



# Sequencing the Rice Genome: Gateway to Agricultural Development

Anindita Paul

## Abstract

For 10,000 years, rice, the most important staple crop in the world, has played a central role in human nutrition and culture. Ensuring a stable supply of this crop to an ever-increasing global population in the face of climate change has become essential. The focus, therefore, has shifted on maximising rice productivity, yield potential and stability. This requires an in-depth understanding of the functional genomics of rice and its breeding pipeline. Spearheaded by the Rice Genome Research Program (Japan), with multinational contribution from ten countries and integration of draft sequences from private organisations, the International Rice Genome Sequencing Project (IRGSP) published a map-based, high-quality genome sequence for *Oryza sativa* ssp. *japonica* variety ‘Nipponbare’ in 2006. With the release of the first crop genome sequence data, the global rice community identified novel genes underlying agronomic traits, developed new tool sets for marker-assisted breeding and positional cloning and advanced towards sequencing other cereal genomes. Enhanced comparative and functional genomic studies delivered crucial insights into genome diversity and evolution, speciation and domestication. Integration of the whole-genome data into diverse omics data like transcriptome, protein-protein interaction network and metabolome allowed high-throughput analysis and orthologous gene identification. The rice genome sequence simultaneously laid the foundation for the international 3000 rice genomes resequencing effort and for identification of candidate loci that can be exploited to breed Green Super Rice. More than a decade later, this milestone continues to serve as an information platform for leveraging the genomics of agroecosystems.

---

A. Paul (✉)

Molecular Biophysics Unit, Indian Institute of Science, Bangalore, Karnataka, India

e-mail: [aninditapaul@iisc.ac.in](mailto:aninditapaul@iisc.ac.in)

© Springer Nature Singapore Pte Ltd. 2020

A. Roychoudhury (ed.), *Rice Research for Quality Improvement: Genomics and Genetic Engineering*, [https://doi.org/10.1007/978-981-15-4120-9\\_6](https://doi.org/10.1007/978-981-15-4120-9_6)

109

---

**Keywords**

Rice genome sequencing · Nipponbare · Systems biology · Map-based cloning · Green Super Rice

---

## 1 Introduction

A comprehensive review of global demographic trends and future prospects published in the *World Population Prospects: The 2017 Revision*, released by the United Nations Department of Economic and Social Affairs (UN DESA), predicted the current world population of 7.6 billion to grow each year by roughly 83 million people (<https://population.un.org/wpp/>). Hence the global population is projected to reach 8.6 billion in 2030, 9.8 billion in 2050 and 11.2 billion in 2100. Within this period from 2017 to 2050, half of the global population growth is expected to be concentrated in a small number of countries, and thus such a status in the poorest countries will present a significant challenge to the respective governments, which strive to implement the 2030 Agenda for Sustainable Development. The nine countries listed in the review that are expected to make a significant contribution to the global population are Indonesia, Uganda, the United States of America, the United Republic of Tanzania, Ethiopia, Pakistan, the Democratic Republic of the Congo, Nigeria and India. The unavoidable question that arises in the face of this global crisis is that how to ensure access of safe, nutritious and sufficient food to all the people?

### 1.1 Rice: The Right Choice

Rice is a very old crop plant, with a long evolutionary and cultivation history. Since time immemorial, it is regarded as an excellent staple crop, as cooking of rice is very simple in contrast to bread making which requires fermentation by yeast. Not only is rice a rich source of carbohydrates but also contains all the essential amino acids except lysine. Rice is mainly harvested and consumed in Asia and Africa, but at a global level, rice is the staple crop for half of the world's population and constitutes nearly 50% of the daily calorie intake. Similar to religion or tradition, rice is deeply ingrained in our lives.

The genus *Oryza* contains 27 species, 2 of which were domesticated independently ~6000 years apart (Wing et al. 2018). The domestication of the Asian rice (*Oryza sativa*) and African rice (*Oryza glaberrima*) occurred on two different continents, at different time points in the evolutionary history of rice. Therefore, rice has an extensive germplasm collection of wild and domesticated species, with some extinct wild species which were progenitors of the present-day rice. Thus the present-day picture of rice evolution is quite complex, as it emerges that divergent ancestral populations gave rise to extant rice populations (Civán et al. 2015). One or

more de novo domestication events may have led to the different varietal groups that exist today, e.g. *Oryza sativa* subspecies, *japonica* and *indica*.

Since its domestication and adaptation, rice has been cultivated for thousands of years in the Old World and for hundreds of years in the New World (Wing et al. 2018). Throughout this duration, rice has undergone significant improvement, adapting to range of geographical locations, soil textures, climate or environmental conditions and also cooking preferences. These have also led to positive selection of agronomically desirable traits, which are still applied to current-day breeding programmes. Identifying these desirable traits and crossing plants to develop new varieties with high-yielding phenotype or with resistance towards adverse climatic conditions are already known. For achieving higher yield and accelerated growth, hybrid vigour or heterosis has also been exploited, where genetically distinct parents are crossed to produce a hybrid offspring with an improved or enhanced function of a desirable biological quality. An early example of rice improvement by hybridisation, hailing from the Neolithic age, is the origin of *O. sativa* varietal group *indica* (Wing et al. 2018). Introgression of genomic regions from subspecies *japonica* to *indica* led to integration of a number of agronomically desirable genes, for example:

- (a) *SH4*: non-shattering allele, enhances sustainability of mature seeds on the panicle.
- (b) *RC*: colour allele, imparts the white grain colour which is desirable.
- (c) *PROGI*: important for erect growth of the crop, imparts lodging resistance.

