Relevance of Bioinformatics and Database in Omics Study

2

Rama Shankar, Vikas Dwivedi, and Gulab Chand Arya

Abstract

Bioinformatics is an interdisciplinary branch of biological sciences that assists biologists to interpret and extract the biological information from the omics data. The biological information is further used to create different databases for annotation of an unknown molecules from the novel organism. In the cell, different types of molecules are present with diverse functions. Based on their type and functions, these molecules are divided into various categories. These molecules are majorly categorized into DNA, RNA, proteins, and metabolites. The bioinformatics tools and techniques are specific to study and analyze the variations and mechanism of these molecules. These molecules are divided into different omics for better understanding. In DNA, majorly two types of variations occur, which is categorized as genetic and epigenetic variations and known as genomics and epigenomics variations, respectively. Diversity in RNA is studied under the transcriptome category, where the level of mRNA, their regulatory molecules and modifications during synthesis and post-synthesis are examined. In addition, synthesis, modification, and interaction of proteins and metabolites are studied in proteome and metabolome categories. These studies are being analyzed by different bioinformatics tools and their respective databases are

R. Shankar (✉)
Department of Pediatrics and Human Development, Michigan State University, Grand Rapids, MI, USA
e-mail: ramashankar@hc.msu.edu

V. Dwivedi
Department of Ornamental Plants and Agricultural Biotechnology, Institute of Plant Sciences, Agricultural Research Organization, Rishon LeZion, Israel

G. C. Arya
Department of Vegetable and Field Crops, Institute of Plant Sciences, Agricultural Research Organization, Rishon LeZion, Israel

A. Kumar et al. (eds.), *Omics Technologies for Sustainable Agriculture and Global Food Security Volume 1*, https://doi.org/10.1007/978-981-16-0831-5_2

used to extract their biological information. Here, we have discussed in brief about the relevance of various bioinformatics tools and databases, which are being used for the analysis of biologically important molecules. This would provide a basic overview of the importance and application of these tools and databases in different omics study.

Keywords

Genomics · Epigenomics · Transcriptomics · Proteomics · Metabolomics · Database · Bioinformatics · Crop improvement

2.1 Introduction

Bioinformatics is an interdisciplinary branch of biological sciences that deals with applications of computational biology for the collection, storage, and analysis of biological data. In recent years, several omics projects in plants have been performed, which were contributed by a vast amount of sequencing data. These omics data generated through the traditional or high-throughput next-generation sequencing (NGS) approaches and belong to genome, transcriptome, proteome, or metabolome of the plants (Knasmüller et al. 2008). The term genome refers to the complete nuclear chromosomal DNA sequence of an organism, whereas the total messenger RNA (mRNA) content in a cell at a time is termed as trancriptome. Its level varied with different plant developmental stages and external environmental condition. The latter produce proteome, which is the result of the translation of the mRNA. During the cell metabolism, primary and secondary metabolites are generated and complete set of metabolites present in the cell are called as metabolome (Lister et al. 2009; Saito and Matsuda 2010). Besides, various inevitable modifications, such as expression of genes without changing original genetic material (DNA) of the organism occurs during lifetime and inherited to next-generation, are termed as epigenetics changes.

The data and related information obtained from the plant omics can be useful for generating high-density linkage maps, allele mining, QTL mapping, genome-wide association studies (GWAS), SNP genotyping, single sequence repeats (SSR), and a better understanding of metabolic pathways and its regulations. All these information may be helpful for better plant breeding and improvement programs.

Besides, bioinformatics with the support of highly advanced experimental evidences, various databases have been developed and curated (Shinozaki and Sakakibara 2009). These databases help to discover the novel and unknown information of novel plants and organisms. The National Center for Biotechnology Information (NCBI) is among the world's largest resource databases, storing a vast amount of data in various categories. Also, there are various other databases related to specific plants are available, such as rice genome annotation project (RGAP) database for rice (Kawahara et al. 2013), The Arabidopsis Information Resource (TAIR) for Arabidopsis (https://www.arabidopsis.org/), Phytozome (https://

phytozome.jgi.doe.gov/pz/portal.html), and OmicsDI (open source platform facilitating the access and dissemination of omics datasets) (https://www.omicsdi.org). The Phytozome and OmicsDI databases are one of the comprehensive omics databases that included information about several datasets including genomic, transcriptomic, proteomic, and metabolomic data (Goodstein et al. 2012). There is one important tool known as ODG (Omics database generator), which is a tool used for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding (Guhlin et al. 2017). A list of various omics integration, software tools, and web applications is provided in Table 2.1

The present chapter describes the available tools and techniques used for curation, interpretation, and functional relevance of biological data using web-based resources. Further, this chapter also describes the online available databases, which can be used to extract the functional and structural information of unknown genes and proteins of novel plants. The relevant resources are also included for validating metabolic pathways. A basic overview is provided for the workflow of different omics analysis (Fig. 2.1).

2.2 Relevance of Bioinformatics in Genomics

DNA polymorphism is the variation of nucleotides in the genomic DNA. These modifications can be originated as a result of single nucleotide polymorphism (SNP), insertion and deletion (InDels), or simple sequence repeats (SSRs). SNPs are locations within the genome, where the original nucleotide is substituted with other nucleotide, whereas InDels are insertion and deletion of nucleotide in the genome, and these changes are inheritable from one generation to other. The length of insertion and deletion in the genomic DNA varies from one to many bases. However, three nucleotide insertion or deletion is very common (Chai et al. 2018; Jain et al. 2014). This could be an evolutionary adaptation as three nucleotides code for an amino acid. SSRs are another genetic variation that occurs in genome and known as simple sequence repeats of single nucleotide to ten nucleotides. However, during the analysis repeats of two nucleotides or more with specific repetition are considered as the SSRs (Agarwal et al. 2015; Daware et al. 2016; Parida et al. 2015; Dwivedi et al. 2017).

Identification of DNA polymorphisms is highly essential for gene mapping, QTL analysis, and marker-assisted breeding. Various techniques have been used to identify DNA polymorphisms including gel-based, like random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), restriction fragment length polymorphism (RFLP), microsatellites, SSR, simple sequence length polymorphism (SSLP), and non-gel-based techniques, like SNPs and InDels. SNPs/InDels are the most popular non-gel-based DNA marker systems, which represent the position of nucleotide(s), where DNA sequence differs by a single or more bases. SNPs/InDels have gained importance due to their ubiquity in the genome coupled with various characteristics, such as stability, robustness, efficiency, and cost-effectiveness (Alkan et al. 2011; Kumar et al. 2012b; McCouch

