domesticated genotypes can be resequenced to classify low-diversity domestication areas. The sequence of whole genomic shotguns can collect species-specific data from similar wild species. This collected information may be used for the design of species-specific PCR primers. For example, maize, sugarcane, and eucalyptus have been used to capture genetic biodiversity for plant enhancement. Vast numbers of individuals can be speedily screened. The new genetic variants of related species are rapidly domesticated and efficiently identified and captured by NGS. Although the next-generation sequencing method is almost 10 years old, the informal way remains to classify highly parallel or high-performance sequencing methods that produce genome-scale

Abhilek Kumar Nautiyal and Vishal Ahuja contributed equally with all other contributors.

A. K. Nautiyal · D. Dasgupta (✉)
Biochemistry and Biotechnology Area, Material Resource Efficiency Division (MRED), CSIR-Indian Institute of Petroleum, Mohkampur, Dehradun, Uttarakhand, India

Academy of Scientific and Innovative Research (AcSIR), CSIR-HRDC Campus, Ghaziabad, India
e-mail: ddgupta@iip.res.in

V. Ahuja · S. Kshirsagar
Biochemistry and Biotechnology Area, Material Resource Efficiency Division (MRED), CSIR-Indian Institute of Petroleum, Mohkampur, Dehradun, Uttarakhand, India

M. K. Gupta, L. Behera (eds.), *Bioinformatics in Rice Research*,
https://doi.org/10.1007/978-981-16-3993-7_26

data or beyond. The introduction of these technologies has also increased the number of applications and procedures using genome-scale power sequencing. This chapter briefly reviews the recent NGS advancements in rice research and its achievements and also presents its bright future.

Keywords

Next-generation sequencing · Rice · Molecular breeding · Marker-assisted selection · Crop improvement

Abbreviations

BIL	Backcross inbred lines
BOLD	Barcode of life data
BP	Bac pool
BSA	Bulked-segregant analysis
CBC	Clone-by-clone
CNN	Convolutional neural network
COI	Cytochrome oxidase I
CSSL	Chromosome segment substitution lines
GBS	Genotyping by sequencing analysis
GO	Gene ontology
GWH	Genome warehouse
IL	Intercrossed lines
MAFB	Marker-assisted forward breeding
miRNAs	Mirnas
MITEs	Miniature inverted-repeat elements
ML	Machine learning
NGS	Next generation sequencing
PCG	Protein coding genes
PM	Physical map integration
QTL	Quantitative trait locus
RF	Random forest
RILs	Recombinant inbred lines
RISC	RNA-induced silencing complex
siRNAs	Small interfering rnas (sirnas)
SNP	Single nucleotide polymorphism
SV	Structural variations
SVM	Single vector method
WGS	Whole genome shotgun

26.1 Introduction

In contrast to temperate countries, tropical countries are typically underdeveloped due to weak farm productivity [1]. The decline of farming productivity in the tropics is attributable to a lack of technological development and various abiotic and biotic factors. The green revolution in the late 1960s, which uplifted many developing countries from drought, contributed to an integrated strategy that uses enhanced cultivars, fertilizers, and pesticides. Hybridization was an effective breeding method in the 1960s and 1980s that created many high-yielding crop varieties [2]. In the 1990s, there was a better awareness of genetics and technical advances in transgenic crops [3]. Compared to traditional breeding methods in which undesirable genes could also be transferred, transgenic technology was preferred since it allowed the transfer of one or extra ideal genes. Several transgenic varieties of insect-resistant cotton, herbicide-tolerant soy, and virus-based papaya were marketed [3]. However, currently, genetically modified (GM) crops are controversial since they induce food allergies and spread antibiotic resistance to intestinal bacteria [3]. The ecological imbalance created by the gene flows with insecticidal proteins and herbicide resistance genes from transgenic plants into wild varieties worries environmentalists. Crop improvement is advocated by using gene pools of related species for enhancement in the fields of agronomy, energy, and biomaterial production [4].

Next-generation sequencing (NGS) is almost 10 years old. However, the exploration of NGS technologies has been started recently for various applications. The technique performs a study of the entire genome to establish phenotypic, genetic basis variations [5] across the species. Although it allows reasonably sized DNA fragments to be sequenced, it is most useful as a quick sequence read. The recent development in NGS platforms and methodologies for various applications has tremendously increased the speed and accuracy of genome-scale sequencing in molecular biology data analysis. The extent of data is developing because the generation adapts to it. NGS is being implemented in crop structures and offers an opportunity to analyze genetic ranges in vegetation and their wild family on a far larger scale than the preceding technology [4]. It helps even the most complex genomes of plants to be addressed [6]. The discovery of new useful variants can also be extended to NGS (Table 26.1). NGS allows the rapid expansion of genomic studies into the non-version species investigations [23]. Before the advent of NGS, the analysis of variation in plant genomes was limited to applying series-based selection [24]. To be able to exhibit different techniques in numerous meals applications (rice), energy (sugarcane), and wooded area (Eucalyptus) species, examples of the usage of NGS inside the discovery of this useful variant to domesticate new genes or species could be mentioned.

Table 26.1 Advanced mapping populations reported between 2007 and 2012 and traits mapped between wild *Oryza* species as the donor parent and an *O. sativa* cultivar as the recurrent parent

Population	Donor parent (accession)	Recurrent parent	Traits introgressed/ mapped	References
BIL	*O. glaberrima* (Tog5675)	*Indica* (IR64)	BPH resistance (*Bph1*)	[7]
BIL	*O. glaberrima* (IRGC96717)	*Japonica* (WAB56-104)	Drought resistance, early vigor	[8]
BIL	*O. glaberrima* (IRGC103544)	*Indica* (Milyang 23)	Yield and yield components	[9]
BIL	*O. nivara* (IRGC105444)	*Japonica* (Koshihikari)	Hybrid breakdown locus (−*hbd 1*(t))	[10]
BIL	*O. rufipogon* (W630)	*Japonica* (Nipponbare)	Drought tolerance	[11]
BIL	*O. rufipogon* (IRGC105491)	Tropical *japonica* (Jefferson)	Early flowering	[12]
BIL	*O. rufipogon* (YJCW)	*Indica* (93-11, restorer line)	Yield-related traits	[13]
BIL	*O. rufipogon* (IRGC105491)	*Indica* (IR64)	Yield and yield components	[14]
BIL	*O. rufipogon* (YJCWR)	*Indica* (TeQing)	Yield and yield components	[15]
BIL	*O. glumaepatula* (RS-16)	*Indica* (BG90-2)	Grain yield, cooking quality	[16]
BIL	*O. minuta* (IRGC101141)	*Indica* (IR31917-45-3-2)	BPH resistance	[7]
BIL	*O. brachyantha* (IRGC101232)	*Indica* (IR56)	Bacterial blight	[7]
BIL/NIL	*O. longistaminata*	*Indica* (RD23)	Pollen/spikelet fertility, plant height	[17]
CCSL	*O. rufipogon*	*Indica* (Teqing)	Small grain panicle and dwarfness	[18]
CSSL	*O. glaberrima*	*Japonica* (Koshihikari)	Glabrous gene	[19]
CSSL	*O. glaberrima*	*Japonica* (Wuyujing-7)	Spreading panicle	[20]
IL	*O. glaberrima* (Tog5681)	*Indica* (IR64)	Drought tolerance	[21]
IL/BIL	*O. nivara* (IRGC105444)	*Japonica* (Taichung 65)	Pollen sterility gene (S27-nivs)	[22]

26.2 Rice: A Staple Food Crop

Although primarily consumed in Asia, rice is also a substantial food source in many nations such as Africa and South America. The average per capita intake rice in the Indian diet is about 72.2 and 48.2 kg annually in rural and urban areas, respectively.

It is being cultivated as a staple crop for more than 2000 years. Crossbreeding and selection carried out by farmers and breeders to fit the particular local conditions has led to thousands of cultivars' production. The full rice genome sequence based on the Nipponbare cultivar has thus resulted in the broad characterization of other *japonica* cultivars, including the well-grown and elite *Koshihikari* cultivar [25] known for their excellent quality. Since the beginning of agriculture, modern-day crops were subjected continuously to genetic selection; hence, their genomes preserve the imprints of killing due to a combination of natural and synthetic selection approaches. Rice was domesticated in China about 9000 years ago [6, 26, 27]. For that reason, it has undergone sizeable geomorphological and physiological alterations via human's synthetic assortment to become one of the most vital cereal vegetation [28]. Apart from its use as a food crop, rice has been extensively used in genetic research due to its small genome size of only 370 million bases (Mb) [28].