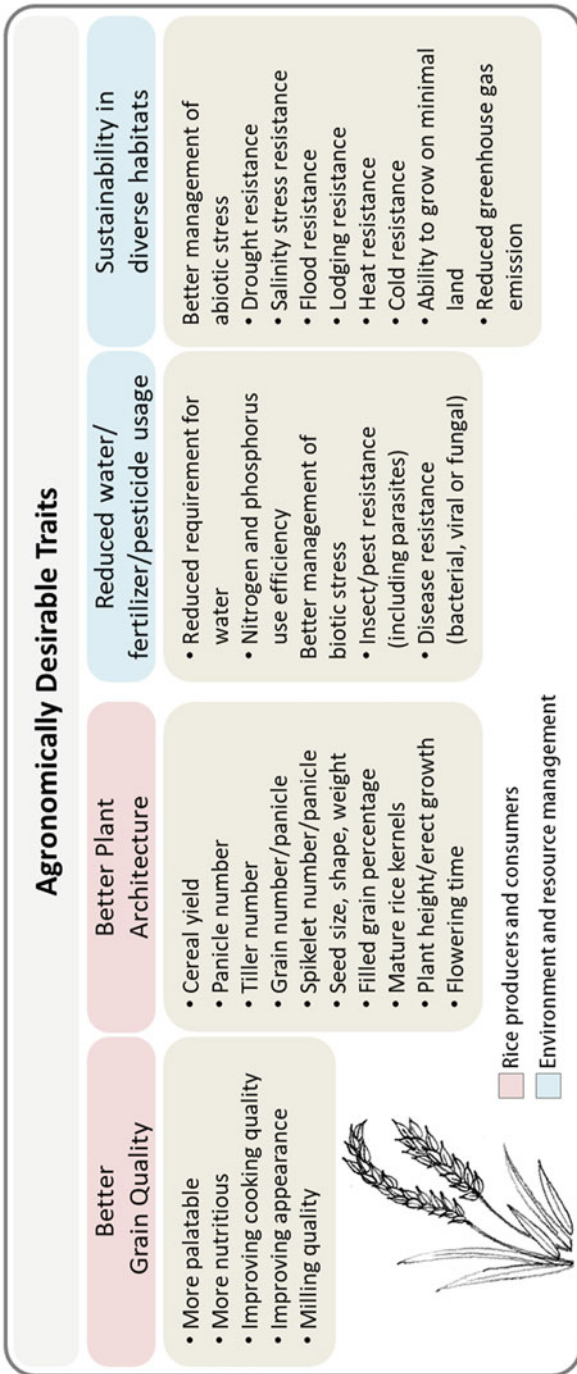
Interestingly the *SH4* orthologue in African rice is associated with mutation, which drives the selection for the non-shattering phenotype in this domesticated crop (Wang et al. 2014). Other genes regulating plant height, grain quality, disease and stress resistance, fertility, nutrient uptake and cooking or eating quality were eventually incorporated as a result of geographical spread and adaptation. Genome-wide association studies have revealed genomic regions that have undergone selection related to geographical adaptation and identified genes of agronomic potential (Meyer et al. 2016) (summarised in Table 1).

With the ever-increasing trend of the global population and the current status of food resources, it is clear that rice will play a crucial role to fulfil the global demands. But increasing quality and quantity only by implementation of traditional crop improvement programmes will undermine the crucial contribution that this cereal can deliver in this lingering crisis. Also there are additional factors associated with breeding of sustainable crops that cannot be ignored (summarised in Fig. 1). With population growth, constant urbanisation and increasing misuse of land, there is an evident dearth of cultivable land. Excessive use of resources such as fertilisers, pesticides and water has had adverse effects on the environment. Increasing rice production by developing high-yielding varieties therefore necessitates incorporating other attributes such as reducing dependency on fertilisers and pesticides, increasing efficiency of water and nutrient use, enhancing tolerance to biotic and abiotic stresses (e.g. disease, drought, salinity, temperature, etc.), being able to grow on limiting lands and reducing emissions of greenhouse gases (Zhang

**Table 1** List of genes that impart agronomically desirable traits and were positively selected during domestication and diversification of cultivated rice

| Genes  | Agronomically desirable traits   |
|--|--|
| <i>SH1/SH4</i><br>( <i>non-shattering allele</i> ) | <i>SH1</i> locus for seed shattering on chromosome 1; <i>SH4</i> locus for grain shattering on chromosome 4; role in retention of mature rice grains on the panicle (Li et al. 2006)                               |
| <i>PROG1</i><br>( <i>prostate growth 1</i> )       | Encodes a single Cys2-His2 zinc-finger protein. Naturally occurring mutants in this gene affect plant architecture (erect growth) and grain yield (Tan et al. 2008)  |
| <i>hd1</i>   | A homolog of CONSTANS from <i>Arabidopsis</i> , encodes a protein with a zinc-finger domain, regulates flowering time and heading time (Zhang et al. 2012)   |
| <i>sd1</i><br>( <i>semidwarf</i> )                 | The semidwarfing varieties contain a defective gibberellin 20-oxidase gene, whereby the deficiency in gibberellin biosynthesis affects plant height (Spielmeier et al. 2002)                                       |
| <i>Gn1a</i>  | Encodes cytokinin oxidase/dehydrogenase, regulates rice grain yield (Wang et al. 2015)   |
| <i>Waxy</i>  | Mutation in this gene leads to change in endosperm starch content and imparts the glutinous or sticky texture in rice (Olsen et al. 2006)  |
| <i>Rc</i>  | A predominant frameshift deletion in this gene is found in the white rice varieties (Sweeney et al. 2007)  |
| <i>HAK5</i>  | Encodes a high-affinity potassium transporter, role in K <sup>+</sup> acquisition by roots and upward transport (Yang et al. 2014)   |
| <i>Cyp2</i><br>( <i>cyclophilin2</i> )             | Upregulated during salinity stress and other types of stresses, serving as a general integrator of environmental stresses (Ruan et al. 2011)   |
| <i>AMT1</i>  | Encodes an ammonium transporter, regulates ammonium uptake and assimilation in rice (Hoque et al. 2006)  |
| <i>Xa26</i>  | Encodes a leucine-rich repeat (LRR) receptor kinase-like protein, which confers resistance against <i>Xanthomonas oryzae</i> pv. <i>oryzae</i> ( <i>Xoo</i> ), at both seedling and adult stages (Sun et al. 2004) |
| <i>Rf-1</i><br>( <i>fertility restorer gene</i> )  | A nuclear-encoded gene which recovers pollen fertility, thereby cytoplasmic male sterility (CMS), inherited by the mitochondrial genome (Kazama and Toriyama 2003)   |

2007; Roychoudhury et al. 2013). Other qualities that will contribute to the better acceptability of rice as a staple diet include improving plant architecture (e.g. number of panicles, number of grains per panicle, grain weight, more productive tillers, spikelet number, uniformly mature kernels, free of empty or half-filled grains); manipulating flowering time, spikelet fertility or heading date; and improving the appearance (uniform size and shape, colour, translucency), milling (high milling recovery, high head rice recovery, no discoloration), cooking (amylose content, gel length and gelling temperature) and nutritional (both micro- and macronutrients) quality of the rice grain (Zhang 2007). To achieve this, scientists and researchers, over the years, had to leverage genomics to obtain a thorough understanding of the genetic components that regulate the abovementioned attributes. Hence this was preceded by deciphering the entire rice genome sequence.



**Fig. 1** Expected deliverables from a typical present-day crop improvement programme that aims for breeding of sustainable rice crop

Rice is the first cereal crop, the first grass species and the first monocot genome to be sequenced. Rice was an attractive target for genome sequence analysis because it has a relatively small genome (430 Mb,  $2n = 24$ ) compared to that of other common cereals like sorghum (750 Mb), maize (3000 Mb), barley (5000 Mb) and wheat (16,000 Mb). Cereal genomes are highly conserved and exhibit considerable synteny among the genes (Freeling 2001). Rice, with a small genome size, was predicted to have higher gene density compared to the other candidates. Moreover, rice has an elaborate germplasm of wild and domesticated species. Also, once genetic markers were identified, they had to be tinkered with to yield the desirable phenotypes, and rice was an easy candidate for genetic manipulation. Therefore, rice emerged to be an excellent model for plant genomics, following *Arabidopsis*.