Table 2.1 Summary of multi-omics integration software tools and web applications

Tools	Omics integrated	Domain	Functionality	Type of license
Omics	Transcriptomics, proteomics, and metabolomics	Medical (human)	Correlation network analysis, co-expression analysis, phenotype generation, KEGG/ Human Cyc, pathway enrichment, GO enrichment, Name to ID conversion	Open
COBRA	Transcriptomics, proteomics, metabolomics, and Fluxomics	Unspecified	Genome scale integrated modeling of cell metabolism and macro molecular expression	Open
Gaggle	Variety of omics platform bioinformatics solutions	Unspecified	Inoperability of the following tools: Bioinformatics resource manager, Cytoscape, Data Matrix Viewer, KEGG, Genome Browser, MeV, PIPE, Bio Tapestry, N-Browse	Open
KaPPA-view	Transcriptomics, and metabolomics	Plants	Integrates transcriptomics and metabolomics data to map pathways	Open
MADMAX	Metagenomics, transcriptomics, and metabolomics	Plants, medical and clinical	Integrates omics data—Statistical analysis and pathway mapping	Open
MapMan	Metagenomics, transcriptomics, and metabolomics	Plants	Compare data across these two species, KEGG classification, classification into KOG clusters, mapping expression responses	Open
MetaboAnalyst	Genomics, transcriptomics, proteomics, metabolomics, and clinical	Plants, microbial, microbiome, medical and clinical	Data processing and statistical analysis, pathway analysis, multi-omics integration	Open
mixOmics (R package)	Metagenomics, transcriptomics, proteomics, and metabolomics	Unspecified	Integration of data, Chemometric analysis (similarity/difference)	Open
Omickriging (R package)	Transcriptomics, proteomics, and metabolomics	Unspecified	Integration and visualization of omics data	Open

(continued)

Table 2.1 (continued)

Tools	Omics integrated	Domain	Functionality	Type of license
PaintOmics	Transcriptomics, and metabolomics	100 top species of different biological kingdoms	Integration and visualization of transcriptomics and metabolomics data	Open
Reactome	Genomics, transcriptomics, proteomics, and metabolomics	Unspecified	Multi-omics data visualization, metabolic map of known biological processes and pathways	Open
SIMCA	Metagenomics, transcriptomics, proteomics, and metabolomics	Unspecified	Integration of data, Chemometric analysis (similarity/difference)	Commercial
VitisNet	Metagenomics, transcriptomics, proteomics, and metabolomics	Grapes	Integration of data - visualization of connectivity	Open
GenBank (database)	Proteomics	Numerous (over 100,000 organisms)	Proteomics database, open access, annotated collection of all publically available nucleotide sequences and their protein transitions.	Open
Plant metabolic network (PMN)	Genomics, proteomics, and metabolomics	Plants	Plant-specific database containing pathways, enzymes, reactions, and compounds	Open
PRIDE	Proteomics	Unspecified	Proteomics database	
KEGG	Genomes, transcriptomics, proteomics, and metabolomics	Plants animals microbes	Collection of databases dealing with genomes, biological pathways, diseases, drugs, and chemical substances	Open and licensed
Yeast metabolome data (YMDB)	Metabolomics	Microbe (yeast)	Metabolite database	Open
VANTED	Metagenomics, transcriptomics, proteomics, and metabolomics	Unspecified	Comparison of multiple omics data sets, visualization of metabolic maps, correlation networks analysis	Open

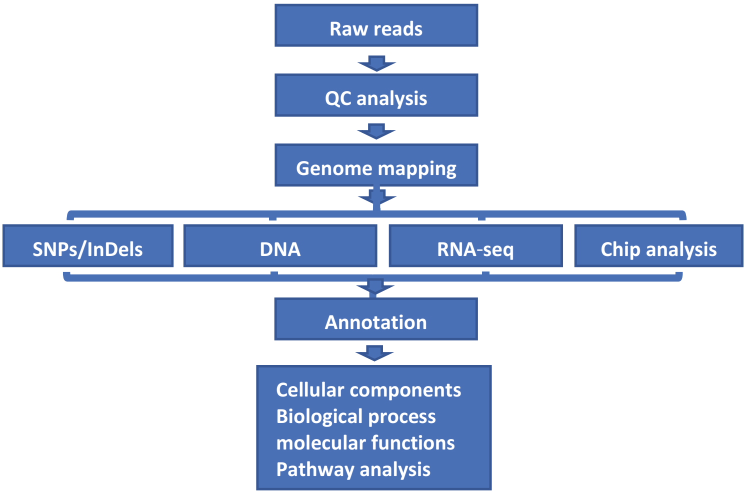

Fig. 2.1 Basic workflow of omics analysis

et al. 2010; Rafalski 2002; Steemers and Gunderson 2007). The next-generation sequencing (NGS) is an easy and cost effective method for discovery of SNPs/ InDels in a population. A large number of SNPs have been discovered from several plant species like Arabidopsis (Atwell et al. 2010), rice (Huang et al. 2010, 2011; Jain et al. 2014; McNally et al. 2009; Meyer et al. 2016; Zhao et al. 2011), maize (Kump et al. 2011; Tian et al. 2011), chickpea (Deokar et al. 2014; Thudi et al. 2014), and soybean (Hwang et al. 2014; Lam et al. 2010) via genome re-sequencing.

Since huge data of SNPs/InDels are being generated using the NGS, a large number of bioinformatics tools are available to validate the biological significance of the aforesaid changes in the genome. For the analysis and validation of SNPs/InDels various bioinformatics tools are available (Li and Wei 2015; Seal et al. 2014), in which GATK and Freebays are the two important tools to discover the SNPs/InDels from the genome mapped files (Garrison and Marth 2012; Van der Auwera et al. 2013). The genome mapping of sequence reads is performed using different tools, mainly TopHat, STAR, and Bowtie tools (Dobin et al. 2013; Trapnell et al. 2009; Wu et al. 2018). Once the DNA polymorphism is identified, it is annotated using the snpEff software (Cingolani et al. 2012). This helps to understand the effect of SNPs/ InDels on various transcriptional, post-transcriptional, and post-translational modifications. These genetic modifications can be further associated with various traits using the genome-wide association (GWAS) study in plants (Marees et al. 2018). The SNPs/InDels associated with various traits can be used for the genetic engineering and crop breeding purposes to improve the crop productivity.

2.3 Application of Bioinformatics in Epigenomics

DNA methylation is one of the epigenetic variations that occur by addition of a methyl group to the genomic DNA. It plays a crucial role in the regulation of chromatin structure and regulates the gene expression in eukaryotes. DNA methylation mainly occurs at the cytosine and adenine nucleotides in DNA; however, methylation in cytosine is specific to higher eukaryotes. In plants, DNA methylation is occurred in three different sequence contexts, CG, CHG, and CHH (where H = A, C or T). This methylation is established and maintained by de novo methyltransferases (DRM1/2/CMT3) via RNA-directed DNA methylation (RdDM) pathway and MET1 proteins (Cao and Jacobsen 2002; Lindroth 2001). Epigenetic modifications are highly stable and heritable, and it regulates cellular and developmental modifications including agronomically important traits in the plants (Manning et al. 2006; Miura et al. 2009; Soppe et al. 2000). DNA methylation analysis has been carried out in different plants to study their role in different developmental processes and stress responses (Chinnusamy and Zhu 2009; Dowen et al. 2012; Gehring et al. 2009; Hsieh et al. 2009; Lang-Mladek et al. 2010; Mirouze et al. 2009; Saze et al. 2003; Zemach et al. 2010).