26.3 Unwrapping the Genetic Structure of the Rice

In 2002, the rice genomes (Fig. 26.1), were correctly sequenced using the whole genome shotgun (WGS) sequencing method [29, 30]. The same *japonica* cultivar was also sequenced by two other private companies: Syngenta, Switzerland, and Monsanto, USA. The genome project initiated by the Chinese Superhybrid Rice Genome Project (CSRGP), on the other hand, has expanded beyond genome sequencing and continued to pursue hybrid studies on rice [31]. Full genomes were released from both rice subspecies in 2005, covering ~95% of the 389-Mb genome [32] (Table 26.2). 37,544 protein-coding genes linked to the non-transposable-element (nTE) have been predicted from the genome, with 34.79% of the genomes being TEs. High-quality genome assemblies with annotation provide excellent insight into rice genomics, evolution, and biology and aid in cloning and molecular studies. NGS has allowed rapid sequencing of the rice genome, and currently the database is updated to ~3000 rice accessions. The Rice Genomes project (3K RGP 2014) has served as an extensive data source for studying rice varieties information [4]. One of the major issues in the plant genome analysis is TE's location, whether within or outside a gene. TE insertions and dynamics play significant roles within narrow taxonomy groups in genome evolution and in the history of related genes [33] and have an enormous impact on higher taxonomy's genome evolution species. The second issue is related to multiplied mutation mechanisms; in this situation, transcript-centric effective GC gradients emerge as glaring in the Gramineae genomes [34, 35]. The GC gradients within the transcription path are not typical, shared only by the grass family of vegetation and heat-blooded vertebrates [34]. Another issue is the polyploidy and ancient whole-genome duplication (WGD) of the plant genome. Whole-genome duplication is a process of genome doubling that supplies raw genetic substances and increases genome complexity. It has currently been discovered that WGD and the next destiny changes of duplicated genes could facilitate phenotypic evolution.

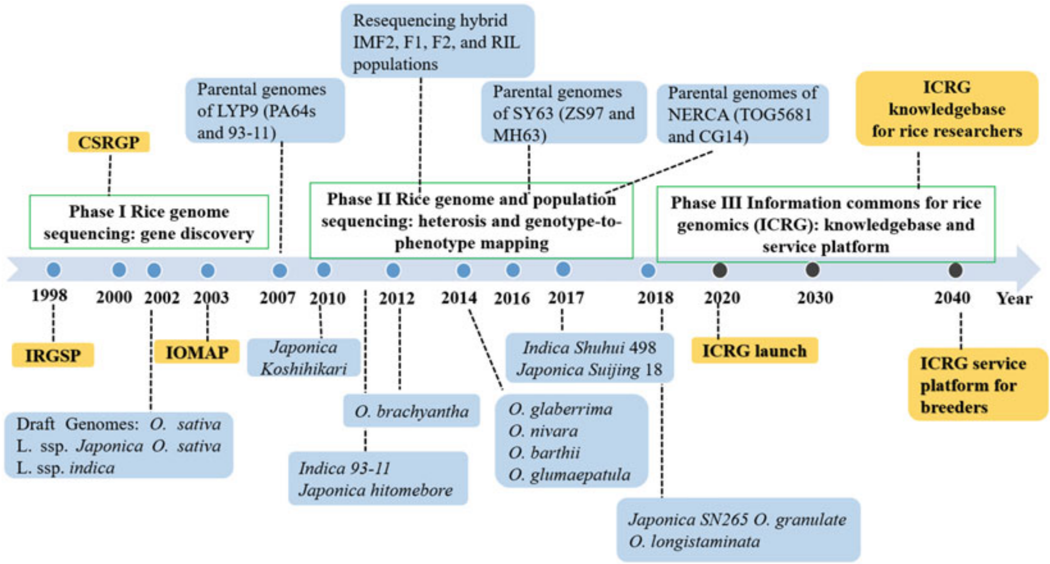

Resequencing hybrid IMF2, F1, F2, and RIL populations
Parental genomes of LYP9 (PA64s and 93-11)
Parental genomes of SY63 (ZS97 and MH63)
Parental genomes of NERCA (TOG5681 and CG14)
ICRG knowledgebase for rice researchers
CSRGP
Phase I Rice genome sequencing: gene discovery
Phase II Rice genome and population sequencing: heterosis and genotype-to-phenotype mapping
Phase III Information commons for rice genomics (ICRG): knowledgebase and service platform
1998
2000
2002
2003
2007
2010
2012
2014
2016
2017
2018
2020
2030
2040
Year
IRGSP
IOMAP
Japonica Koshihikari
Indica Shuhui 498 *Japonica Suijing* 18
ICRG launch
ICRG service platform for breeders
Draft Genomes: *O. sativa* L. ssp. *Japonica O. sativa* L. ssp. *indica*
O. brachyantha
Indica 93-11 Japonica hitomebore
O. glaberrima
O. nivara
O. barthii
O. glumaepatula
Japonica SN265 O. granulate
O. longistaminata

Fig. 26.1 The timeline of rice genomics and geographical distribution of sequenced rice genomes for four decades. The solid circles suggest past rice genomics events, including section I for rice genome sequencing and section II for rice genome and population sequencing. The open circles indicate projected future events for phase III

Table 26.2 Sequencing status of 16 *Oryza* genomes and *Leersia perrieri* (outgroup species) based on their Genome size (in increasing order)

Sl. No	Species (genome type)	Genome size (Mb)	Sequencing method	Status
1.	*O. brachyantha* (FF)	260	WGS/PM	Reference genome
2.	*L. perrieri* (outgroup)	323	WGS/PM	Draft genome
3.	*O. longistaminata* (AA)	352	WGS	Draft genome
4.	*O. glaberrima* (AA)	354	BP	Reference genome
5.	*O. sativa* ssp. *indica* (AA)	400	WGS	Draft genome
6.	*O. sativa* ssp. *japonica* (AA)	400	CBC/PM	Reference genome
7.	*O. barthii* (AA)	411	WGS/PM	Reference genome
8.	*O. punctata* (BB)	423	BP/WGS/PM	Reference genome
9.	*O. meridionalis* (AA)	435	WGS/PM	Draft genome
10.	*O. rufipogon* (AA)	445	WGS	Draft genome
11.	*O. nivara* (AA)	448	BP/WGS/PM	Assembly IP
12.	*O. glumaepatula* (AA)	464	WGS/PM	Assembly IP
13.	*O. eichingeri* (CC)	650	WGS	Sequencing IP
14.	*O. rhizomatis* (CC)	650	WGS	Sequencing IP
15.	*O. officinalis* (CC)	653	WGS/PM	Sequencing IP
16.	*O. granulata* (GG)	862	WGS/PM	Sequencing IP
17.	*O. australiensis* (EE)	960	WGS/PM	Sequencing IP

It has been inferred that high-quality sequence data and a close reference genome are essential for gene-level comparative evaluation for its landraces and subspecific and gene-precise members. The updated 2012 Nipponbare Reference Genome assembly (Os-Nipponbare-Reference-IRGSP-1.0) was a fantastic instance of the way to use quite many technologies to attain concerted efforts, from optical mapping to CBC mapping, from clone-based assembly and heavy WGS coverage to Roche GS FLX long-examine sequences, to short-study sequences for the Illumina Genome Analyzer II assembly [36]. NGS has also been extensively used to categorize domestication gene homologues in wild-type rice [37], which we will discuss in the section below.

26.3.1 Domestication of Rice: An NGS-Based Approach

Rice became a primary crop to be used as a reference genome in 2005 [29, 32]. It enabled researchers to use NGS as a tool to pick out variations in the genome and the gene pool among different *Oryza* species. Despite these significant efforts, knowledge of rice's domestication stays tough, in all likelihood due to a complicated foundation involving a few gene movements among the *japonica* and *indica* species [38, 39]. The whole-genome resequencing has shown that both *Japonica* and *Indica* rice possibly share mutual areas of low variety primarily because of genetic transfer from one population to another [40]. Many low variety areas only own a single domestication gene. This gene is associated with low diversity domesticated rice to shatter upright growth dependency and white grain pericarp [41]. These studies may offer treasured information on rice domestication and the connections between rice populations in contemporary history. NGS from wild rice populations has been used these days to explain the distinction between wild rice populations in Asia and Australia [42]. This study used the complete genome sequence of chloroplasts observed inside the shotgun series of general plant samples of DNA [43]. The *Oryza rufipogon* turned into domesticated variety in Asia. The Australian network of *O. rufipogon* appears to show closeness to *O. meridionalis* than to the Asian population's *O.rufipogon.* Each of these varieties is a genetic asset for rice development. NGS of wild rice relatives' nuclear genome will apprehend the genetic variant of capability advantage for rice improvement and clarify phylogenetic relationships. Hybridization and heterosis contribute drastically to higher productivity and sustainable food manufacturing. In properly studied structures that include maize, NGS has been used to assess and forecast heterosis [44]. Gene expression styles in hybrids [45] with NGS have also been evaluated. The viable participants to rice miRNA and little RNAs to heterosis in hybrid rice have also been demonstrated [46]. Complete genome data and transcriptome collection may be used to compare rice varieties' genomes and transcriptomes to select individuals with a hybrid functionality. It enables the technology to choose new crop species as an efficient hybrid crop alternative.

Genus *Oryza* encompasses more than 21 wild varieties categorized into four species complexes: *O. granulate, O. officinalis, O. ridleyi,* and *O. sativa*—species based on their morphological and genomic characteristics. Available literature has emphasized rice's origin from wild grasses that have slowly become an integral part of human food and civilization with the domestication of *O. barthii* and *O. rufipogon* in Africa (3500 years ago) in the Asian subcontinent (around 10,000 years ago). Asian rice further laid the foundation for *indica* and *japonica* [47, 48]. Previous researches have restricted the exploration of the domestication process and identification of critical factors.