The following sections aim to highlight how the International Rice Genome Sequencing Project (IRGSP) was conceived as well as its contribution and impact on the present-day understanding of rice genomics. Recently, resequencing efforts through high-throughput platforms and mapping this data into high-quality reference sequence genomes is helping to identify novel targets for genomic breeding and making development of Green Super Rice (GSR) varieties scientifically and technically feasible.

---

## 2 Mapping the Rice Genome

The profound importance of rice as a food crop and as a model plant by virtue of its syntenic relationships with other cereal crops has significantly stimulated rice genome analysis. Genome analysis in rice comprised of the following goals: (1) mapping of rice genome, (2) sequencing the rice genome and (3) functional association of the mapped genes. Large volumes of structural and functional data of the rice genome were generated in the twentieth century which facilitated mapping of the rice genome. This created a strong foundation for subsequent sequencing of the rice genome, which was possible with major advancements in sequencing strategies. Initial efforts included cytogenetic mapping, linkage mapping, genetic mapping and physical mapping, which have been utilised extensively to build sequence-ready physical map of the rice genome (Tyagi et al. 2004).

1. Cytogenetic maps conventionally were based on microscopic examination of chromosomes. With the advent of novel effective techniques like fluorescence in situ hybridisation (FISH) and genomic in situ hybridisation (GISH), an enhanced-quality rice chromosome map could be generated (Heng et al. 1997). GISH was used extensively to dissect the genomic constitution of rice somatic hybrids and tetraploid species, as this technique could efficiently identify a chromosome complement belonging to a particular genome (Fukui et al. 1997). FISH, on the contrary, allowed mapping of restriction fragment length polymorphism (RFLP) markers, rDNA loci, bacterial artificial chromosome (BAC) and yeast artificial chromosome (YAC) clones on the respective rice chromosomes (Jiang et al. 1995).

2. Khush and Brar proposed 12 linkage groups which corresponded to the haploid number of chromosomes in rice (Khush and Brar 2001). These linkage groups were associated with the rice chromosomes by primary trisomics and reciprocal translocations, while centromere positions and orientation of the linkage groups were determined using secondary trisomics and telotrisomics. All these efforts culminated into a comprehensive molecular linkage map of rice.
3. The advent of molecular genetic markers provided significant impetus to the construction of genetic maps, which were being eventually supplemented with phenotypic markers. Several genetic maps of rice were being constructed from RFLP, amplified fragment length polymorphism (AFLP), random amplified polymorphic DNA (RAPD), cleaved amplified polymorphic sequence (CAPS) and microsatellite or simple sequence length polymorphism (SSLP) markers (Mohan et al. 1997). Most of these markers were developed from expressed sequence tag (EST) clones derived from cDNA libraries of rice.
4. The immense progress that was feasible while constructing the physical maps was accomplished mainly due to the availability of libraries of large insert clones in vectors like BAC, YAC and P1-derived artificial chromosomes (PACs) and the information about DNA markers on genetic maps. The YAC clones were originally mapped onto the rice chromosomes by chromosome landing, integrating information from the genetic maps. Such integrated YAC physical map was subsequently used for positional cloning of several genes (e.g. rice blast resistance gene, *Pib*; bacterial blight resistance gene, *Xa-1*; and gibberellin-insensitive dwarf mutant gene, *d1*), for the assignment of chromosomal locations of more than 6000 EST markers to generate a rice transcript map and also has been utilised as a backbone for construction of PAC- and BAC-derived physical maps (Yoshimura et al. 1996; Ashikari et al. 1999; Wang et al. 1999; Wu et al. 2002). Tao et al. (2002) developed one large insert plant-transformation-competent BIBAC library and three BAC libraries for the temperate *japonica* rice cv. Nipponbare, which significantly aided functional analysis of the genome, comparative genomics of grass species and subspecies and molecular breeding in rice and other cereals (Tao et al. 2002).
5. With constructions of these abovementioned maps, the next approach was to integrate all the information obtained from the cytogenetic, genetic and physical maps to constitute a comprehensive physical-genetic map, which will cover a large fraction of the genome and facilitate the study of organisation and functional aspects of the genome. For instance, a standardised rice karyotype was constructed from meiotic pachytene chromosomes of *O. sativa* spp. *japonica* rice cv. Nipponbare, using centromere-specific DNA probes and chromosomal arm-specific BACs, following which it was fully integrated with the most saturated rice genetic linkage maps in which Nipponbare was used as one of the mapping parents (Cheng et al. 2001a). A physical-genetic map of rice chromosome 10 was developed by integrating the pachytene chromosome-based FISH mapping of BAC clones with a genetic linkage map, which revealed the precise genetic position of the centromere on chromosome 10 and reflected on

the genetic recombination frequencies among the chromosomal arms (Cheng et al. 2001b).

## 2.1 Rice Genome Research Program (RGP), Japan, Established 1991

In 1991, the Japanese government initiated its Rice Genome Research Program (RGP). The aim of this programme was to clarify the genome sequence of rice in entirety (Sasaki 1998). The RGP was implemented in two phases: the first phase of 7 years dealt with the mapping of the rice genome which included large-scale cDNA analysis, genetic mapping by construction of fine-scale RFLP map and YAC-based physical mapping; the second phase was initiated in 1998, when RGP stepped into the new era of genome sequencing and matured into the International Rice Genome Sequencing Project (IRGSP). This section highlights the first phase when the main motive was to construct a genome-wide physical map. This phase was coupled with advances in the field of computation in terms of next-generation web technologies and bioinformatics analysis platforms, which became extremely important for scientists and bench biologists to visualise, annotate and analyse the high-throughput experimental data.

### 2.1.1 Linking Genomics and Genetics

RGP adopted cDNA analysis, a quick and easy strategy to clone several genes expressed in rice. This involved random cloning, partial sequencing and developing cDNA libraries from various tissues at different developmental stages, for example, green and etiolated seedlings, young roots, panicles at the flowering stage and calluses cultured with 2,4-dichlorophenoxyacetic acid (Yamamoto and Sasaki 1997). By 1997, 36,000 cDNA clones from 15 main cDNA libraries were sequenced for 400–500 5'-terminal bases (Sasaki 1998). Additionally, more than 40,000 cDNA clones were partially sequenced to generate ESTs. Therefore, sequencing the cDNA from these samples helped: (1) to designate ESTs for genomic regions of expressed genes, (2) to understand the expression profile of genes from various tissues at separate growing stages and (3) to functionally annotate the genes depending on similarity search in public databases. Generating ESTs was of profound importance as it was exploited for linkage analysis by utilising RFLP markers and also for determining hinge markers for YAC contigs. So to link the information in the DNA sequence to phenotypic traits, the basic tool required for rice genome analysis is a genome-wide physical-genetic map.