To study the genome-wide DNA methylation, various techniques have been developed (HPLC, mass spectrometry, Sssl methyltransferase tritium labeling and methyl sensitive restriction enzyme). Initially, these methods were low throughput because they could capture the DNA methylation only in few genes (Karan et al. 2012; Wang et al. 2011). Later, microarray has been proved as first high-throughput technique to study the DNA methylation (Schumacher et al. 2006). Further, next-generation sequencing (NGS) based technique has also been evolved to capture the DNA methylation at the single-base resolution and has been used to study the DNA methylation in various plants including Arabidopsis and rice (Dowen et al. 2012; Garg et al. 2015; Rajkumar et al. 2020; Wang et al. 2011). This technique provides more in-depth knowledge about the DNA methylation, its distribution, and regulation.

Bioinformatics tools such as Bismark and Methylkit are highly efficient tools to analyze the DNA methylation data. Bisulfite sequencing is widely used technique to study the DNA methylation, in which nonmethylated thymine is changed into a cytosine but methylated thymine nucleotide does not modify (Li and Tollefsbol 2011). The first step of bisulfite sequencing is NGS based sequencing. Further, the sequencing data needs to be mapped on genomic DNA. Specific sequence aligner is required to align the sequence reads on the genome. The most widely used sequence aligner is Bismark (Krueger and Andrews 2011). Further, the mapped reads are mined by another bioinformatic tool widely known as Methylkit (Akalin et al. 2012). It extracts the methylated cytosine from the data throughout the genome. This information is further used to annotate and study the biological relevance of methylation on the various biological processes and metabolic pathways using different databases.

2.4 Bioinformatic Tools to Identify the Transcriptomic Alterations

2.4.1 RNA-Seq Analysis

Transcriptome can be defined as the total mRNA in a cell at a particular time. mRNA is derived from one strand of genomic DNA. Further, it translates into a protein with the help of the ribosomes. Transcriptome of the cell can be studied by the microarray and RNA sequencing (RNA-seq) (Jain 2012; Wang et al. 2011). Microarray has low throughput and various limitations as compared to the RNA-seq.

Microarray is based on the hybridization of the DNA probe designed for every gene (Page et al. 2007). They are very specific for the genes. mRNA in one condition is labeled with the green color and mRNA in other condition is tagged with red color. These labeled mRNAs are hybridized on a chip containing DNA probes for various genes. Once the labeled mRNA hybridized with the probe, it emits a fluorescent color, which is detected by the highly sensitive camera. Further, these patterns of color overlap between two conditions and based on the intensity, the differential expression between two conditions is estimated. To analyze these data, GeneSpring GX is one of the most widely used bioinformatics tool provided by the Agilent (Agapito 2019). It is a combination of different utilities that provides powerful, accessible statistical tools for data analysis and visualization. It is designed basically for the need of biologist and enables understanding of transcriptomics, genomics, proteomics, metabolomics, and NGS data within the biological context. It allows the researchers to quick and reliable identification of the biologically significant genes and pathways.

RNA-seq is one of the most advanced techniques based on next-generation sequencing (NGS) to study the transcriptome (Børsting and Morling 2015; Jain 2012; Lister et al. 2009). It has various advantages over microarray, as it can be used to study alternative splicing, polyadenylation, and novel genes or transcript discovery (Rao et al. 2018). During the RNA-seq library preparation process, mRNA is converted into cDNA to enhance stability. The cDNA is mechanically fragmented into small fragments (100–500 nucleotides). These fragments are attached with the adopter sequences present on the sequencing chip. The attached fragments further PCR amplified using the primers based on the adopter sequences to enhance the number of fragments for each molecule. These cDNA fragments are further sequenced by the sequencing technology (Kumar et al. 2012a; Zhong et al. 2011). The sequencing platform uses the sequence by synthesis approach. Based on the sequence length, these techniques are divided into two groups, i.e. short reads and long reads (Berbers et al. 2020). Both of these groups have advantages and disadvantages. The short reads sequencing technology can provide more read depth, whereas long reads technology provides the longer reads but shallow read depth (Reinert et al. 2015).

Once the sequencing is complete, the sequencing reads are mapped on the genome sequence of the respective plant. Mapping of sequencing reads is done by various bioinformatics tools, such as Tophat, SOAP, STAR, Salmon, Bowtie (Dobin

Table 2.2 Databases for the study of promoter sequences and regulatory elements of a gene

Database	Description	URL
TRANSFAC	Transcription factor database	http://transfac.gbf.de/TRANSFAC/
PlantCARE	Plant cis-acting regulatory elements database	http://sphinx.rug.ac.be:8080/PlantCARE/
PLACE	Plant cis-acting regulatory elements database	http://www.dna.affrc.go.jp/htdocs/PLACE/
SignalP 4.0	Identification of signal peptides	http://www.cbs.dtu.dk/services/SignalP/
TargetP	Subcellular localization of sequences	http://www.cbs.dtu.dk/services/TargetP/
LOCTREE3	Subcellular localization of sequences	https://www.rostlab.org/services/loctree3/
Plant-mPLoc	Subcellular localization of sequences	www.csbio.sjtu.edu.cn/bioinf/plant-multi/
PSI-Pred	Prediction of transmembrane regions of the gene	http://bioinf.cs.ucl.ac.uk/psipred/
DNASTAR	Making of sequence assembly	http://www.dnastar.com/
PromPredict	Promoter analysis	http://nucleix.mbu.iisc.ernet.in/prompredict/prompredict.html
CTDB	Transcriptome	http://www.nipgr.ac.in/ctdb.html

et al. 2013; Kim et al. 2013; Patro et al. 2017; Trapnell et al. 2009; Xie et al. 2014). Among all, STAR is the better alignment tool and it provides the normalized count of reads mapped on each gene in every sample (Dobin et al. 2013). Normalized mapped read count is used to estimate the differential gene expression between two samples or conditions. To estimate the differential gene expression, various bioinformatics tools are being used including EdgeR, DESeq, Limma, cufflinks/Cuffdiff, RSEM, and Salmon (Ghosh and Chan 2016; Li and Dewey 2011; Love et al. 2014; Patro et al. 2017; Pollier et al. 2013; Ritchie et al. 2015; Robinson et al. 2010). Edger, DEseq, and Limma are the most used tools for identification of differentially expressed (DE) genes (Love et al. 2014; Ritchie et al. 2015; Robinson et al. 2010).

The DE genes are further used to discover the biological processes and pathways regulated by them. The biological processes were discovered by the EnrichR and BinGO tools (Kuleshov et al. 2016; Maere et al. 2005). For the annotation of DE genes, these tools used the functional annotation from the ontology databases. To discover the role of DE genes in biological pathways KEGG pathway database (https://www.genome.jp/kegg/pathway.html) is used. DE genes were also used to discover the transcription regulatory elements using different databases (Table 2.2). Among all, plant cis-acting regulatory elements database (PlantCARE) and PLACE are the most suitable and highly used database (Guo et al. 2008).

For transcriptomic studies, there are several public databases available to store the transcriptomic data, such as Genevestigator, NASCArrays, ArrayExpress, Stanford Microarray Database, Omics DI, and Gene Expression Omnibus (Bhardwaj and Somvanshi 2015). An example of the database is Chickpea Transcriptome Database (CTDB), which has information about the tools used for transcriptome sequence, transcription factor families, conserved domain(s), and molecular markers in chickpea (Verma et al. 2015) (Table 2.2).