In recent years, the evolution of NGS took domestication to the molecular level. NGS has revealed the domestication's fundamental basis, identified the relativeness and closeness among new and close related varieties, and discovered domestication genes in crop efficiently [49]. Rice genome foot-printing traces three different epicenters for rice domestication, contributing to current Asian rice gene pools. The rice populations from southern Yangtze Valley China represent the *japonica* rice gene pool, while those from Brahmaputra Valley and Indo-China become a source for *indica* rice gene pool [47, 50, 51]. Initially, morphological characteristics like shattering, seed dispersal, seed yield, seed size, mating habits, tiller number, pericarp structure, pericarp color, and dormancy period were identified as domestication keys. NGS has made it possible to link domestication with genes and quantitative trait locus (QTLs). WGS of *O. rufipogon, indica,* and *japonica* population identified low diversity regions and domestication genes for shattering upright growth habit and white grain pericarp within this region [40, 41, 52]. reported the disruption of the Black hull4 (*Bh4*) gene in *Oryza rufipogon* and *Oryza nivara.* It has been further revealed that the *Bh4* gene, located on chromosome IV, has become nonfunctional (due to a loss of 22 bp sequence), which altered the seed color of *Oryza sativa* to become white instead of black. Comparative analysis of Illumina sequence reads of Dongxiang wild rice (DXWR) and Nipponbare whole genome revealed 2536 structural variations (SVs) in Nipponbare, mainly represented by 1568 deletions with 731 QTLs, especially linked to crop yield, vigor, anatomy, quality, biochemical, and development. Nipponbare also acquired genes associated with plant height, spikelet number, panicle number, leaf senescence, panicle length, biomass yield, seedling vigor, leaf width, and tiller number during domestication [53]. Whole-chloroplast genome analysis confirmed the origin of both Asian and Australian rice from *Oryza rufipogon,* but Australian populations are closer to *O. meridionalis* than Asian populations [43, 54]. Initially, it was suggested that *japonica* and *indica* rice originated from the same lineage due to evolutionary selection. The recent archaeogenetic analysis of modern-age rice and ancient rice grains of the Liangzhu Period, collected from China, suggested that both O. *japonica* and O. *indica* have different maternal lineages [55].

26.3.2 NGS-Based Identification of New Varieties: Improvement of DNA Barcoding

A short-standardized collection of DNA 400–800 bp is used as DNA barcodes to discover the latest plant species, previously acknowledged plant description, and classification. DNA barcoding using a short DNA fragments system for rapid identification of plants was first proposed by Hebert et al. [56]. Hollingsworth et al. [57]. suggested the matK+rbcL two-locus mixture system as a core barcode for land plants identification. The unique molecular markers rbcL, matK, trnH-psbA, and ITS2 were used as DNA barcodes. DNA barcoding system has provided an efficient tool using selective markers for plant identification and conservation of the world’s biodiversity.

The barcode of life data system (BOLD; www.boldsystems.org) platform was used by many researchers to match reference DNA barcodes with the specimen for identification [58]. However, it has been observed that the BOLD data system has many limitations for the unsampled specimens. Despite these sanger sequencing and other identification, systems-based DNA barcoding for some closely related species and different advanced application DNA barcode is not enough [59]. To address this, Wilkinson et al. [60] reported NGS-based DNA barcoding as the possible solution for unsampled specimen cases. Next-generation sequencing has many advantages over Sanger sequencing-based in the DNA barcoding system [60]. The concept of sanger sequencing and NGS are similar. The main difference varies in sequencing volume. In sanger sequences, a single DNA fragment is sequenced at a time, while in

NGS, sequencing of lots of genes can be done. NGS has many advantages over sanger sequencing like high sensitivity, low detection limit, and rare genes with detailed sequencing. Shokralla et al. [61] reported the first de novo next-generation sequencing for Lepidopteran specimen of mitochondrial gene cytochrome c oxidase subunit I (COI) 189 of 190 using 454 pyrosequencing platform. It has been studied that in the initial stages, the NGS platform also has many limitations for barcoding applications. However, the NGS Illumina platform has resolved the many shortfalls and benefitted many researchers for DNA barcoding applications by pairing end sequencing of two overlapping COI amplicons and merging reads to assemble full-length barcodes [61].

The conventional plant DNA barcodes (matK, rbcL, psbA-trnH, and ITS) have been used to identify the rice variants. Li et al. [62] proposed the advanced DNA super barcode system for rice identification with the complete application package. A DNA super barcode provides the complete information to distinguish between the species of interest in the form of a complete genome or parts of a genome. It constitutes the complete chloroplast or mitochondrial genomes, and their mixtures and assemblies of single nucleotide polymorphisms represent the DNA super barcodes. The application of DNA super barcodes can help to identify haplotypes and seeds of closely related species. The common chloroplast gene fragments cannot differentiate the A and C haploid genome types to address these DNA super barcodes with a complete chloroplast genome [63]. Zhang et al. [63] reported the detailed study description between the conventional markers as DNA barcode, rice-specific barcodes, rice- specific nuclear DNA barcodes, and super DNA barcodes (entire chloroplast genome) using NGS techniques. Chloroplast genomes as a DNA barcodes sequenced using NGS-based Illumina platform as a useful tool for rice variety discrimination have been proposed by Song et al. [64].

26.4 Machine-Learning in Gene Identification and Annotation

The advent of computation technology, coupled with sequencing platforms and bioinformatics tools, has made genome assembly and sequence analysis a relatively easy task. However, the functional annotation of the genomic regions is still a challenge [65]. Even for model organisms, such as *Arabidopsis thaliana,* ~20% of the predicted genes could not be assigned any functional role and elucidate their role in biological pathways [66]. The high-volume sequencing data generated by numerous scientists and researchers worldwide is only adding to the already existent challenge to derive biological interpretation of the discovered sequences. Machine-learning (ML) has become a powerful and popular technique to analyze high throughput data with considerable background noise to tackle gene function discovery in recent years.

ML algorithms are broadly classified into two categories, namely, the supervised and unsupervised methods. A supervised learning model utilizes a training dataset to deduce the plausible function, which can be further used to analyze test samples. In an unsupervised form, the dataset does not contain any predefined label or class typically used to discover hidden signal patterns within the data. This learning method is mainly applied to clump the big dataset into clusters on which supervised (trained) algorithms can be run to classify new samples [67]. Supervised machine-learning approaches, such as the Single vector method (SVM), Random forest (RF) technique, etc., have been developed to detect genomic features, such as protein-coding genes (PCG), miRNAs, and non-coding RNAs. They either use the binary classification setting of "true" or "false," which implies that the genome contains any coding genes or not. Alternatively, machine-learning methods that approach using multi-classification methods extract the feature of genes, transposons, and miRNA. The Hidden Markov Model (HMM)-based tools (Augustus, SNAP, etc.) founded on the Artificial Neural Network performs best for structural feature prediction problems and can accurately predict the exonic regions, splice sites, UTR's, and regulatory regions within the genome. Ben-Hur et al. [67] reported that the SVM-based approach using the mGene software could incorporate heterogeneous datasets to detect genic regions, such as the transcription initiation site and the splicing regions with higher precision HMM [67]. Gan et al. [68] reported that gene expression information coupled with sequence-based prediction could be successfully used to analyze multiple *Arabidopsis* plants' genomes and differentiate between gene homologues using the data heterogeneity as a modifier.

Once the coding and non-coding segments of a rice genome are identified, machine-learning approaches can provide greater insight into the expression of annotated genes, the location of proteins within the cell, and their interactions. Scientists have crucially analyzed the promoter regions of up- and down-regulated genes across the sequences deposited in the databases to identify crucial motifs using the SVM. These motifs are used to form a training set to predict the up or downregulation in unknown gene sets [69]. alternatively devised a unique strategy for gene expression analysis. Using a deep-learning method such as a Convolutional neural network (CNN) with promoter and terminator sequences, the authors achieved ~85% accuracy in predicting whether a gene was getting expressed or not under different treatment conditions in maize [69]. The authors concluded that the 3′-UTR was highly informative in providing the information regarding transcript abundance. This learning method may be extended with modifications for gene expression analysis in non-model organisms. Subcellular localization can be predicted using N terminal regions within the protein sequences using ANN techniques. For example, SignalP v 5.0 utilized a dataset of known proteins obtained from Expasy Prosite as a training set and was tested to predict signal motifs on unknown sequence sets to identify their location within the cell. The technique has varied prediction accuracy for different cellular locations and highlighted the importance of a sufficient training dataset for prediction accuracies, which involves considerable time and resources. Protein–protein interaction has been extensively studied for model organisms such as *A. thaliana* using ANN and RF-based methods with prediction accuracy of over 0.95 [70]. The training parameters have been applied to predict protein–protein interactions localized within maize [71]. In his publication, Zhu et al. [72] identified two different sets of proteins that were either interacting physically or functionally using the SVM approach with a prediction accuracy of more than 0.85 [72].