At the time, a high-density genetic linkage map based on polymorphisms within DNA sequences, such as RFLPs, CAPSs and simple sequence repeats (SSRs), was derived from F2 plants from a single cross between the *japonica* variety Nipponbare and the *indica* variety Kasalath (Harushima et al. 1998). It was composed of 2275 genetic markers with a cumulative genetic distance of 1550 cM for 12 linkage groups. Nearly 70% of the DNA markers were rice ESTs (clones from Nipponbare callus, root and shoot libraries), while the remaining were clones from genomic



DNA and cDNA of rice and other cereal crops. The position of the centromere, assigned using the secondary trisomics and telotrisomics developed at the International Rice Research Institute (Philippines), revealed the chromosomal orientation of individual linkage groups and suggested that the meiotic recombination frequency is very low at each centromere but reasonably high in the flanking regions (Singh et al. 1996). Additionally the clarity on the knowledge of DNA markers on the genetic map was used for accurate genotyping of candidate progenies obtained by backcrossing for a particular trait (Yano et al. 1997).

A physical map assembled using YACs was also available. The YAC library was derived from Nipponbare variety and comprised of ~7000 clones with an average insert size of 350 kb (Umehara et al. 1995). Thus the library size was 5.5 times of the rice genome (430 Mb), indicating overlapping YAC clones. Also chimerism (non-contiguous DNA fragments present within the same YAC clone) and difficulty of separating YACs from other yeast chromosomes posed challenges in using YACs as templates for DNA sequencing. This redundancy was removed by the following strategies: (1) colony hybridisation of YACs using RFLP markers, (2) using sequence-tagged site markers for identifying positive YACs and (3) mapping many ESTs onto YAC physical map, by assembling them using primers designed from the 3' untranslated region (UTR), specific for each gene (Antonio et al. 1996, Saji et al. 1996, Shimokawa et al. 1996, Umehara et al. 1996, 1997, Wang et al. 1996, Koike et al. 1997, Tanoue et al. 1997). BAC/PAC vectors were also used extensively to construct new rice genomic libraries (Budiman et al. 1999; Baba et al. 2000). Clones were generated using several combinations of restriction enzymes (e.g. Sau3AI for PACs and HindIII and EcoRI for BACs), partially sequenced from 5' and 3' termini and subsequently fingerprinted to develop a BAC physical map. Similarly as mentioned before, PAC contigs were identified by EST and confirmed by fingerprinting. Sequencing of the selected PAC contigs resulted in a significant coverage of the rice genome and indicated a higher gene density than expected.

The combination of ESTs obtained from extensive cDNA analysis, DNA markers identified on RFLP linkage maps and the information derived from the physical maps generated from YAC clones and BAC/PAC libraries indicated to a large extent the exact location of the expressed genomic region. A properly annotated genome-wide physical-genetic map was the need of the hour, such that each locus, supplemented with positional information on the chromosomes, could be associated with a phenotypic trait, thereby facilitating the cloning of biologically and agronomically important genes. RGP was a prerequisite for generating a reliable sequence-ready physical map. These efforts not only prepared the foundation for rice genome sequencing, but triggered and inspired the entire sequencing project that was to be undertaken in the subsequent years worldwide (Sasaki 1998; Sasaki and Burr 2000).

### 2.1.2 Genomics Meets Informatics

The sequences that were generated by the concerted efforts were subsequently annotated by searching with various software and prediction tools against databases. The following approaches were adopted to generate a final annotation of genes and associated elements, along with assigning their proper coordinates in the genome

sequence: (a) BLASTX to search the nonredundant protein database, (b) BLASTN to search the rice EST database, (c) GENSCAN to predict open reading frames (ORFs) within the sequence, (d) SplicePredictor to identify potential exon-intron splice sites, (e) Miropeats to predict inverted or tandem repeats and assemble the shotgun sequence and (f) using *gag* and *pol* genes as references to recognise transposable elements, followed by identification of long terminal repeats (LTRs) (Parsons 1995; Kleffe et al. 1996; Altschul et al. 1997; Burge and Karlin 1997). Similarly other tools such as Gene Finder, GeneMark and NetPlantGene were also employed (Hebsgaard et al. 1996; Solovyev and Salamov 1997; Lukashin and Borodovsky 1998). Even after integration and correlation of all this genetic information with the genome sequence, powerful information retrieval system was necessary to easily access and manipulate the data and link this information to other genetic resources.

A new rice genome database, called INtegrated rice genome Explorer (INE, pronounced as 'i-ne', also referring to rice plant in Japanese language), was developed to accelerate the release of all this genetic information to the public (Sakata et al. 2000). This web interface based on a Java applet allowed rapid viewing capability of the integrated maps, accompanied by smooth navigation options. INE incorporated and integrated large data volumes from various experimental outputs to enhance the overall density of the markers within the genome. The high-density linkage map with 2275 DNA markers also included the image of each probe used for RFLP and the sequence of the corresponding markers (Harushima et al. 1998). The physical map generated using the YAC clones covered a significant portion of the genome and also showed the actual physical distance between genetic markers. Incorporation of ESTs in the INE allowed review of syntenic relationships and also enhanced the significance of high-density markers. Also a physical map constructed from sequenced PAC clones was incorporated to enhance the reliability of the data. Additional quality control was also applied to ensure good-quality data. INE provided a page for accessing 'Low Quality Information', which harboured sequences that failed to meet the standards or any thresholds set up by RGP. INE not only increased its extent of data incorporation but also allowed integrated view of the data for efficient data mining. Each chromosome was associated with an integrated map which showed the linkage map, the physical map constructed from ordered YAC clones, the EST map and the PAC contigs. Therefore the value of genetic information was greatly enhanced with this integrated display. The following examples corroborate the usefulness of the genetic map: by integrating the PAC contig map or an EST map into the linkage map, (1) the existence of a PAC or EST adjacent to the genetic marker of interest was verified, (2) the exact positions of genetic markers and other genetic elements were evaluated, and (3) functional annotation of a PAC near a genetic marker revealed an EST within a PAC contig responsible for a particular trait. These would facilitate subsequent map-based cloning of agronomically desirable genes. INE was later modified to meet the demands of the genome sequencing project (Sasaki and Burr 2000). Upon further elucidation of signal transduction pathways in rice, more valuable information on the physiological and biochemical aspects of rice genes was incorporated into INE. For

exploration of syntenic relationships of rice with other cereal crops, INE was subsequently linked to other genomic resources of important cereal crop species.