2.4.2 Tools and Databases for Transcription Factor Binding Site

Chromatin immunoprecipitation (ChIP)-sequencing (ChIP-seq) is the method to analyze the protein DNA interaction. It is a combination of chromatin immunoprecipitation (ChIP) coupled with NGS to identify the binding sites of DNA associated proteins. It could be useful to discover the binding sites of any protein and has primarily been used to study the transcription factor (TF) binding sites and chromatin-associated proteins (Mundade et al. 2014).

ChIP-seq includes a few critical steps before the sequencing of the DNA-fragments attached with TF/protein. It starts with the crosslinking of protein with the DNA using formaldehyde (Hoffman et al. 2015; Klockenbusch and Kast 2010; Nadeau and Carlson 2007). However, along with the protein DNA crosslinking there are chances of contamination of RNA-protein complexes in the reaction mixture. This crosslinked sample was fragmented to get the DNA-protein crosslinked fragments and pull-down using antibody. The DNA fragments are then sequenced using the deep short-read sequencing platform. The first step in the ChIP-seq data analysis is known as the peak calling.

The most popular bioinformatics tool for peak calling is MACS (Feng et al. 2012; Zhang et al. 2008). This empirically models the shift size of ChIP-seq tags and uses it to improve the spatial resolution of predicted binding sites. Once the binding sites in the whole genome are predicted, these binding sites must be annotated to find out the respective genes, which are present at the downstream. This can be performed by HOMER and various other databases available to annotate these binding sites and related TFs (Table 2.3) (Heinz et al. 2010, 2018). It provides information about the binding sites and their regulating genes and pathways. This information can be used to identify genes and relevant pathways that can be used to implement in the crop improvement.

2.4.3 Tools and Databases for Analysis of Post-Transcriptional Modifications

Another important event known as alternative splicing is also studied in transcriptome analysis as the post-transcriptional event. Alternative splicing is divided into five categories such as exon skipping, mutually exclusive exon, alternative 5′ donor site, alternative 3′ acceptor site, and intron retention (Bedre et al. 2019; Eckardt 2013; Shang et al. 2017; Shankar et al. 2016). Intron retention is the most common alternative splicing events that happened during the transcription process under normal or any stress condition (Shankar et al. 2016). The recommended tools to identify the alternative splicing are TopHat, MapSplice, SpliceMap, HMMsplicer, STAR, and HISAT (Au et al. 2010; Dimon et al. 2010;

Table 2.3 Database for transcription factor prediction

AGRIS, AtTFDB	Arabidopsis	http://arabidopsis.med.ohio-state.edu/AtTFDB/
DRTF	Rice	http://drtf.cbi.pku.edu.cn/
DPTF	Poplar	http://dptf.cbi.pku.edu.cn/
TOBFAC	Tobacco	http://compsysbio.achs.virginia.edu/tobfac/
PlantTFDB	Plant species	http://planttfdb.cbi.pku.edu.cn/22
PlnTFDB	Plant species	http://plntfdb.bio.uni-potsdam.de/v3.0/20
GRASSIUS, GrassTFDB	Maize, rice, sorghum, and sugarcane	http://grassius.org/grasstfdb.html
LegumeTFDB	Soybean, lotus japonicas, and Medicago truncatula	http://legumetfdb.psc.riken.jp/
DBD	700 species	http://dbd.mrc-lmb.cam.ac.uk/DBD/index.cgi?Home
PlantTFDB	83 species	http://planttfdb.cbi.pku.edu.cn/

Dobin et al. 2013; Kim et al. 2015; Trapnell et al. 2009; Wang et al. 2010). These tools provide information about the alternative splicing in mRNA. Various bioinformatics tools are available for computing the differential expression of transcript isoforms produced as a result of alternative splicing (Kim et al. 2013; Patro et al. 2017). This will help to identify a specific isoform produced during the stress or different developmental stages (Akhter et al. 2018; Jiang et al. 2015; Shankar et al. 2016). A biologist to understand the deeper knowledge of plant development and stress responses will use this information.

RNA secondary structure is another post-transcriptional changes happened in the RNA during the post-transcriptional event (Ding et al. 2014; Wang et al. 2019b; Yang et al. 2018). It is known that genomic DNA is folded into specific shapes in the nucleus. Similar folding is reported in RNA also after post-transcriptional process to deliver its function or stability. It is well established that ribosomal RNA folded into distinct three-dimensional shape including internal loops and helices. It binds with the ribosomal protein and make ribosomal subunit required for protein synthesis. Various studies have been carried out to discover the mRNA secondary structure in plants using the NGS techniques (Ding et al. 2014; Wang et al. 2019b; Yang et al. 2018). It has been observed that mRNA with variations in RNA secondary structure lead to affect various transcriptional and post-transcriptional events (Li et al. 2012). There are several bioinformatics tools available, which can provide the secondary structure of the RNA (Gruber et al. 2008; Reuter and Mathews 2010; Wang et al. 2019a). It has been observed that RNA secondary structure predicted using the bioinformatics tools and structure detected using the NGS technique are very similar (Li et al. 2012).

2.5 Importance of Bioinformatics in Proteomics and Metabolomics

Proteins regulate various biochemical and physiological functions in the cells. The dysregulation of proteins may result in various diseases like cancer, neurodegenerative disease, and metabolic imbalance. Protein is synthesized from the mRNA during the translation process and folded into three-dimensional structure after protein synthesis. If the 3D structure is not folded properly, the protein will not be able to perform its activity and will not be able to interact with other proteins as well. The knowledge of protein–protein interactions and structure can be obtained from various databases (Table 2.4).

One of the most advanced techniques available for proteomic analysis is known as mass spectrometry (Di Falco 2018; Reinders et al. 2004). All the proteins from a sample are needed to be extracted and digested using specific proteases to generate a defined peptide. The peptides obtained are analyzed by the liquid chromatography coupled to mass spectrometry (GC-MS) (Lluveras-Tenorio et al. 2017). During the analysis, peptides eluted from the chromatography are selected and data is recorded as a mass spectrometer. The resulted tandem spectra provide information about the sequence of the peptide. These proteins are further used for functional annotation using the gene ontology (GO) terms and KEGG pathways database. The GO term provides the information about the cellular component, biological process, and molecular functions of the respective genes and proteins. The cellular component GO term provides information about the protein location in the cell compartment. The biological process GO terms provide information about the biological processes and molecular functions GO terms represent activities rather than the entities (molecules or complexes) performed by the genes or proteins (Hill et al. 2008). Similarly, the KEGG pathways database provides knowledge about the metabolic pathways regulated by these proteins. This information is further used by the research scientist to conclude the pathways regulated by these genes and used it to translate into genetic engineering and crop improvement.

There are different public databases available for MS proteomics research. These databases are Global Proteome Machine Database (GPMDB), Mass Spectrometry Interactive Virtual Environment (MassIVE), PRIDE, PeptideAtlas, PeptideAtlas SRM Experiment Library (PASSEL), and Proteomics DB. Moreover, for more integration and sharing of public databases, the Proteome Xchange consortium has been made recently to take its advantage for the scientific community (Perez-Riverol et al. 2015).