Prediction of biological functions, such as categorization based on Gene Ontology (GO) and gene classification based on pathways, has been carried out by machine-learning with considerable success. Typically, GO analysis is performed by sequence similarity-based annotation. However, recent research has shown that machine-learning-based algorithms provide better results than the traditional approach. Studies demonstrated that, by utilizing structural information from the protein database coupled with gene expression analysis, GO for unknown genes could be predicted with high precision. Also, for proteins that share similar motif patterns, the machine-learning software could assign a biological function to the protein based on the motif's function. For predicting metabolic pathways, a combination of inputs starting with gene to protein structure, properties, and homology is required to associate a gene to a pathway and explain its activity [73]. reported significant improvements over classical non-ML-based prediction tools by utilizing several features. However, the dataset, particularly the seed for the ML-based approach, is still limited. Extensive data curation needs to be undertaken to associate them with the enzymes and the reactions they catalyze. An alternative approach would be to combine transcriptome and metabolomics data to integrate them into an RF framework to identify novel pathway genes.

Despite the recent advances, machine-learning-based approaches are still limited due to the scarcity of datasets. Lack of available positive training datasets is one of the significant impediments which results in inaccurate predictions. With the growing databases, for ML-based approaches to be successful, a strong foundation of data validated by wet-lab experiments is essential.

26.5 Development of Novel Hybrid Crops

NGS fueled the establishment on single nucleotide polymorphism (SNP) for the identification of genes and plant species with or without any reference genome (orphan species). Also, it enabled gene expression analysis, development of genetic resources, large sequences assemblies, genetic improvement, and so on [74–76]. In the upcoming section, we will discuss the role of NGS technology and tools that have been explicitly used for rice improvement:

26.5.1 Biotic and Abiotic Stress Tolerance

Environmental conditions like submergence, extreme temperature variations, salinity, and pathogenesis significantly affect crop survival and productivity. Its influence can be seen clearly in cellular functions like transcription, translation, and metabolism. In plants, two types of sRNA moieties have been reported with distinct biogenesis, structure, and function, i.e., microRNAs (miRNAs) and small interfering RNAs (siRNAs). miRNAs mainly form RNA-induced silencing complex (RISC), while some miRNAs are also known to play regulatory roles and stress responses. To determine the role of miRNAs in stress tolerance, the rice population was segregated into four groups and cultivated under control conditions and drought, cold, or salt stress. For analysis, vegetative tissue, that is, such as inflorescences, was collected, and separate small RNA libraries were constructed by Illumina deep-sequencing technology. Out of 227 miRNAs, 62 miRNAs were validated from pre-published literature and databases like DCL1, DCL3, and RDR2 RNAi lines 43 miRNAs were identified that govern stress tolerance rice (18 droughts, 15 cold, and 10 salt stress). Besides, 80 miRNAs were also identified that originated from transposable elements, including miniature inverted-repeat elements (MITEs) [77]. A sudden drop in temperature and cold environment lowered crop productivity and survival. Recombinant inbred lines (RILs) were established with a cross between a cold-tolerant and cold-sensitive variety to trace the responsible QTL for cold susceptibility and tolerance, Dongnong422 and Kongyu131, respectively. A novel QTL *qPSST6* on 28.4 cM intervals on chromosome 6 was identified by integrated application of bulked-segregant analysis (BSA) and next-generation sequencing (NGS) technology (Seq-BSA). Within the segment, two genes, LOC_Os06g39740 and LOC_Os06g39750, were identified to control plant response to cold; however, LOC_Os06g39750 have a higher response rate under cold stress [78].

The plant's ability to reprogram transcriptional networks and associated factors like NAM, ATAF1-2, and CUC2 (NAC) to cope with stress was inferred from the modulation of drought-responsive transcription factors in rice. OsNAC14 is reported as a drought-responsive transcription factor in rice, usually expressed during the meiosis stage. However, its expression was greatly influenced by drought, high salinity, ABA, and low temperature in leaves [79]. demonstrated that the overexpression of the OsNAC14 gene led to induced drought tolerance in the vegetative stage and improved DNA damage repair with strigolactone biosynthesis. ERF family transcription factor *OsLG3* was also reported to participate in stress response in rice (*Oryza sativa*) under drought stress, which was earlier reported to determine rice grain length without affecting grain quality. Suppression and overexpression of *OsLG3* determine the suppression and tolerance of rice plants to drought. It was also reported that *OsLG3* involves reactive oxygen species scavenging system to improve drought stress tolerance [80]. Gene pyramiding approach has been implemented to construct a resistant rice variety against gall midge (insect), blast (disease), submergence, and salinity through MAS (Fig. 26.2). For example, an *indica* rice variety "Lalat" with four bacterial blight resistance genes *Xa4, Xa21, xa13*, and *xa5* (CRRI, India, Annual Report 2011-12) was selected as a recurrent parent (Table 26.3).

Marker-assisted forward breeding (MAFB) methods were used to develop resistant rice varieties for high grain yield, resilient for abiotic and biotic stresses, including blast, bacterial leaf blight, brown planthopper, gall midge, and drought tolerance. Eleven genes/QTLs: *Pi9* for against blast, *Xa4, xa5, xa13, Xa21* for bacterial leaf blight, *Bph3, Bph17* for brown planthopper, *Gm4, Gm8* for gall midge, and *qDTY1.1, qDTY3.1* for drought tolerance were targeted to prepare seven introgression lines (ILs). The performance of all seven lines was superior to

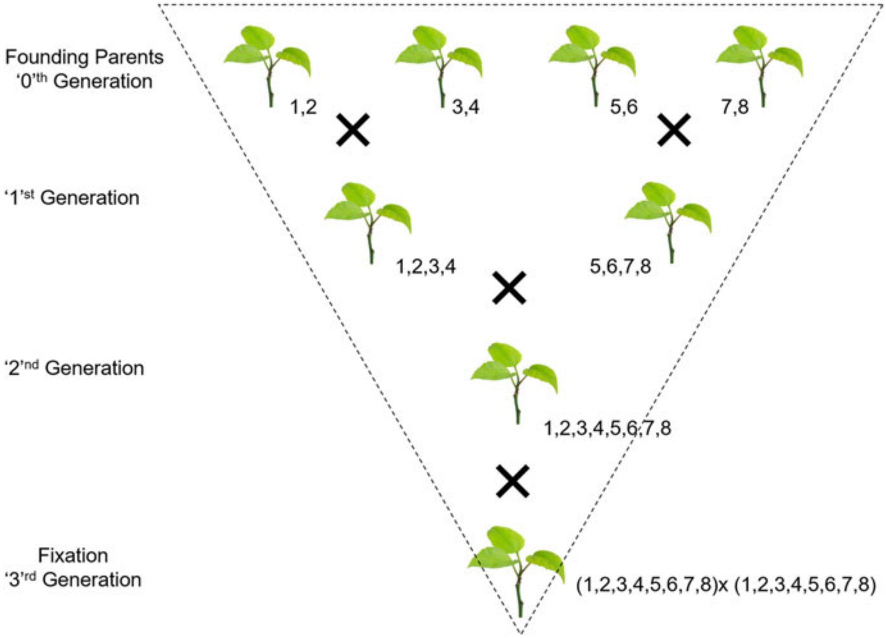


Fig. 26.2 Gene pyramiding approach to introduce and express multiple genes responsible for different characters like various biotic and abiotic stress resistance in a single plant from multiple parents

the respective recurrent parents (Swarna+drought) in terms of grain quality and high amylose content (AC) (23–26%). Among these seven lines, three lines including, viz., IL1 (*Pi9*+ *Xa4*+ *xa5*+ *Xa21*+ *Bph17*+ *Gm8*+ *qDTY1.1*+ *qDTY3.1*), IL6 (*Pi9*+ *Xa4*+ *xa5*+ *Xa21*+ *Bph3*+ *Bph17*+ *Gm4* + *Gm8*+ *qDTY1.1*+ *qDTY3.1*) and IL7 (*Pi9*+ *Xa4*+ *xa5*+ *Bph3*+ *Gm4*+ *qDTY1.1*+ *qDTY3.1*) exhibited resiliency against multiple abiotic and biotic stresses in both glasshouse and field trails. Trials confirm the superiority of ILs under both stressed and non-stressed conditions; however, under drought conditions, the yield advantage extended up to 1.0 t ha^{-1} without affecting grain quality [92].