---

### 3 Sequencing the Rice Genome

The genome-wide physical-genetic maps could leverage genomics if the nucleotide sequence of the rice genome was available. In fact, the DNA sequence can bridge the gap between the structural organisation of the genome and its functional dimensions. Towards the end of the twentieth century, with the progress in sequencing methodologies, adoption of automation and advances in computation, several genome sequencing projects of higher organisms were undertaken, including *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana* and *Homo sapiens* (Mewes et al. 1997; The C. elegans Sequencing Consortium 1998; Adams et al. 2000; The Arabidopsis Genome Initiative 2000; International Human Genome Sequencing Consortium 2001; Venter et al. 2001). In that new era of sequence-based genomic research, the obvious choice for a genome sequencing project on a cereal crop was rice. By then the rice genome was well mapped because of the decade-long extensive research on rice genomics, thereby providing a solid foundation for the upcoming sequencing project. Among the major cereal crop genomes, rice has the smallest genome of 430 Mb. Moreover, rice serves as an excellent model as it shares a large extent of synteny with other grass genomes, like rye, wheat, maize, barley and sorghum, and can be easily transformed and genetically manipulated (Tyagi and Mohanty 2000). With the establishment of the International Rice Genome Sequencing Project, in the second phase of RGP, rice became the first crop and the first monocot genome to be sequenced. The following section discusses the inception of IRGSP, inclusion of the participating nations, adoption of optimal methodologies for sequencing, choice of rice cultivar, nucleotide sequence accuracy, finishing standards and sequence release policies.

#### 3.1 International Rice Genome Sequencing Project (IRGSP), Established 1997

















The International Rice Genome Sequencing Project (IRGSP) was initiated at a workshop in September 1997, at the International Symposium on Plant Molecular Biology held at Singapore. Scientists and representatives from several nations attended the workshop and agreed to collaborate internationally to sequence the rice genome (Eckardt 2000). To avoid any instance of allelic polymorphism, a single variety of rice cultivar was used as the common source of DNA: *Oryza sativa* ssp. *japonica* cultivar Nipponbare. Participants from Japan, Korea, China, the United Kingdom and the United States decided to share and contribute materials and agreed to release physical maps and DNA sequences to the public databases in a timely manner. The IRGSP eventually evolved to include 11 nations (listed in Table 2) and

also revised their policies and finishing standards to ensure a high-quality sequence-ready genome map of rice. Like other large-scale genome sequencing projects accomplished at the time, the modus operandi for this project was similar (Sasaki and Burr 2000; International Rice Genome Sequencing Project and Sasaki 2005). Large insert genomic libraries constructed in BACs and PACs were used as the primary sequence templates. Using polymerase chain reaction (PCR) screening, fingerprinting, end-sequencing and physical mapping, minimal tiling paths were constructed such that a contiguous set of minimally overlapping BAC/PAC clones can be anchored to physical positions along the length of the chromosome, thereby generating a sequence-ready BAC/PAC contig. These clones belonging to a contig were subjected to a map-based hierarchical clone-by-clone shotgun approach, to produce shotgun libraries, which were then sequenced and assembled using *in silico* tools to reconstitute the entire intact insert within the corresponding BAC/PAC clone (see Fig. 2). The second focus of this project was annotating the raw nucleotide sequence data and associating useful biological information with it, for example, positioning of genes, prediction of splice sites, transcriptional start sites and regulatory sites, repeat elements, prediction of tRNAs, analyses of the predicted proteins to dissect any functional domains or motifs, etc. Therefore this mammoth task, inclusive of so many dimensions, finally culminated into a high-quality genome sequence of rice. The following sections discuss the methodologies adopted by IRGSP and some significant outcomes of the rice genome analysis.

### 3.1.1 Physical Map and Sequencing










The hierarchical clone-by-clone methodology adopted by IRGSP used a high-density genetic linkage map generated using a single F2 population (2275 markers), ESTs, YAC- and BAC-based physical maps and 2 draft sequences from 2 separate private sources, Monsanto (now Pharmacia, New Jersey) and Syngenta (San Diego) (Yamamoto and Sasaki 1997; Harushima et al. 1998; Mao et al. 2000; Barry 2001; Saji et al. 2001; Chen et al. 2002; Goff et al. 2002; Wu et al. 2002). In total, nine genomic libraries from *Oryza sativa* ssp. *japonica* cultivar Nipponbare were utilised to establish the physical map (International Rice Genome Sequencing Project and Sasaki 2005). Marker-aided PCR screening, fingerprinting with restriction enzymes and end-sequencing of approximately 3400 BAC/PAC clones were used to construct the minimal tiling paths. Majority of the physical gaps in the BAC/PAC tiling path were filled with the help of PCR fragments, 10-kb plasmids and 40-kb fosmid clones. A typical workflow with BAC clones included the following steps: (1) fingerprinting of HindIII and EcoRI BAC libraries, (2) assembly of the fingerprinted BACs into contiguous contigs, (3) anchorage of these BACs onto the physical map with RFLP and end-sequencing analysis and, finally, (4) connection and extension of contigs by chromosome walking. Another technique that was adopted for extending contigs was the use of 'overgo probes' (Eckardt 2000). 24-bp sequences were designed from BAC end-sequences with an 8-bp overlap. The 24-bp sequences were then joined to create a 40-bp 'overgo', which was then used to probe a high-density BAC library in order to search for additional BAC clones that may potentially extend a contig. Using similar workflows, plant ESTs were also mapped onto the physical map of rice chromosomes. The PAC, BAC and fosmid clones on the