Metabolomics is another direction of omics included in the comprehensive assessment and quantification of metabolites present in the cell. Metabolites represent a diverse group of low molecular weight molecules including lipids, amino acids, peptides, nucleic acids, organic acids, vitamins, thiols, and carbohydrates. These metabolites have a different role in the biological systems and their role in various plant stress and development processes needed to be understood (Hussein and El-Anssary 2019; Bartwal et al. 2013; Jwa et al. 2006; Saito and Matsuda 2010; Shankar et al. 2016). Further, this information can be used by the biologist to

Table 2.4 Important computational tools for predicting protein structure and protein–protein interactions

S. No.	Software/ server	URL	Description
1	SWISS-MODEL	http://swissmodel.expasy.org/	Automated protein homology modeling server
2	YASARA	http://www.yasara.org/	Molecular modeling tool
3	ESyPred3D	http://www.unamur.be/sciences/biologie/urbm/bioinfo/easypred/	Homology modeling with increased alignment performance
4	ROSETTA	http://boinc.bakerlab.org/resetta/	3D structure prediction
5	RaptorX	http://raptorx.uchicago.edu/	Protein structure prediction
6	HHPred	http://toolkit.tuebingen.mpg.de/hhpred	Homology detection and structure prediction server
7	Phyre2	http://www.sbg.bio.ic.ac.uk/~phyre2/html/page.cgi?id=index	3D structure prediction
8	Bhageerah	http://www.scfbio-iitd.res.in/bhageerath/index.jsp	Energy-based protein structure prediction server
	3DJigsaw	http://bmm.cancerresearchuk.org/~3djigsaw/	Predict structure and function of protein
9	I-TASSER	http://zhanglab.ccmb.med.umich.edu/I-TASSER/	Predict structure and function of protein
10	3DJigsaw	http://bmm.cancerresearchuk.org/~3djigsaw/	Predict structure and function of protein
11	MODELLER	http://salilab.org/modeller/	Comparative modeling of protein 3D structures
12	PIPE2	http://cgmlab.carleton.ca/PIPE2	PIPE2 queries the protein interactions between two proteins based on specificity and sensitivity
13	HomoMINT	http://mint.bio.uniroma2.it/HomoMINT	HomoMINT predicts interaction in human based on ortholog information in model organisms
14	MirrorTree	http://csbg.cnb.csic.es/mtserver/	The MirrorTree allows graphical and interactive study of the coevolution of two protein families and assesses their interactions in a taxonomic context
15	COG	http://www.ncbi.nlm.nih.gov/COG/	COG shows phylogenetic classification of proteins encoded in genomes
16	PreSPI	http://code.google.com/p/prespi/	PreSPI predicts protein interactions using a combination of domains

(continued)

Table 2.4 (continued)

S. No.	Software/ server	URL	Description
17	InPrePPI	http://inpreppi.biosino.org/InPrePPI/index.jsp	InPrePPI predicts protein interactions in prokaryotes based on genomic context
18	STRING	http://string.embl.de	STRING database includes protein interactions containing both physical and functional associations
19	InterPreTS	http://gabrmn.uab.es/interpret/	InterPreTS uses tertiary structure to predict interactions
20	iWARP	http://groups.csail.mit.edu/cb/iwrap/	iWARP is a threading-based method to predict protein interaction from protein sequences
21	Coev2Net	http://groups.csail.mit.edu/cb/coev2net/	Coev2Net is a general framework to predict, assess, and boost confidence in individual interactions inferred from a high-throughput experiment

perform genetic engineering or plant breeding to improve the crop plants. Various methods have been developed to study the metabolites including GC, HPLC, UPLC, CE coupled to MS and NMR spectroscopy (Boizard et al. 2016; Boros et al. 2018; Garcia-Perez et al. 2020; Lluveras-Tenorio et al. 2017; Patel et al. 2017; Yang et al. 2013, 2020). This could help in separation, detection, characterization, and quantification of such metabolites and their related pathways. However, the diverse group of molecules makes it more challenging to study the metabolites using a single technique. Thus, more than one technique is used to identify the different metabolites in the plant system.

2.6 Challenges and Opportunity in Omics Study

Various advancements have been achieved in the field of omics study. Now we can detect the maximum number of RNA, DNA, and protein content present in the cell. However, different challenges are still persisted, which need to be answered. Even today, during library preparation of DNA or RNA sequencing, we are not able to capture all the DNA and RNA molecules. A large number of RNA and DNA have become degraded during the sample preparation. Genome re-sequencing with advanced technology is not able to cover 100% of the genome of any organism. We used to get a lot of redundancy during the mapping of the sequencing reads on the genome and/or transcriptome. This problem is more prominent in the plants with genome $\geq$2n (diploid). Study of proteomics and metabolomics are at very early stage and recent development in large scale proteomics data impose a substantial challenge for available bioinformatics tools to validate these results (Cho 2007; Hongzhan et al. 2007; Reinders et al. 2004; Schubert et al. 2017). During the proteomic analysis, a large number of challenges needed to be resolved besides sample preparation such as data assembly and database search for the functional annotation (Reinders et al. 2004; Schubert et al. 2017). We can annotate only those proteins, whose information is present in the database, but identifying a novel protein is very challenging.

To capture all the DNA and RNA new methods and techniques are being developed. Single molecule sequencing is evolving, as the new approach is developed to improve the genome coverage. The analysis for these molecules is also being improved. It provides a complete sequence information of all the mRNA expressed in a cell or tissue. This will also enable to get a deeper understanding of the post-transcriptional modifications occurred in RNA. Implementation of this method can solve the limitation of protein sequencing and quantification. During the sample preparation, one part of tissue is used to extract either DNA or RNA or proteins and metabolites. This adds the batch effect in the analysis. Now molecular signature is being analyzed from single cell, so developing methods to extract the entire molecular signature from the same cell or tissue has a great opportunity. Recently, few protocols have been developed to extract the DNA and RNA from same tissue but still need a lot of optimization. In bioinformatics analysis, all the tools and techniques come with few limitations. To solve all these limitations, novel techniques and methods are being developed. Hopefully, in future we will be able to develop more advanced technology to solve all these challenges and limitations.

References

Agapito G (2019) Computer tools to analyze microarray data. Methods Mol Biol 1986:267–282

Agarwal G, Sabbavarapu MM, Singh VK, Thudi M, Sheelamary S, Gaur PM, Varshney RK (2015) Identification of a non-redundant set of 202 in silico SSR markers and applicability of a select set in chickpea (Cicer arietinum L.). Euphytica 205:381–394

Akalin A, Kormaksson M, Li S, Garrett-Bakelman FE, Figueroa ME, Melnick A, Mason CE (2012) methylKit: a comprehensive R package for the analysis of genome-wide DNA methylation profiles. Genome Biol 13:87

Akhter S, Kretzschmar WW, Nordal V, Delhomme N, Street NR, Nilsson O, Emanuelsson O, Sundström JF (2018) Integrative analysis of three RNA sequencing methods identifies mutually exclusive exons of MADS-box isoforms during early bud development in Picea abies. Front Plant Sci 9:1625

Alkan C, Coe BP, Eichler EE (2011) Genome structural variation discovery and genotyping. Nat Rev Genet 12:363–376

Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT et al (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465:627–631

Au KF, Jiang H, Lin L, Xing Y, Wong WH (2010) Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Res 38:4570–4578

Bartwal A, Mall R, Lohani P, Guru SK, Arora S (2013) Role of secondary metabolites and Brassinosteroids in plant defense against environmental stresses. J Plant Growth Regul 32:216–232