26.5.2 Quality and Yield Improvement

In the current situation, an increase in crop yield is necessary to satisfy society's hunger. Identification of genetic structure, sequences, SNPs, etc., are advantageous for crop improvement and molecular breeding. It will help to understand the expression level of genes, regulatory sequences' role, and response pattern for different environmental/growth conditions. In recent efforts, Genotyping by sequencing analysis (GBS) revealed the role of *GS3* and *GIF1* genes in governing the grain size and yield after the screening of 1,07,140 Indels and 82,59,639 SNPs

Table 26.3 Molecular markers developed in rice for yield improvement

Molecular marker	Gene/QTL/SNP	Encoded protein	Trait controlled	Citation
SSR/InDel markers	*Pi54* and *Pita*	–	Blast resistant	[81]
qGW8F/R POT1F/R-POT10F/R	*OsSPL16/qGW8*	–	Grain quality	[82]
–	*OSAGO17*	Argonaute (AGO)' regulatory proteins	Higher yield and growth	[83]
RM3572	LOC_Os08g01490, and *LOC_Os08g01680*	–	Grain quality	[84]
78 SSR and STS markers	*Pi54, Pi1* and *Pita*	(a) α-amylase; (b) Triosephosphate isomerase; (c) 19 kDa globulin; (d) *S*-(+)-linalool synthase	Blast resistant	[85]
RM6970 and M1	*Gn1a/OsCKX2*	Cytokinin oxidase/ dehydrogenase	Grain quality	[86]
GW8	*GS3* and *OsSPL16*	Squamosa promoter binding protein-like 16	Grain quality	[87]
HGW	*Xa13*	Ubiquitin-associated domain protein	Grain weight	[88]
–	*LAZY*1	Specific herb protein	Panicle	[89]
–	*OsPIN2*	Auxin efflux transporter		
–	*LAX2*	Nuclear protein with a plant-specific conserved domain	Panicle	[90]
RM22475 and RM5556, Ind8-47 and Ind8-15	*DTH8/Ghd8*	OsHAP3 subunit of a CCAAT-box-binding protein	Panicle	[91]

from 16.3 × 10^7 reads [93]. Some hybrid rice varieties have exhibited higher yield and delayed flowering. Around 2000 F_2 generation progenies, obtained from two hybrids with a heading date of ~130 days, were segregated into eight pools based on the delayed flowering pattern. The comparative whole-genome analysis identified responsible genes *Hd1* (Heading date1) and *Ghd8/DTH8* (grain number, plant height, and heading date 8) on chromosomes 6 and 8 for delayed flowering. Besides, delayed flowering, *Ghd8/DTH8*, also contributes to higher grain number and product yield when overexpressed [94]. Another class of regulatory proteins, Argonaute (AGO), was overexpressed in reproductive tissue, specifically among the non-dicot AGO1 clade members in monocots, resulting in higher crop yield and robust growth, higher-fertility, and panicle length [83]. Bommisetty and colleague mapped major genomic regions/QTLs for grain weight on chromosomes 1, 7, and 8, viz., *qGW1* (35–40 Mb), *qGW7* (10–18 Mb), and *qGW8* (2–5 Mb), respectively. However,

qGW8 was the dominant factor among the three, which occupied two candidate genes, LOC_Os08g01490 (Cytochrome P450) and LOC_Os08g01680 (WD domain, G-beta repeat domain-containing protein). These genes were highly expressed in reproductive organs during grain development [84]. Later, the same QTL *OsSPL16/qGW8* was engineered through CRISPR/Cas9 to enhance grain size. Mutants were reported to have higher crop yield without affecting other agronomic traits. The change in genetic structure resulted in upregulation and downregulation of 33 and 11 proteins [82].

26.6 Commercialization and IP Protection: A Case Study

In-plant manufacturing and processing of food or biomaterials identification of plant genotypes may be tremendous. The conservation of intellectual property linked to plant genotypes often includes discriminatory identification tools. NGS provides a new level of protection to distinguish essential genes that decide functional characteristics [41]. Recently, the cost-effective approach for identifying plants has been demonstrated by NGS, with a useful technique being to analyze the genome sequence of chloroplast from the entire genome of shotgun sequencing of plant DNA [43]. Nuclear sequences decided on at some stage in domestication can also protect newly shaped crop species with NGS to efficiently monitor germplasm protection. Transcriptome sequencing of barley genotypes was used to discover many SNPs that differentiate between the genotypes and define genetic variants in the genotypes.

Similarly, plants like barley show a genetic variety of the species because of the breeding strategies utilized in producing the assortment's basis population. NGS determines even minor variants based on SNPs. NGS advocates this variety and any shifts in crop production over the years because of genetic waft. NGS era can enlarge into wild populations and facilitate the introgression of specific genome regions with desirable traits. The dry barley resistance becomes related to wild barley's distinct chromosome zones (*Hordeum spontaneum*) [95]. NGS needs to characterize the component in the drought-accepting domesticated genotypes produced with the aid of wild barley germplasm.

Genetically modified beta-carotene-containing rice, generally known as Golden Rice (GR), is no longer being delivered in any area. Vitamin A deficiency (VAD) has been developed in low-income rice purchasers but wishes to be advanced and checked before being implemented into farmer's fields. Due to preliminary publicity, GR is the most well-known biofortification attempt with modern biotechnology. An incredible technical fulfillment is the successful layout of the carotenoid biosynthetic pathway (i.e., genes) in rice endosperm with the expression of seasoned vitamin A (i.e., beta-carotene). This technological and clinical issue indeed suggests the complexity of the IP of Golden Rice (GR), which, in turn, poses this sort of mission to the developing world. It turned into a lightning rod to discuss GMOs' use to fulfill nutritional needs [41]. GR exhibits both the dramatic dietary advantages of modern-day biotechnology and the fundamental barriers to eventual recognition and effect.

The polished rice grain does not include beta-carotene, a precursor to vitamin A that the frame transforms to vitamin A. In low-income communities wherein rice is the primary staple, many micronutrient deficiencies are persisting, such as lack of nutrition. These deficiencies are mainly widespread in youngsters who need a higher dietary density to attain higher nutrient desires. Simultaneously, as the correlation between VAD and blindness captures media attention, VAD also reduces the immune reaction and increases the demise charge of commonplace formative years illnesses in developing countries [43]. As such, VAD is frequently taken into consideration, especially in terms of early life mortality consequences [41]. The concept of rice as a medium to restore micronutrient deficiencies goes back at the least to the early 1980s. This idea originated within the context of the Consultative Group on International Agricultural Research (CGIAR) system. It contributed to traditional breeding efforts to boost iron and zinc in rice in the 1990s. The development of beta-carotene rice changed until the arrival of contemporary biotechnology strategies [96]. The preliminary research of golden rice has been supported by the Rockefeller basis (RF) via its Rice Biotechnology network explicitly designed for fundamental biotechnological research in this critical food crop, which is all likely to be disregarded with the aid of the non-public quarter in industrialized countries. In the 1990s, with the RF's assist, Ingo Potrykus at the Swiss Federal Institute of technology and Peter Beyer on Freiburg, Germany, worked collectively to combine daffodil genes to the rice. At the time, studies were complicated as it became an early example of using pathway engineering. Their progress has been celebrated as a significant advance within the utility of present-day biotechnology, and the work has appeared in *Science* [96].

The management of intellectual property (IP) rights wishes to be tackled for farm biotechnology (agri-biotech) to play a significant role in developing sustainable agricultural systems. These troubles are not confined to developing countries. As globalization increases, agricultural-biotech IP rights control impacts both developing and developed countries [41]. For instance, in developed countries, IP rights threat control includes the defense of innovations with the aid of robust patent portfolios. In developing countries, IP rights management requires the acquisition of privileges needed to use technologies essential to the population's fundamental welfare [40]. Techniques are necessary to move these different IP control paradigms to facilitate the powerful transition of agrobiotechnology from an industrial supply to a growing country recipient.

Golden rice is an apt model for agri-biotech goods that may be transferred to the developing world. Genetic engineering of cost-brought nutritional content material in Golden rice is a whirling theme in technology and global technology adjustment. Scientifically, the engineering of plant metabolism (in this example, rice) to enhance the buildup of carotene is a pioneering leap forward. The transition of this promising technology becomes need for developing countries but its exploitation from business point of view is itself problematic. In the long term, improved international harmonization of IP laws and control should help increase lots of these risks, and consequently inspire the continued switch of Golden rice and capability agri-biotech advances.

Due to the diverse viewpoints and conditions of the advanced and growing countries, it is not easy to lay the foundation for the same and unbiased IP rights negotiations. Many growing nations will not forget the criminal reputation of international patents applicable. Growing nations usually are at a downside in terms of licensing phrases because the key workforce is poorly known. Maximum developing countries have an insufficient wide variety of licensing officers and IP managers. Therefore, as agriculture appears to be a vital part of national sovereignty in individual growing nations, agro-biotech piracy is more likely to be considered reasonable [43]. The unlucky legacy of colonialism makes the scenario even greater complicated. From a commercial standpoint, traders consider the IP portfolio of a commercial enterprise as a sincerely critical aspect. The IP rights management scheme influences generation transfer and funding decisions. Industrialized nations also strongly agree that the reliable patent machine promotes invention in a supportable way [97]. Notwithstanding the differences of opinion between developing and evolved countries, the improved harmonization of IP rights appears to be a critical part of each countries' economic boom.

26.7 Conservation Strategies

While progress in genomics and breeding techniques would allow the effective transfer of complex traits found in wild rice relatives, the lengthy-term availability of these wild genetic assets is not always assured [98]. Dangers to populations of wild rice relatives and plant biodiversity as a whole encompass threats to human activities and weather alternate [99]. The expansion of city and rural regions contributes to habitat loss and fragmentation. In India, as an instance (Andhra Pradesh), the population of *O. officinalis* ssp. *Malampuzhaensis* is endemic to the Nallamalais of the Eastern Ghats and desires pressing choice and protection as its constrained range renders it extra liable to habitat disruption [100]. *O. rufipogon* populations were threatened by means of overgrazing and its implications for the drift of water in Queensland, Australia [101], and using the development of buildings within the imperative Plains of Thailand [97]. Mounting sea tiers may be a situation for the population of *Oryza* in the wetlands of Northern Australia [101]. For this reason, the sustainable conservation of wild spouses and children of rice is an urgent issue that desires to be tackled before the cited, and undescribed populations that could offer good traits of interest are misplaced for all time.