**Table 2** Chromosome assignments for sequencing for IRGSP participants

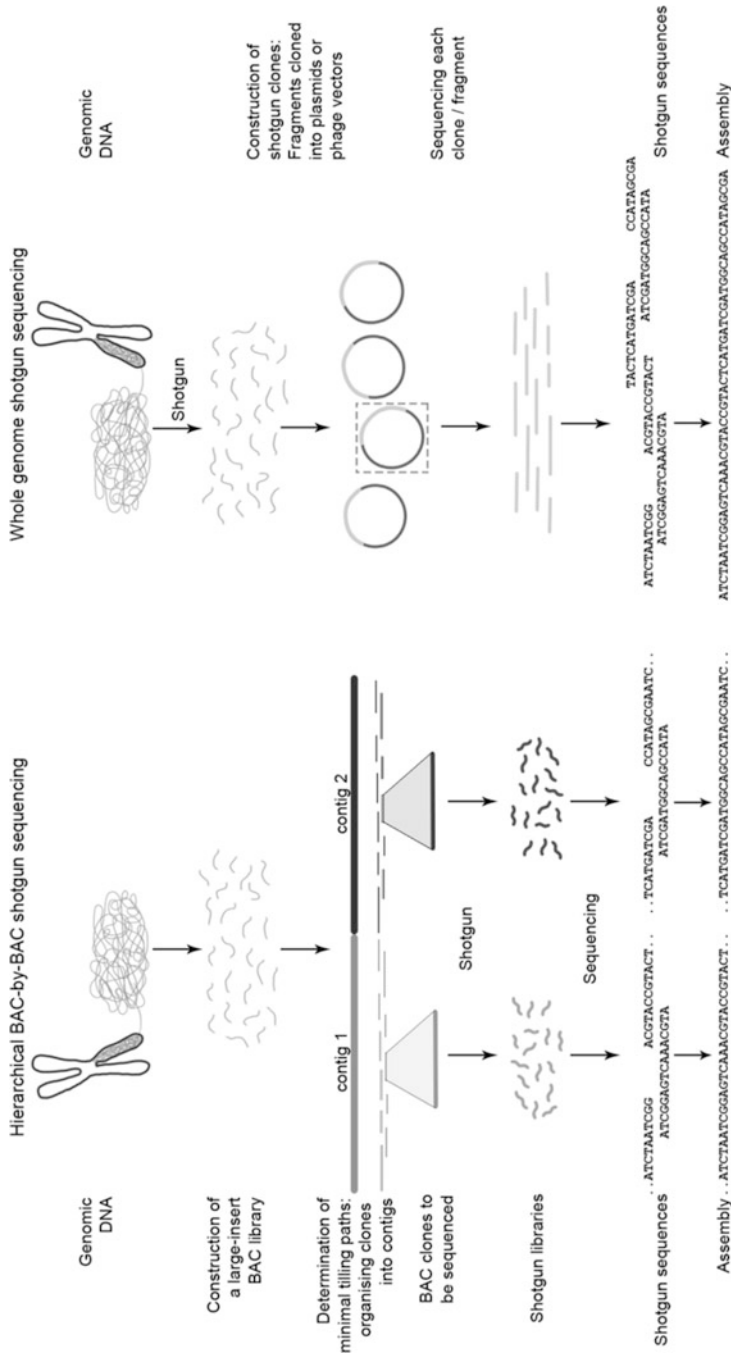
| Rice chromosomes | Participating nations  | Initiatives and/or institutions (acronym)   |
|------------------|--|---|
| 1,6,7,8          | Japan<br><br>  | Rice Genome Research Program (RGP)  |
| 3,10,11          | United States of America<br><br><br><br> | <ul style="list-style-type: none"> <li>• Arizona Genomics Institute (AGI) and Arizona Genomics Computational Laboratory (AGCol)</li> <li>• Cold Spring Harbor Laboratory (CSHL)</li> <li>• Washington University School of Medicine Genome Sequencing Center</li> <li>• University of Wisconsin–Madison</li> <li>• The Institute for Genomic Research (TIGR)</li> <li>• Clemson University Genomics Institute (CUGI)</li> <li>• Plant Genome Initiative at Rutgers (PGIR), Waksman Institute, Rutgers University</li> </ul> |
| 1                | Korea<br><br>  | <p>Korea Rice Genome Research Program (KRGRP)</p> <ul style="list-style-type: none"> <li>• National Institute of agricultural science and technology</li> </ul>   |
| 2                | United Kingdom<br><br>   | <ul style="list-style-type: none"> <li>• John Innes Centre (JIC)</li> </ul>   |
| 4                | China<br><br>  | <p>National Center for Gene Research (NCGR)</p> <ul style="list-style-type: none"> <li>• Shanghai Institutes for Biological Sciences Chinese Academy of Sciences (CAS)</li> </ul>   |
| 5                | Taiwan<br><br>   | <p>Academia Sinica Plant Genome Center (ASPGC)</p> <ul style="list-style-type: none"> <li>• National Cheng Kung University</li> <li>• National Yang-Ming University</li> </ul>  |
| 9                | Thailand<br><br>   | <p>National Center for Genetic Engineering and Biotechnology (BIOTEC)</p> <ul style="list-style-type: none"> <li>• Rice Gene Discovery Unit, Kasetsart University</li> </ul>  |

(continued)

**Table 2** (continued)

| Rice chromosomes | Participating nations   | Initiatives and/or institutions (acronym)   |
|------------------|---|---|
| 9                | Canada<br><br><br> | <ul style="list-style-type: none"> <li>• McGill University</li> <li>• York University</li> </ul>  |
| 11               | India<br><br>   | Indian Initiative for Rice Genome Sequencing (IIRGS) <ul style="list-style-type: none"> <li>• University of Delhi South Campus (UDSC)</li> <li>• Indian Agricultural Research Institute (IARI)</li> </ul> |
| 12               | Brazil<br><br>  | Brazilian Rice Genome Initiative (BRIGI) <ul style="list-style-type: none"> <li>• Centro de Genomica e Fitomelhoramento, UFPel</li> </ul>   |
| 12               | France<br><br>   | Genoscope <ul style="list-style-type: none"> <li>• Centre National de Séquençage, INRA-URGV and CNRS</li> </ul>   |

physical map were next subjected to a shotgun sequencing approach, originally employed by RGP. In this procedure, the extracted DNA from individual PAC/BAC clones (ranging between 100 and 200 kb) belonging to a sequence-ready contig was subjected to random shearing by sonication or nebulisation, following which the fragments were subcloned to produce shotgun libraries with an average insert size of 1 to 3 kb (see Fig. 2). Random clones from the shotgun libraries were then sequenced, using both universal primers and the dye-terminator or dye-primer methods, to acquire the desired degree of ‘coverage’ of the total sequence. The shotgun sequencing approach finally led to an assembled, ordered and finished quality rice genome, with approximately tenfold sequence coverage and less than 1 error per 10,000 bases. As mentioned above, small physical gaps were bridged by long-range PCRs. FISH and optical mapping were the other two techniques that were adopted for the remaining gaps. The profound application of FISH to rice mapping was already well established (Jiang et al. 1995; Cheng et al. 2001b). FISH



**Fig. 2** Schematic workflow of two distinct shotgun sequencing approaches used in rice genome sequencing: hierarchical clone-by-clone approach was adopted by IRGSP and Monsanto (*left*), while whole-genome shotgun approach was employed by Syngenta (*right*)