Bedre R, Irigoyen S, Petrillo E, Mandadi KK (2019) New era in plant alternative splicing analysis enabled by advances in high-throughput sequencing (HTS) technologies. Front Plant Sci 10:740

Berbers B, Saltykova A, Garcia-Graells C, Philipp P, Arella F, Marchal K, Winand R, Vanneste K, Roosens NHC, De Keersmaecker SCJ (2020) Combining short and long read sequencing to characterize antimicrobial resistance genes on plasmids applied to an unauthorized genetically modified bacillus. Sci Rep 10:4310

Bhardwaj T, Somvanshi P (2015) Plant systems biology: insights and advancements. In: Barh D, Khan MS, Davies E (eds) PlantOmics: the omics of plant science. New Delhi, Springer, pp 791–819

Boizard F, Brunchault V, Moulos P, Breuil B, Klein J, Lounis N, Caubet C, Tellier S, Bascands J-L, Decramer S et al (2016) A capillary electrophoresis coupled to mass spectrometry pipeline for long term comparable assessment of the urinary metabolome. Sci Rep 6:34453

Boros E, Pinkhasov OR, Caravan P (2018) Metabolite profiling with HPLC-ICP-MS as a tool for in vivo characterization of imaging probes. EJNMMI Radiopharm Chem 3:2

Børsting C, Morling N (2015) Next generation sequencing and its applications in forensic genetics. Forensic Sci Int Genet 18:78–89

Cao X, Jacobsen SE (2002) Role of the Arabidopsis DRM methyltransferases in De novo DNA methylation and gene silencing. Curr Biol 12:1138–1144

Chai C, Shankar R, Jain M, Subudhi PK (2018) Genome-wide discovery of DNA polymorphisms by whole genome sequencing differentiates weedy and cultivated rice. Sci Rep 8:14218

Chinnusamy V, Zhu J-K (2009) Epigenetic regulation of stress responses in plants. Curr Opin Plant Biol 12:133–139

Cho WCS (2007) Proteomics technologies and challenges. Genomics Proteomics Bioinformatics 5:77–85

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly (Austin) 6:80–92

Daware A, Das S, Srivastava R, Badoni S, Singh AK, Agarwal P, Parida SK, Tyagi AK (2016) An efficient strategy combining SSR markers- and advanced QTL-seq-driven QTL mapping unravels candidate genes regulating grain weight in rice. Front Plant Sci 7:1535

Deokar AA, Ramsay L, Sharpe AG, Diapari M, Sindhu A, Bett K, Warkentin TD, Tar'an B (2014) Genome wide SNP identification in chickpea for use in development of a high density genetic map and improvement of chickpea reference genome assembly. BMC Genomics 15:708

Di Falco MR (2018) Mass spectrometry-based proteomics. Methods Mol Biol 1775:93–106

Dimon MT, Sorber K, DeRisi JL (2010) HMMSplicer: a tool for efficient and sensitive discovery of known and novel splice junctions in RNA-Seq data. PLoS One 5:e13875

Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM (2014) In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 505:696–700

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21

Dowen RH, Pelizzola M, Schmitz RJ, Lister R, Dowen JM, Nery JR, Dixon JE, Ecker JR (2012) Widespread dynamic DNA methylation in response to biotic stress. Proc Natl Acad Sci U S A 109:E2183–E2191

Dwivedi V, Parida SK, Chattopadhyay D (2017) A repeat length variation in myo-inositol monophosphatase gene contributes to seed size trait in chickpea. Sci Rep 7:4764

Eckardt NA (2013) The plant cell reviews alternative splicing. Plant Cell 25:3639

Feng J, Liu T, Qin B, Zhang Y, Liu XS (2012) Identifying ChIP-seq enrichment using MACS. Nat Protoc 7:1728–1740

Garcia-Perez I, Posma JM, Serrano-Contreras JI, Boulangé CL, Chan Q, Frost G, Stamler J, Elliott P, Lindon JC, Holmes E et al (2020) Identifying unknown metabolites using NMR-based metabolic profiling techniques. Nat Protoc 15:2538–2567

Garg R, Narayana Chevala V, Shankar R, Jain M (2015) Divergent DNA methylation patterns associated with gene expression in rice cultivars with contrasting drought and salinity stress response. Sci Rep 5:14922

Garrison E, Marth G (2012) Haplotype-based variant detection from short-read sequencing. ArXiv:1207.3907 [q-Bio]

Gehring M, Bubb KL, Henikoff S (2009) Extensive demethylation of repetitive elements during seed development underlies gene imprinting. Science 324:1447–1451

Ghosh S, Chan C-KK (2016) Analysis of RNA-seq data using TopHat and cufflinks. Methods Mol Biol 1374:339–361

Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N et al (2012) Phytozome: a comparative platform for green plant genomics. Nucleic Acids Res 40:D1178–D1186

Gruber AR, Lorenz R, Bernhart SH, Neuböck R, Hofacker IL (2008) The Vienna RNA websuite. Nucleic Acids Res 36:W70–W74

Guhlin J, Silverstein KAT, Zhou P, Tiffin P, Young ND (2017) ODG: omics database generator—a tool for generating, querying, and analyzing multi-omics comparative databases to facilitate biological understanding. BMC Bioinformatics 18:367

Guo A-Y, Chen X, Gao G, Zhang H, Zhu Q-H, Liu X-C, Zhong Y-F, Gu X, He K, Luo J (2008) PlantTFDB: a comprehensive plant transcription factor database. Nucleic Acids Res 36:D966–D969

Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol Cell 38:576–589

Heinz S, Texari L, Hayes MGB, Urbanowski M, Chang MW, Givarkes N, Rialdi A, White KM, Albrecht RA, Pache L et al (2018) Transcription elongation can affect genome 3D structure. Cell 174:1522–1536

Hill DP, Smith B, McAndrews-Hill MS, Blake JA (2008) Gene ontology annotations: what they mean and where they come from. BMC Bioinformatics 9:5

Hoffman EA, Frey BL, Smith LM, Auble DT (2015) Formaldehyde crosslinking: a tool for the study of chromatin complexes. J Biol Chem 290:26404–26411

Hongzhan H, Shukla HD, Cathy W, Satya S (2007) Challenges and solutions in proteomics. Curr Genomics 8:21–28

Hsieh T-F, Ibarra CA, Silva P, Zemach A, Eshed-Williams L, Fischer RL, Zilberman D (2009) Genome-wide demethylation of Arabidopsis endosperm. Science 324:1451–1454

Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, Li C, Zhu C, Lu T, Zhang Z et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 42:961–967

Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, Li W, Guo Y, Deng L, Zhu C et al (2011) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 44:32–39

Hussein RA, El-Anssary AA (2019) Plants secondary metabolites: the key drivers of the pharmacological actions of medicinal plants. In: Builders PF (ed) Herbal medicine. IntechOpen, London

Hwang E-Y, Song Q, Jia G, Specht JE, Hyten DL, Costa J, Cregan PB (2014) A genome-wide association study of seed protein and oil content in soybean. BMC Genomics 15:1

Jain M (2012) Next-generation sequencing technologies for gene expression profiling in plants. Brief Funct Genomics 11:63–70