Step one in conservation is the characterization and selection of species, and populations well stated in many world components, where the wild relatives of rice are determined [102, 103]. However, complete identity and collection are nevertheless required in a few regions, including Australia [101] and Venezuela [104]. There is an immediate need to consider the genetic range and composition of populations to incorporate conservation measures with a gold standard outcome. A vital connection between genetic and geographical distances, in addition to an excessive diploma of enzymatic polymorphism, became observed among *O. glumaepatula* accessions in the South USA [105]. This kind of information is required to determine the need to guard unique populations, and accelerated worldwide efforts need to be made to systematically verify genetic diversity and experiment with gene germplasm for beneficial genes [98]. The evaluation of present threats and the conservation popularity of species and ecosystems is an essential step. Alas, the IUCN red list, which applies the maximum commonplace criteria for determining threats to wild taxa, presently covers a low percentage of wild crop family. There are the best three *Oryza* species (*O. neocaledonica*, *O. bureinalis*, *O. rufipogon*), in which *O. neocaledonica* is considered threatening.

Similarly, new standards should be set up to determine the genetically varied species [99]. While such information is available, a complete rice circle of relative conservation plan can be enforced. Classic conservation has already been properly evolved, with ample seed gathered in GenBank, and the two most significant are the International Rice Research Institute of the Philippines (4370 wild and hybrid species at IRRI) and Oryzabase in Japan (1703) (http://www.shigen.nig.jp). Genetic degradation caused by a lack of exposure to environmental variability, pressure and genetics are linked to germplasm regeneration and is a disadvantage of ex-situ conservation [106, 107]. Assessment of ex-situ and in-situ conservation overall performance of *O. rufipogon* populations in Dongxiang, China, has proven that ex-situ does not preserve genetic diversity, lowering allelic polymorphism by 34% and genetic heterozygosity by 16% in 13 years [106]. On the divergent, the in-situ protection of germplasm, along with its habitat, is a complex form of protection. It allows populations to reply to environmental stress, mainly to improve genetic novelties to be beneficial for future studies and breeding practices. To efficiently maintain the environment's biodiversity, it is vital to complement ex-situ with in-situ conservation. Of the nine populations of *O. rufipogon* studied in Dongxiang in 1978, the most effective three remained in 1995 [106]. A brick fence was changed into built to shield two of those populations susceptible to human interest. The long-term costs and the need for cooperation among national governments with overseas agricultural studies programs lead to demanding situations for introducing and managing new protected sites. A brief-term answer will be able to enhance the control of wild relative rice populations in currently included regions.

26.8 Future Prospects

As indicated in the introductory section, improving rice yields, reducing the environmental effect, and improving nutrition are vital targets to address nine billion people by 2050. This persistent hassle has approximately 25 years to explain if we are to deliver the plasma breeders to the sector that has to be tailored for unique developing conditions. The worldwide *Oryza* Map Alignment Project aims to grow a basic and translational research community, which provides immediate genomic and functionally high-quality reference genome sequences related to populace sequencing information and clone tools. It presents instant entry to any part of the collective *Oryza* genome (i.e., getting access to phenotyped advanced interspecific mapping populations and in situ conserved herbal populations). Any such platform will help classify genes, molecular markers, and germplasm from an evolutionary viewpoint, effortlessly converting various potentially extensive trends into cultivated rice.

Advances in sequencing technologies have greatly affected crop genetics, which allows genome-sequencing and multiple crops transcriptome. While reference genomes have been obtained for many significant crops, large-scale resequencing and gene expression studies are essential to identify the key genes responsible for the desired trait. This knowledge in crop breeding would strengthen the development of better crop varieties and lead to a second green revolution. It would reduce the hunger of billions and revolutionize the economies of developing tropical countries. The usage of NGS for gene evaluation in plants like rice can recognize genes that may be transformed into the domesticated gene pool. Plant breeders using NGS would incorporate variability in their varieties more, unlike before while preserving plant efficiency and product quality objectives. Wild germplasm can be anticipated to be domesticated very quickly. New traits or alleles are added, while significant domestication genes are preserved at the same time. These innovations are essential in adapting agriculture and crops to climate change. The use of NGS and innovative methods for screening the diverse wild resources allow for automation of genetic discoveries and the transition to commercial genotypes.

Conflicts of Interest None

References

1. Gallup JL, Sachs JD. Agriculture, climate, and technology: why are the tropics falling behind? Am J Agric Econ. 2000;82:731–7.
2. Guimarães EP. Rice breeding. In: Carena MJ, editor. Cereals [Internet]. New York: Springer; 2009 [cited 2021 Mar 21]. p. 99–126. https://doi.org/10.1007/978-0-387-72297-9_2.
3. Mannion AM, Morse S. GM crops 1996–2012: a review of agronomic, environmental and socio-economic impacts. Guildford/Reading: Centre for Environmental Strategy, University of Surrey/Department of Geography and Environmental Science, University of Reading; 2013.
4. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-seek database of SNPs derived from 3000 rice genomes. Nucleic acids research. Oxford University Press. 2015;43:D1023–7.
5. Goodman RM, Hauptli H, Crossway A, Knauf VC. Gene transfer in crop improvement. Science. 1987;236:48–54.
6. Huang X, Kurata N, Wang Z-X, Wang A, Zhao Q, Zhao Y, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature. 2012;490:497–501.
7. Ram T, Deen R, Gautam SK, Ramesh K, Rao YK, Brar DS. Identification of new genes for brown planthopper resistance in rice introgressed from O. glaberrima and O. minuta. Rice Genet Newsl. 2010;25:67–9.
8. Ndjiondjop MN, Manneh B, Cissoko M, Drame NK, Kakai RG, Bocco R, et al. Drought resistance in an interspecific backcross population of rice (Oryza spp.) derived from the cross WAB56-104 (O. sativa)×CG14 (O. glaberrima). Plant Sci. 2010;179:364–73.
9. Kang J-W, Suh J-P, Kim D-M, Oh C-S, Oh J-M, Ahn S-N. QTL Mapping of agronomic traits in an advanced backcross population from a cross between Oryza sativa L. cv. Milyang 23 and O. glaberrima. 한국육종학회지. 2008;40:243–9.

10. Miura K, Yamamoto E, Morinaka Y, Takashi T, Kitano H, Matsuoka M, et al. The hybrid breakdown 1 (t) locus induces interspecific hybrid breakdown between rice Oryza sativa cv. Koshihikari and its wild relative O. nivara. Breed Sci. 2008;58:99–105.
11. Thanh PT, Phan PDT, Mori N, Ishikawa R, Ishii T. Development of backcross recombinant inbred lines between *Oryza sativa* Nipponbare and *O. rufipogon* and QTL detection on drought tolerance. Breed Sci. 2011;61:76–9.
12. Maas LF, McClung A, McCouch S. Dissection of a QTL reveals an adaptive, interacting gene complex associated with transgressive variation for flowering time in rice. Theor Appl Genet. 2010;120:895–908.
13. Fu Q, Zhang P, Tan L, Zhu Z, Ma D, Fu Y, et al. Analysis of QTLs for yield-related traits in Yuanjiang common wild rice (Oryza rufipogon Griff.). J Genet Genomics. 2010;37:147–57.
14. Cheema KK, Bains NS, Mangat GS, Das A, Brar DS, Khush GS, et al. Introgression of quantitative trait loci for improved productivity from Oryza rufipogon into O sativa. Euphytica. 2008;160:401–9.
15. Tan L, Liu F, Xue W, Wang G, Ye S, Zhu Z, et al. Development of Oryza rufipogon and O. sativa introgression lines and -assessment for yield-related quantitative trait loci. J Integr Plant Biol. 2007;49:871–84.
16. Rangel PN, Brondani RPV, Rangel PHN, Brondani C. Agronomic and molecular characterization of introgression lines from the interspecific cross Oryza sativa (BG90-2)$\times$ Oryza glumaepatula (RS-16). Genet Mol Res. 2008;7:184–95.
17. Chen Z, Hu F, Xu P, Li J, Deng X, Zhou J, et al. QTL analysis for hybrid sterility and plant height in interspecific populations derived from a wild rice relative, Oryza longistaminata. Breed Sci. 2009;59:441–5.
18. Shan J-X, Zhu M-Z, Shi M, Gao J-P, Lin H-X. Fine mapping and candidate gene analysis of spd6, responsible for small panicle and dwarfness in wild rice (Oryza rufipogon Griff.). Theor Appl Genet. 2009;119:827–36.
19. Angeles-Shim RB, Asano K, Takashi T, Kitano H, Ashikari M. Mapping of the glabrous gene in rice using CSSLs derived from the cross Oryza sativa subsp. Japonica cv. Koshihikari × O. glaberrima. In: Proceedings of 6th intertnational rice genetic symposium, Manila; 2009. p. 16–9.
20. Luo J-J, Hao W, Jin J, Gao J-P, Lin H-X. Fine mapping of Spr3, a locus for spreading panicle from African cultivated rice (Oryza glaberrima Steud.). Mol Plant. 2008;1:830–8.
21. Bocco R, Lorieux M, Seck PA, Futakuchi K, Manneh B, Baimey H, et al. Agro-morphological characterization of a population of introgression lines derived from crosses between IR 64 (Oryza sativa indica) and TOG 5681 (Oryza glaberrima) for drought tolerance. Plant Sci. 2012;183:65–76.
22. Win KT, Yamagata Y, Miyazaki Y, Doi K, Yasui H, Yoshimura A. Independent evolution of a new allele of F1 pollen sterility gene S27 encoding mitochondrial ribosomal protein L27 in Oryza nivara. Theor Appl Genet. 2011;122:385–94.
23. Song B-H, Mitchell-Olds T. Evolutionary and ecological genomics of non-model plants. J Syst Evol. 2011;49:17–24.
24. Rafalski JA. Novel genetic mapping tools in plants: SNPs and LD-based approaches. Plant Sci. 2002;162:329–33.
25. Yamamoto T, Nagasaki H, Yonemaru J, Ebana K, Nakajima M, Shibaya T, et al. Fine definition of the pedigree haplotypes of closely related rice cultivars by means of genome-wide discovery of single-nucleotide polymorphisms. BMC Genomics. 2010;11:267.
26. Molina J, Sikora M, Garud N, Flowers JM, Rubinstein S, Reynolds A, et al. Molecular evidence for a single evolutionary origin of domesticated rice. Proc Natl Acad Sci U S A. 2011;108:8351–6.
27. Gross BL, Zhao Z. Archaeological and genetic insights into the origins of domesticated rice. Proc Natl Acad Sci U S A. 2014;111:6190–7.

28. Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, et al. Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice. 2013;6:1–10.
29. Goff SA, Ricke D, Lan T-H, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science. 2002;296:92–100.
30. Yu J, Hu S, Wang J, Wong GK-S, Li S, Liu B, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science. 2002;296:79–92.
31. Yu J, Ka-Shu Wong G, Liu S, Wang J, Yang H. A comprehensive crop genome research project: the Superhybrid Rice Genome Project in China. Philos Trans R Soc B Biol Sci. 2007;362:1023–34.
32. Yu J, Wang J, Lin W, Li S, Li H, Zhou J, et al. The genomes of Oryza sativa: a history of duplications. PLoS Biol. 2005;3:e38.
33. Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm Phoenix dactylifera L. Nat Commun. 2013;4:1–9.
34. Wong GK-S, Wang J, Tao L, Tan J, Zhang J, Passey DA, et al. Compositional gradients in Gramineae genes. Genome Res. 2002;12:851–6.
35. Wang J, Zhang J, Li R, Zheng H, Li J, Zhang Y, et al. Evolutionary transients in the rice transcriptome. Genomics Proteomics Bioinformatics. 2010;8:211–28.
36. Song S, Tian D, Zhang Z, Hu S, Yu J. Rice genomics: over the past two decades and into the future. Genomics Proteomics Bioinformatics. 2018;16:397–404.
37. Schatz MC, Maron LG, Stein JC, Wences AH, Gurtowski J, Biggers E, et al. Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. Genome Biol. 2014;15:506.
38. Sang T, Ge S. The puzzle of Rice domestication. J Integr Plant Biol. 2007;49:760–8.
39. Zhao K, Wright M, Kimball J, Eizenga G, McClung A, Kovach M, et al. Genomic diversity and introgression in O. sativa reveal the impact of domestication and breeding on the rice genome. PLoS One. 2010;5:e10780.
40. Yang C, Sakai H, Numa H, Itoh T. Gene tree discordance of wild and cultivated Asian rice deciphered by genome-wide sequence comparison. Gene. 2011;477:53–60.
41. He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, et al. Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet. 2011;7:e1002100.
42. Krishnan SG, DLE W, Henry RJ. Australian wild rice reveals pre-domestication origin of polymorphism deserts in rice genome. PLoS One. 2014;9:e98843.
43. Nock CJ, Waters DLE, Edwards MA, Bowen SG, Rice N, Cordeiro GM, et al. Chloroplast genome sequences from total DNA for plant identification. Plant Biotechnol J. 2011;9:328–33.
44. Martienssen RA, Rabinowicz PD, O'Shaughnessy A, McCombie WR. Sequencing the maize genome. Curr Opin Plant Biol. 2004;7:102–7.
45. Guo M, Yang S, Rupe M, Hu B, Bickel DR, Arthur L, et al. Genome-wide allele-specific expression analysis using massively parallel signature sequencing (MPSS™) reveals cis- and trans-effects on gene expression in maize hybrid meristem tissue. Plant Mol Biol. 2008;66:551–63.
46. Jeong D-H, Park S, Zhai J, Gurazada SGR, De Paoli E, Meyers BC, et al. Massive analysis of Rice small RNAs: mechanistic implications of regulated MicroRNAs and variants for differential target RNA cleavage[W][OA]. Plant Cell. 2011;23:4185–207.
47. Awan TH, Ahmadizadeh M, Jabran K, Hashim S, Chauhan BS. Domestication and development of rice cultivars. In: Chauhan BS, Jabran K, Mahajan G, editors. Rice production worldwide [Internet]. Cham: Springer; 2017 [cited 2021 Mar 16]. p. 207–16. https://doi.org/10.1007/978-3-319-47516-5_9.
48. Kim H, Lee KK, Jeon J, Harris WA, Lee Y-H. Domestication of Oryza species eco-evolutionarily shapes bacterial and fungal communities in rice seed. Microbiome. 2020;8:1–17.
49. Henry RJ. Next-generation sequencing for understanding and accelerating crop domestication. Brief Funct Genomics. 2012;11:51–6.

50. Sweeney M, McCouch S. The complex history of the domestication of rice. Ann Bot. 2007;100:951–7.
51. Callaway E. Domestication: the birth of rice. Nature. 2014;514:S58–9.
52. Zhu B-F, Si L, Wang Z, Zhu YZJ, Shangguan Y, Lu D, et al. Genetic control of a transition from black to straw-white seed hull in rice domestication. Plant Physiol. 2011;155:1301–11.
53. Zhang F, Xu T, Mao L, Yan S, Chen X, Wu Z, et al. Genome-wide analysis of Dongxiang wild rice (Oryza rufipogon Griff.) to investigate lost/acquired genes during rice domestication. BMC Plant Biol. 2016;16:103.
54. Waters DLE, Nock CJ, Ishikawa R, Rice N, Henry RJ. Chloroplast genome sequence confirms distinctness of Australian and Asian wild rice. Ecol Evol. 2012;2:211–7.
55. Tanaka K, Zhao C, Wang N, Kubota S, Kanehara M, Kamijo N, et al. Classification of archaic rice grains excavated at the Mojiaoshan site within the Liangzhu site complex reveals an Indica and Japonica chloroplast complex. Food Prod Process Nutr. 2020;2:15.
56. Hebert PDN, Cywinska A, Ball SL, deWaard JR. Biological identifications through DNA barcodes. Proc R Soc Lond Ser B Biol Sci. 2003;270:313–21.
57. Hollingsworth ML, Clark AA, Forrest LL, Richardson J, Pennington RT, Long DG, et al. Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour. 2009;9:439–57.
58. Sickel W, Ankenbrand MJ, Grimmer G, Holzschuh A, Härtel S, Lanzen J, et al. Increased efficiency in identifying mixed pollen samples by meta-barcoding with a dual-indexing approach. BMC Ecol. 2015;15:20.
59. Hollingsworth PM, Graham SW, Little DP. Choosing and using a plant DNA barcode. PLoS One. 2011;6:e19254.
60. Wilkinson MJ, Szabo C, Ford CS, Yarom Y, Croxford AE, Camp A, et al. Replacing sanger with next generation sequencing to improve coverage and quality of reference DNA barcodes for plants. Sci Rep. 2017;7:1–11.
61. Shokralla S, Gibson JF, Nikbakht H, Janzen DH, Hallwachs W, Hajibabaei M. Next-generation DNA barcoding: using next-generation sequencing to enhance and accelerate DNA barcode capture from single specimens. Mol Ecol Resour. 2014;14:892–901.
62. Li X, Yang Y, Henry RJ, Rossetto M, Wang Y, Chen S. Plant DNA barcoding: from gene to genome. Biol Rev. 2015;90:157–66.
63. Zhang W, Sun Y, Liu J, Xu C, Zou X, Chen X, et al. DNA barcoding of Oryza: conventional, specific, and super barcodes. Plant Mol Biol. 2021;105:215–28.
64. Song M, Dong G-Q, Zhang Y-Q, Liu X, Sun W. Identification of processed Chinese medicinal materials using DNA mini-barcoding. Chin J Nat Med. 2017;15:481–6.
65. Lathe W, Williams J, Mangan M, Karolchik D. Genomic data resources: challenges and promises. Nat Educ. 2008;1:2.
66. Mahood EH, Kruse LH, Moghe GD. Machine learning: a powerful tool for gene function prediction in plants. Appl Plant Sci. 2020;8:e11376.
67. Ben-Hur A, Ong CS, Sonnenburg S, Schölkopf B, Rätsch G. Support vector machines and kernels for computational biology. PLoS Comput Biol. 2008;4:e1000173.
68. Gan X, Stegle O, Behr J, Steffen JG, Drewe P, Hildebrand KL, et al. Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature. 2011;477:419–23.
69. Washburn JD, Mejia-Guerra MK, Ramstein G, Kremling KA, Valluru R, Buckler ES, et al. Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. Proc Natl Acad Sci U S A. 2019;116:5542–9.
70. Rodgers-Melnick E, Culp M, DiFazio SP. Predicting whole genome protein interaction networks from primary sequence data in model and non-model organisms using ENTS. BMC Genomics. 2013;14:608.
71. Ding Z, Kihara D. Computational identification of protein-protein interactions in model plant proteomes. Sci Rep. 2019;9:1–13.
72. Zhu G, Wu A, Xu X-J, Xiao P-P, Lu L, Liu J, et al. PPIM: a protein-protein interaction database for maize. Plant Physiol. 2016;170:618–26.