was established as a useful technique for easy identification of rice chromosomes, determination of physical positions of uncertain clones and examination of the physical nature of large linkage gaps. The last feature facilitated sequencing at chromosomal ends, centromere and telomere regions and turned out to be effective for characterising BAC clones that contain complex repetitive DNA sequences, very common in rice (Moore et al. 1995). The optical mapping technique was previously used to create whole-genome restriction maps of the microorganisms *Deinococcus radiodurans* and *Plasmodium falciparum* (Lin et al. 1999; Lai et al. 1999). In this technique, fluid flow capillary action was employed to extend and align DNA molecules onto a specially prepared glass surface. DNA was then digested with restriction enzymes, and fluorescence microscopy imaging was finally used to map the fragments into an ordered array. Typically, charting a minimal tiling path (i.e. a subset of clones with minimally overlapping sequences) which encompasses a chromosome or a region of interest along a chromosome is dependent on the physical map and the assembly of sequences from a selected subset of clones in an unambiguous fashion with help from their overlapping regions. Unfortunately, the centromeric regions of majority multicellular eukaryotic organisms emerge to be resistant to this method, largely because they contain highly repetitive satellite DNA sequences and transposable elements (Eckardt 2004). The recalcitrance of these regions impeded the efforts to understand their functionality. Findings from previous cytological work indicated that rice centromeres contain multiple repeats of a 155-bp satellite DNA sequence called CentO and many rice chromosomes limited amounts of the satellite DNA repeat sequences compared to other species (Dong et al. 1998; Cheng et al. 2002). The centromere of rice chromosome 8 (Cen8) had the least amount of satellite DNA sequence (~64 kb) among the 12 rice chromosomes and was an obvious choice for obtaining the sequence for this region. Two groups put significant efforts towards this goal: Wu et al. (2004) employed repeated subcloning, transposon-based sequencing and fingerprinting of BAC/PAC clones, while Nagaki et al. (2004) used similar methods, mapped the Cen8 region using CentO repeat sequences and centromere-specific retrotransposon sequence probes and constructed a minimal tiling path of 12 BAC clones encompassing this region (Nagaki et al. 2004; Wu et al. 2004). Breakthrough findings reported from two groups indicated a large fraction of the Cen8 sequence to contain known repetitive elements, like inverted repeats, transposons, *gypsy*-type retrotransposons, CentO satellite repeats, etc. Therefore to fill in the large physical gaps arising from the centromeric regions in the sequences, they were estimated on the basis of the CentO satellite DNA content of the respective chromosomes (International Rice Genome Sequencing Project and Sasaki 2005). The length of CentO arrays therefore provided approximate measures of centromere gaps, telomere gaps and chromosomal arms.

### **3.1.2 Integration of Draft Sequences from Monsanto and Syngenta with IRGSP**

In early 2000, Monsanto announced that the company was set to release a draft sequence of the rice genome. The sequences would be made available to academic scientists under an access agreement with Monsanto. A separate agreement between



the company and IRGSP facilitated incorporation of the Monsanto BAC clones and sequences into the IRGSP sequences. The Monsanto rice genome sequencing project also undertook the BAC-to-BAC sequencing approach on ~3400 BAC clones of the cultivar Nipponbare (Barry 2001). The draft sequence represented 393 Mb of the rice genome, with ~5X coverage. This development significantly accelerated public efforts to sequence the genome. Drafts of the rice genome sequence from ssp. *japonica* cultivar Nipponbare were completed and published separately by Syngenta (San Diego) in 2002 (Goff et al. 2002). This draft, generated using whole-genome shotgun sequencing approach, represented 93% of the rice genome (~390 Mbp), with ~6X coverage, and a 44% GC content. The Syngenta draft sequence, also referred to as Syd, provided useful insights into the rice genome and proteome. Gene predictions on the assembled sequences revealed 32,000 to 50,000 genes contained by the genome, with homologs in maize, wheat, barley and other cereal crops. Extensive synteny and gene homology was also indicated between rice and other cereal genomes. Another draft sequence of the *O. sativa* subsp. *indica* cultivar 93-11 was released by the Beijing Genomics Institute (BGI), which was also generated through a whole-genome shotgun sequencing approach (Yu et al. 2002). This represented ~360 Mb of the genome and emerged to be an important resource for gene discovery, comparative genomics, syntenic associations and SNP discovery. BGI sequence analysis suggested that the rice genome encodes for 46,000–55,000 proteins, which was consistent with the estimate made from the Syd data. The BGI data were made publicly available through GenBank/DDBJ/EMBL, unlike Monsanto or Syngenta. Contigs from BGI and Syngenta whole-genome assemblies were aligned against the IRGSP Nipponbare pseudomolecules using MUMmer (Delcher et al. 1999). The number of IRGSP full-length cDNA-supported gene models that aligned with the contigs was tabulated. To estimate the location of centromeres, the 155-bp CentO consensus sequence was used as BLAST query against 93-11 and Nipponbare whole-genome shotgun contigs, and their coordinates were recorded (International Rice Genome Sequencing Project and Sasaki 2005). The CentO sequence positions on each IRGSP pseudomolecule for a particular chromosome served as the benchmark to estimate the location of centromeres for each *indica* chromosome.

### 3.1.3 Assembly and Annotation

Fingerprinted BACs were assembled into contigs using the software called FingerPrinted Contigs (FPC) (Soderlund et al. 2000). The shotgun sequences were assembled into intact inserts within the BAC/PAC clone using PHRED, PHRAP (<http://www.phrap.org/phredphrapconsed.html>) and CONSED software packages or using the TIGR Assembler (Sutton et al. 1995; Ewing et al. 1998; Ewing and Green 1998; Gordon et al. 1998). The primary goal of IRGSP was to obtain a complete and accurate sequence of the rice genome. Consistent with the Bermuda sequencing standards (the international human genome sequencing community organised meetings in Bermuda in 1996 and 1997 to set finishing standards for DNA sequences), ‘complete’ was initially defined as less than 1 error in 10,000 bases (Eckardt 2000; International Rice Genome Sequencing Project and Sasaki 2005). An

empirically determined PHRAP score of 30 or above was found to be consistent with the level of accuracy. Sequence ambiguities, therefore, were indicated by PHRAP scores less than 30 and were usually regions containing repetitive DNA sequences. These problem regions were resolved by changing sequencing chemistries or using alternate polymerases. Assemblies of BAC and PAC clones were checked for accuracy by comparing the experimentally determined fingerprint patterns with computationally derived patterns of restriction enzyme digests. Sequence quality was also tested by comparisons of overlapping sequences obtained independently.

Another important step in the rice genome sequencing project was the construction of pseudomolecules, which were generated by using an in-house, database-based, semiautomated process (<http://rice.plantbiology.msu.edu/>). These are virtual contigs that are constructed by selecting an optimal tiling path of BAC/PAC clones representing all the 12 chromosomes and by resolving discrepancies between overlapping BAC/PAC clones. Overlap regions were trimmed in accordance with information from similarity searches and tiling path information, while physical gaps were represented by a stretch of 1000 nucleotides. The overlapping clones were also validated by comparing the pseudomolecule sequence with the optical map for rice. The unique sequences were then linked to form a contiguous sequence for each rice chromosome. Location of centromeres was ascertained on the basis of gaps and presence of the CentO repeats. Information on gene models and features associated with the BAC/PAC clones were transferred to the pseudomolecules, which underwent quality control using the whole-genome shotgun sequences of Nipponbare. All BAC/PAC clones were not incorporated into the pseudomolecules due to different reasons: some belonged to subspecies *indica*, some clones were unfinished or redundant or overlapping, some were released later than the release of the version of pseudomolecules, while some were difficult to map to their correct positions. Sequences from BACs and the Syngenta whole-genome shotgun assembly of Nipponbare that could not be aligned on the existing pseudomolecules were represented on separate pseudomolecules, the Unanchored and the Syngenta chromosomes. Release of this assembly of pseudomolecules, termed Os-Nipponbare-Reference-IRGSP-1.0, was made in collaboration with the Agrogenomics Research Center at the National Institute of Agrobiological Sciences, Tsukuba, Japan. The annotation files for the pseudomolecules exclude partial genes, pseudogenes and small gene models (<50 amino acids).