Jain M, Moharana KC, Shankar R, Kumari R, Garg R (2014) Genomewide discovery of DNA polymorphisms in rice cultivars with contrasting drought and salinity stress response and their functional relevance. Plant Biotechnol J 12:253–264

Jiang J, Zhang C, Wang X (2015) A recently evolved isoform of the transcription factor BES1 promotes brassinosteroid signaling and development in Arabidopsis thaliana. Plant Cell 27:361–374

Jwa N-S, Agrawal GK, Tamogami S, Yonekura M, Han O, Iwahashi H, Rakwal R (2006) Role of defense/stress-related marker genes, proteins and secondary metabolites in defining rice self-defense mechanisms. Plant Physiol Biochem 44:261–273

Karan R, DeLeon T, Biradar H, Subudhi PK (2012) Salt stress induced variation in DNA methylation pattern and its influence on gene expression in contrasting rice genotypes. PLoS One 7: e40203

Kawahara Y, de la Bastide M, Hamilton JP, Kanamori H, McCombie WR, Ouyang S, Schwartz DC, Tanaka T, Wu J, Zhou S et al (2013) Improvement of the Oryza sativa Nipponbare reference genome using next generation sequence and optical map data. Rice (N Y) 6:4

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:36

Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360

Klockenbusch C, Kast J (2010) Optimization of formaldehyde cross-linking for protein interaction analysis of non-tagged integrin β 1. J Biomed Biotechnol 2010:1–13

Knasmüller S, Nersesyan A, Misík M, Gerner C, Mikulits W, Ehrlich V, Hoelzl C, Szakmary A, Wagner K-H (2008) Use of conventional and -omics based methods for health claims of dietary antioxidants: a critical overview. Br J Nutr 99:3–52

Krueger F, Andrews SR (2011) Bismark: a flexible aligner and methylation caller for bisulfite-Seq applications. Bioinformatics 27:1571–1572

Kuleshov MV, Jones MR, Rouillard AD, Fernandez NF, Duan Q, Wang Z, Koplev S, Jenkins SL, Jagodnik KM, Lachmann A et al (2016) Enrichr: a comprehensive gene set enrichment analysis web server 2016 update. Nucleic Acids Res 44:90–97

Kumar R, Ichihashi Y, Kimura S, Chitwood DH, Headland LR, Peng J, Maloof JN, Sinha NR (2012a) A high-throughput method for Illumina RNA-Seq library preparation. Front Plant Sci 3:202

Kumar S, Banks TW, Cloutier S (2012b) SNP discovery through next-generation sequencing and its applications. Int J Plant Genomics 2012:831460

Kump KL, Bradbury PJ, Wisser RJ, Buckler ES, Belcher AR, Oropeza-Rosas MA, Zwonitzer JC, Kresovich S, McMullen MD, Ware D et al (2011) Genome-wide association study of quantitative resistance to southern leaf blight in the maize nested association mapping population. Nat Genet 43:163–168

Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B et al (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42:1053–1059

Lang-Mladek C, Popova O, Kiok K, Berlinger M, Rakic B, Aufsatz W, Jonak C, Hauser M-T, Luschnig C (2010) Transgenerational inheritance and resetting of stress-induced loss of epigenetic gene silencing in Arabidopsis. Mol Plant 3:594–602

Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323

Li Y, Tollefsbol TO (2011) DNA methylation detection: bisulfite genomic sequencing analysis. Methods Mol Biol 791:11–21

Li L, Wei D (2015) Bioinformatics tools for discovery and functional analysis of single nucleotide polymorphisms. Adv Exp Med Biol 827:287–310

Li F, Zheng Q, Vandivier LE, Willmann MR, Chen Y, Gregory BD (2012) Regulatory impact of RNA secondary structure across the Arabidopsis transcriptome. Plant Cell 24:4346–4359

Lindroth AM (2001) Requirement of CHROMOMETHYLASE3 for maintenance of CpXpG methylation. Science 292:2077–2080

Lister R, Gregory BD, Ecker JR (2009) Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. Curr Opin Plant Biol 12:107–118

Lluveras-Tenorio A, Vinciguerra R, Galano E, Blaensdorf C, Emmerling E, Perla Colombini M, Birolo L, Bonaduce I (2017) GC/MS and proteomics to unravel the painting history of the lost Giant Buddhas of Bāmiyān (Afghanistan). PLoS One 12:e0172990

Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550

Maere S, Heymans K, Kuiper M (2005) BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics 21:3448–3449

Manning K, Tör M, Poole M, Hong Y, Thompson AJ, King GJ, Giovannoni JJ, Seymour GB (2006) A naturally occurring epigenetic mutation in a gene encoding an SBP-box transcription factor inhibits tomato fruit ripening. Nat Genet 38:948–952

Marees AT, de Kluiver H, Stringer S, Vorspan F, Curis E, Marie-Claire C, Derks EM (2018) A tutorial on conducting genome-wide association studies: quality control and statistical analysis. Int J Methods Psychiatr Res 27:e1608

McCouch SR, Zhao K, Wright M, Tung C-W, Ebana K, Thomson M, Reynolds A, Wang D, DeClerck G, Ali ML et al (2010) Development of genome-wide SNP assays for rice. Breed Sci 60:524–535

McNally KL, Childs KL, Bohnert R, Davidson RM, Zhao K, Ulat VJ, Zeller G, Clark RM, Hoen DR, Bureau TE et al (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. Proc Natl Acad Sci U S A 106:12273–12278

Meyer RS, Choi JY, Sanches M, Plessis A, Flowers JM, Amas J, Dorph K, Barretto A, Gross B, Fuller DQ et al (2016) Domestication history and geographical adaptation inferred from a SNP map of African rice. Nat Genet 48:1083–1088

Mirouze M, Reinders J, Bucher E, Nishimura T, Schneeberger K, Ossowski S, Cao J, Weigel D, Paszkowski J, Mathieu O (2009) Selective epigenetic control of retrotransposition in Arabidopsis. Nature 461:427–430

Miura K, Agetsuma M, Kitano H, Yoshimura A, Matsuoka M, Jacobsen SE, Ashikari M (2009) A metastable DWARF1 epigenetic mutant affecting plant stature in rice. Proc Natl Acad Sci U S A 106:11218–11223

Mundade R, Ozer HG, Wei H, Prabhu L, Lu T (2014) Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell Cycle 13:2847–2852

Nadeau OW, Carlson GM (2007) Protein interactions captured by chemical cross-linking: one-step cross-linking with formaldehyde. CSH Protoc 2007:4634

Page GP, Zakharkin SO, Kim K, Mehta T, Chen L, Zhang K (2007) Microarray analysis. Methods Mol Biol 404:409–430

Parida SK, Verma M, Yadav SK, Ambawat S, Das S, Garg R, Jain M (2015) Development of genome-wide informative simple sequence repeat markers for large-scale genotyping applications in chickpea and development of web resource. Front Plant Sci 6:645

Patel DP, Krausz KW, Xie C, Beyoğlu D, Gonzalez FJ, Idle JR (2017) Metabolic profiling by gas chromatography-mass spectrometry of energy metabolism in high-fat diet-fed obese mice. PLoS One 12:e0177953

Patro R, Duggal G, Love MI, Irizarry RA, Kingsford C (2017) Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 14:417–419