73. Dale JM, Popescu L, Karp PD. Machine learning methods for metabolic pathway prediction. BMC Bioinformatics. 2010;11:1–14.
74. Andersen JR, Lübberstedt T. Functional markers in plants. Trends Plant Sci. 2003;8:554–60.
75. Varshney RK, Mahendar T, Aggarwal RK, Börner A. Genic molecular markers in plants: development and applications. In: Varshney RK, Tuberosa R, editors. Genomics-assisted crop improvement, vol. 1: Genomics approaches and platforms [Internet]. Dordrecht: Springer; 2007 [cited 2021 Mar 13]. p. 13–29. https://doi.org/10.1007/978-1-4020-6295-7_2.
76. Varshney RK, Nayak SN, May GD, Jackson SA. Next-generation sequencing technologies and their implications for crop genetics and breeding. Trends Biotechnol. 2009;27:522–30.
77. Barrera-Figueroa BE, Gao L, Wu Z, Zhou X, Zhu J, Jin H, et al. High throughput sequencing reveals novel and abiotic stress-regulated microRNAs in the inflorescences of rice. BMC Plant Biol. 2012;12:1–11.
78. Sun J, Yang L, Wang J, Liu H, Zheng H, Xie D, et al. Identification of a cold-tolerant locus in rice (Oryza sativa L.) using bulked segregant analysis with a next-generation sequencing strategy. Rice. 2018;11:24.
79. Shim JS, Oh N, Chung PJ, Kim YS, Choi YD, Kim J-K. Overexpression of OsNAC14 improves drought tolerance in rice. Front Plant Sci [Internet]. 2018 [cited 2021 Mar 13];9. https://www.frontiersin.org/articles/10.3389/fpls.2018.00310/full.
80. Xiong H, Yu J, Miao J, Li J, Zhang H, Wang X, et al. Natural variation in OsLG3 increases drought tolerance in rice by inducing ROS scavenging. Plant Physiol. 2018;178:451–67.
81. Mishra A, Wickneswari R, Bhuiyan MAR, Jena KK, Shamsudin NAA. Broad spectrum blast resistance alleles in newly developed Malaysian rice (Oryza sativa L.) genotypes. Euphytica. 2021;217:1–17.
82. Usman B, Nawaz G, Zhao N, Liao S, Qin B, Liu F, et al. Programmed editing of rice (Oryza sativa L.) OsSPL16 gene using CRISPR/Cas9 improves grain yield by modulating the expression of pyruvate enzymes and cell cycle proteins. Int J Mol Sci. 2021;22:249.
83. Pachamuthu K, Swetha C, Basu D, Das S, Singh I, Sundar VH, et al. Rice-specific Argonaute 17 controls reproductive growth and yield-associated phenotypes. Plant Mol Biol. 2021;105:99–114.
84. Bommisetty R, Chakravartty N, Bodanapu R, Naik JB, Panda SK, Lekkala SP, et al. Discovery of genomic regions and candidate genes for grain weight employing next generation sequencing based QTL-seq approach in rice (Oryza sativa L.). Mol Biol Rep. 2020;47:8615–27.
85. Khan GH, Shikari AB, Vaishnavi R, Najeeb S, Padder BA, Bhat ZA, et al. Marker-assisted introgression of three dominant blast resistance genes into an aromatic rice cultivar Mushk Budji. Sci Rep. 2018;8:1–13.
86. Li S, Zhao B, Yuan D, Duan M, Qian Q, Tang L, et al. Rice zinc finger protein DST enhances grain production through controlling Gn1a/OsCKX2 expression. Proc Natl Acad Sci U S A. 2013;110:3167–72.
87. Wang S, Wu K, Yuan Q, Liu X, Liu Z, Lin X, et al. Control of grain size, shape and quality by OsSPL16 in rice. Nat Genet. 2012;44:950–4.
88. Li C, Wei J, Lin Y, Chen H. Gene silencing using the recessive rice bacterial blight resistance gene xa13 as a new paradigm in plant breeding. Plant Cell Rep. 2012;31:851–62.
89. Chen Y, Fan X, Song W, Zhang Y, Xu G. Over-expression of OsPIN2 leads to increased tiller numbers, angle and shorter plant height through suppression of OsLAZY1. Plant Biotechnol J. 2012;10:139–49.
90. Tabuchi H, Zhang Y, Hattori S, Omae M, Shimizu-Sato S, Oikawa T, et al. LAX PANICLE2 of rice encodes a novel nuclear protein and regulates the formation of axillary meristems. Plant Cell. 2011;23:3276–87.
91. Wei X, Xu J, Guo H, Jiang L, Chen S, Yu C, et al. DTH8 suppresses flowering in rice, influencing plant height and yield potential simultaneously. Plant Physiol. 2010;153:1747–58.
92. Dixit S, Singh UM, Singh AK, Alam S, Venkateshwarlu C, Nachimuthu VV, et al. Marker assisted forward breeding to combine multiple biotic-abiotic stress resistance/tolerance in rice. Rice. 2020;13:1–15.

93. Vasumathy SK, Peringottillam M, Sundaram KT, Kumar SHK, Alagu M. Genome- wide structural and functional variant discovery of rice landraces using genotyping by sequencing. Mol Biol Rep. 2020;47:7391–402.
94. Liu J, Gong J, Wei X, Yang S, Huang X, Li C, et al. Dominance complementation of Hd1 and Ghd8 contributes to extremely late flowering in two rice hybrids. Mol Breed. 2020;40:1–10.
95. Lakew B, Eglinton J, Henry RJ, Baum M, Grando S, Ceccarelli S. The potential contribution of wild barley (Hordeum vulgare ssp. spontaneum) germplasm to drought tolerance of cultivated barley (H. vulgare ssp. vulgare). Field Crops Res. 2011;120:161–8.
96. Ye X, Al-Babili S, Klöti A, Zhang J, Lucca P, Beyer P, et al. Engineering the provitamin A (beta-carotene) biosynthetic pathway into (carotenoid-free) rice endosperm. Science. 2000;287:303–5.
97. Nonomura K-I, Morishima H, Miyabayashi T, Yamaki S, Eiguchi M, Kubo T, et al. The wild *Oryza* collection in National BioResource Project (NBRP) of Japan: history, biodiversity and utility. Breed Sci. 2010;60:502–8.
98. Ford-Lloyd BV, Schmidt M, Armstrong SJ, Barazani O, Engels J, Hadas R, et al. Crop wild relatives—undervalued, underutilized and under threat? Bioscience. 2011;61:559–65.
99. Maxted N, Kell S, Toledo Á, Dulloo E, Heywood V, Hodgkin T, et al. A global approach to crop wild relative conservation: securing the gene pool for food and agriculture. Kew Bull. 2010;65:561–76.
100. Elangovan M, Kiran BP, Tonapi VA, Subba RLV, Sivaraj N. Cultivated grasses and their wild relatives in andhra pradesh and their conservation concerns. Indian J Plant Genet Resour. 2012;25:166–73.
101. Henry RJ, Rice N, Waters DL, Kasem S, Ishikawa R, Hao Y, et al. Australian Oryza: utility and conservation. Rice. 2010;3:235–41.
102. Vaughan DA. The wild relatives of rice: a genetic resources handbook. Los Banos: International Rice Research Institute; 1994.
103. Vaughan DA, Morishima H, Kadowaki K. Diversity in the Oryza genus. Curr Opin Plant Biol. 2003;6:139–46.
104. Berlingeri C, Crespo MB. Inventory of related wild species of priority crops in Venezuela. Genet Resour Crop Evol. 2012;59:655–81.
105. Veasey EA, de Andrade Bressan E, Zucchi MI, Vencovsky R, Cardim DC, da Silva RM. Genetic diversity of American wild rice species. Sci Agricola. 2011;68:440–6.
106. Xie J, Agrama HA, Kong D, Zhuang J, Hu B, Wan Y, et al. Genetic diversity associated with conservation of endangered Dongxiang wild rice (Oryza rufipogon). Genet Resour Crop Evol. 2010;57:597–609.
107. Sun J-C, Cao G-L, Ma J, Chen Y-F, Han L-Z. Comparative genetic structure within single-origin pairs of rice (Oryza sativa L.) landraces from in situ and ex situ conservation programs in Yunnan of China using microsatellite markers. Genet Resour Crop Evol. 2012;59:1611–23.