Perez-Riverol Y, Alpi E, Wang R, Hermjakob H, Vizcaíno JA (2015) Making proteomics data accessible and reusable: current state of proteomics databases and repositories. Proteomics 15:930–950

Pollier J, Rombauts S, Goossens A (2013) Analysis of RNA-Seq data with TopHat and cufflinks for genome-wide expression analysis of jasmonate-treated plants and plant cultures. Methods Mol Biol 1011:305–315

Rafalski A (2002) Applications of single nucleotide polymorphisms in crop genetics. Curr Opin Plant Biol 5:94–100

Rajkumar MS, Shankar R, Garg R, Jain M (2020) Bisulphite sequencing reveals dynamic DNA methylation under desiccation and salinity stresses in rice cultivars. Genomics 112:3537–3548

Rao MS, Van Vleet TR, Ciurlionis R, Buck WR, Mittelstadt SW, Blomme EAG, Liguori MJ (2018) Comparison of RNA-Seq and microarray gene expression platforms for the toxicogenomic evaluation of liver from short-term rat toxicity studies. Front Genet 9:636

Reinders J, Lewandrowski U, Moebius J, Wagner Y, Sickmann A (2004) Challenges in mass spectrometry-based proteomics. Proteomics 4:3686–3703

Reinert K, Langmead B, Weese D, Evers DJ (2015) Alignment of next-generation sequencing reads. Annu Rev Genomics Hum Genet 16:133–151

Reuter JS, Mathews DH (2010) RNAstructure: software for RNA secondary structure prediction and analysis. BMC Bioinformatics 11:129

Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015) Limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res 43:e47

Robinson MD, McCarthy DJ, Smyth GK (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140

Saito K, Matsuda F (2010) Metabolomics for functional genomics, systems biology, and biotechnology. Annu Rev Plant Biol 61:463–489

Saze H, Mittelsten Scheid O, Paszkowski J (2003) Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. Nat Genet 34:65–69

Schubert OT, Röst HL, Collins BC, Rosenberger G, Aebersold R (2017) Quantitative proteomics: challenges and opportunities in basic and applied research. Nat Protoc 12:1289–1294

Schumacher A, Kapranov P, Kaminsky Z, Flanagan J, Assadzadeh A, Yau P, Virtanen C, Winegarden N, Cheng J, Gingeras T et al (2006) Microarray-based DNA methylation profiling: technology and applications. Nucleic Acids Res 34:528–542

Seal A, Gupta A, Mahalaxmi M, Aykkal R, Singh TR, Arunachalam V (2014) Tools, resources and databases for SNPs and indels in sequences: a review. Int J Bioinform Res Appl 10:264–296

Shang X, Cao Y, Ma L (2017) Alternative splicing in plant genes: a means of regulating the environmental fitness of plants. Int J Mol Sci 18:432

Shankar R, Bhattacharjee A, Jain M (2016) Transcriptome analysis in different rice cultivars provides novel insights into desiccation and salinity stress responses. Sci Rep 6:23719

Shinozaki K, Sakakibara H (2009) Omics and bioinformatics: an essential toolbox for systems analyses of plant functions beyond 2010. Plant Cell Physiol 50:1177–1180

Soppe WJ, Jacobsen SE, Alonso-Blanco C, Jackson JP, Kakutani T, Koornneef M, Peeters AJ (2000) The late flowering phenotype of fwa mutants is caused by gain-of-function epigenetic alleles of a homeodomain gene. Mol Cell 6:791–802

Steemers FJ, Gunderson KL (2007) Whole genome genotyping technologies on the BeadArray platform. Biotechnol J 2:41–49

Thudi M, Upadhyaya HD, Rathore A, Gaur PM, Krishnamurthy L, Roorkiwal M, Nayak SN, Chaturvedi SK, Basu PS, Gangarao NVPR et al (2014) Genetic dissection of drought and heat tolerance in chickpea through genome-wide and candidate gene-based association mapping approaches. PLoS One 9:e96758

Tian F, Bradbury PJ, Brown PJ, Hung H, Sun Q, Flint-Garcia S, Rocheford TR, McMullen MD, Holland JB, Buckler ES (2011) Genome-wide association study of leaf architecture in the maize nested association mapping population. Nat Genet 43:159–162

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111

Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, Del Angel G, Levy-Moonshine A, Jordan T, Shakir K, Roazen D, Thibault J et al (2013) From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. Curr Protoc Bioinformatics 43:1–33

Verma M, Kumar V, Patel RK, Garg R, Jain M (2015) CTDB: an integrated chickpea transcriptome database for functional and applied genomics. PLoS One 10:e0136880

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM et al (2010) MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res 38:e178

Wang W-S, Pan Y-J, Zhao X-Q, Dwivedi D, Zhu L-H, Ali J, Fu B-Y, Li Z-K (2011) Drought-induced site-specific DNA methylation and its association with drought tolerance in rice (Oryza sativa L.). J Exp Bot 62:1951–1960

Wang L, Liu Y, Zhong X, Liu H, Lu C, Li C, Zhang H (2019a) DMfold: a novel method to predict RNA secondary structure with pseudoknots based on deep learning and improved base pair maximization principle. Front Genet 10:143

Wang Z, Wang M, Wang T, Zhang Y, Zhang X (2019b) Genome-wide probing RNA structure with the modified DMS-MaPseq in Arabidopsis. Methods 155:30–40

Wu DC, Yao J, Ho KS, Lambowitz AM, Wilke CO (2018) Limitations of alignment-free tools in total RNA-seq quantification. BMC Genomics 19:510

Xie Y, Wu G, Tang J, Luo R, Patterson J, Liu S, Huang W, He G, Gu S, Li S et al (2014) SOAPdenovo-trans: de novo transcriptome assembly with short RNA-Seq reads. Bioinformatics 30:1660–1666

Yang S-O, Lee SW, Kim YO, Sohn S-H, Kim YC, Hyun DY, Hong YP, Shin YS (2013) HPLC-based metabolic profiling and quality control of leaves of different Panax species. J Ginseng Res 37:248–253

Yang X, Yang M, Deng H, Ding Y (2018) New era of studying RNA secondary structure and its influence on gene regulation in plants. Front Plant Sci 9:671

Yang G, Liang K, Zhou Z, Wang X, Huang G (2020) UPLC-ESI-MS/MS-based widely targeted metabolomics analysis of wood metabolites in teak (Tectona grandis). Molecules 25:2189

Zemach A, Kim MY, Silva P, Rodrigues JA, Dotson B, Brooks MD, Zilberman D (2010) Local DNA hypomethylation activates genes in rice endosperm. Proc Natl Acad Sci U S A 107:18729–18734

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al (2008) Model-based analysis of ChIP-Seq (MACS). Genome Biol 9:R137

Zhao K, Tung C-W, Eizenga GC, Wright MH, Ali ML, Price AH, Norton GJ, Islam MR, Reynolds A, Mezey J et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in Oryza sativa. Nat Commun 2:467

Zhong S, Joung J-G, Zheng Y, Chen Y, Liu B, Shao Y, Xiang JZ, Fei Z, Giovannoni JJ (2011) High-throughput illumina strand-specific RNA sequencing library preparation. Cold Spring Harb Protoc 2011:940–949