Many bioinformatics resources were used to leverage sequence annotation of the rice genome. Several gene prediction software like GENSCAN, FGENESH, RiceHMM, GlimmerM and MZEF were employed to predict gene models (Burge and Karlin 1997; Zang 1997; Sakata et al. 1999; Salamov and Solovyev 2000; Pertea and Salzberg 2002). Monocot trained versions of the tools were used for accuracy. Gene models that corresponded to organellar DNA, those with incomplete open reading frames and those encoding proteins less than 50 amino acids were omitted. The SplicePredictor programme was employed to correctly ascertain the exon-intron boundaries (Kleffe et al. 1996). Another automated annotation system, RiceGAAS (the Rice Genome Automated Annotation System), was also used extensively to predict genes and long terminal repeat sequences on the basis of homology search

(Sakata et al. 2002). The predicted proteome was analysed, using various tools from RiceGAAS, for the presence of any functional domains (ProfileScan and MOTIF), intracellular localisation (PSORT) and solubility (SOSui). Conserved domains and motif searches and deciphering the gene ontology (GO) associations were completed using InterproScan in combination with Interpro2Go programme (Zdobnov and Apweiler 2001; Camon et al. 2005). The predicted proteome of rice was also searched using BLASTP against the proteome of several model species, including *Arabidopsis*, to identify homologs that could reflect on the potential functions of the genes. MUMmer and RepeatMasker software were used to delineate repetitive sequences in the rice genome (Delcher et al. 1999; Smit et al. 1996–2000). The Simple Sequence Repeat Identification Tool was used to identify SSR motifs and estimate the copy number of SSR markers (Temnykh et al. 2001). The TIGR Oryza Repeat Database, together with other sequence data on rice transposable elements, was used to create a rice transposable element database (RTEdb) (Juretic et al. 2004). The coordinates of these transposable elements were determined on the rice pseudomolecules. Organellar DNA sequences were also used to mask pseudomolecules, using BLASTN and MUMmer. Prediction of noncoding RNAs was also addressed. Prediction of transfer RNA genes was executed by the programme tRNA-scanSE (Lowe and Eddy 1997). miRNAs, spliceosomal and snoRNA sequences were retrieved from the Rfam database (<http://rfam.xfam.org/>) and used as BLASTN queries. In a different approach, experimentally validated miRNAs from other species, excluding *Arabidopsis*, were used for BLASTN queries against the pseudomolecules. To identify single nucleotide polymorphisms in the Nipponbare cultivar, end-sequencing of BAC clones from an *O. sativa* ssp. *indica* var. Kasalath BAC library was carried out, and the sequences were subjected to BLASTN analysis against the pseudomolecules (International Rice Genome Sequencing Project and Sasaki 2005).

### 3.1.4 Salient Features of the Rice Genome Reported by IRGSP

In 2005, the IRGSP presented the map-based, finished quality sequence of the rice genome that at the time included virtually all the euchromatin and two complete centromeres (International Rice Genome Sequencing Project and Sasaki 2005). The estimated genome size of the rice nuclear genome was 388.8 Mb (~389 Mb). The draft sequence was used to construct pseudomolecules representing the 12 rice chromosomes (haploid number), which were expected to cover 95.3% of the entire genome and 98.9% of the euchromatin. In a separate representation, 8391 of 8440 unique EST markers (99.4%) could be identified in these pseudomolecules. The centromere allocations were done for two chromosomes, 4 and 8. As discussed previously, all rice centromeres contain the repetitive CentO satellite DNA sequences, along with flanking centromere-specific retrotransposons (Dong et al. 1998; Cheng et al. 2002). Complete sequencing of the centromeric regions for chromosomes 4 and 8 indicated these chromosomes to contain ~59 kb and ~69 kb of clustered CentO repeats, respectively, arranged in tandem head-to-tail arrays within the clusters (Nagaki et al. 2004; Wu et al. 2004; Zhang et al. 2004). These CentO clusters varied in length and orientation between the two chromosomes and

were frequently associated with centromere-specific transposable elements, like the retrotransposon RIRE7.

To figure out the gene content, distribution and density in the genome, the pseudomolecules were masked for repetitive sequences, and the gene prediction tools were employed to identify potential non-transposable element-related genes. At the time, a total of 37,544 non-transposable element-related protein-coding sequences were predicted, of which 22,840 genes could be correlated with ESTs or full-length cDNA models (International Rice Genome Sequencing Project and Sasaki 2005). A lower gene density of one gene per 9.9 kb was predicted in rice. Homolog search in cereals and dicots like *Arabidopsis* indicated a total of 2859 genes unique to rice and other cereals, which provided useful cues for differentiating between monocots and dicots. While majority of the genes encode for proteins with unknown function or hypothetical proteins, families of seed storage proteins called prolamins, hormone response proteins and defence proteins, such as proteinase inhibitors, chitinases, pathogenesis-related proteins and seed allergens, are absent in dicots. But with improvements in gene annotation methods, the number of true protein-coding genes in rice has been revised in the present day. Functional classification of 2296 rice genes according to agronomically desirable traits indicated the following percentage of genes in rice to have distinct functional roles (Wing et al. 2018):

Stress response 12.2% (367), disease resistance 7.4% (221), insect resistance 1.0% (31), growth and development 17.1% (513), phytohormone 15.8% (472), flower organ and heading date 9.2% (276), fertility 5.8% (174), yield 6.3% (189), grain quality 2.1% (63), domestication 1.0% (28), nutrient-use efficiency 6.9% (207) and others 15.2% (455).

Functional analysis of genes and relating them to phenotypes is the basic essence of rice research, and gene disruption with transposable elements is still an ongoing trend. *Tos17* is an endogenous *copia*-like retrotransposon in rice, which prefers to insert into gene-rich regions, making it a suitable tool for functional analysis of the rice genes by gene knockout (Hirochika et al. 1996). A total of 11,487 *Tos17* retrotransposon insertion sites were identified within the rice genome at the time, of which 3243 were in genes, suggesting the density of *Tos17* insertions to be higher in euchromatic regions, compared to retrotransposons, which were frequently found in the paracentromeric regions (International Rice Genome Sequencing Project and Sasaki 2005). Another interesting outcome of the annotation was a large percentage of genes were arranged in tandem repeats, for example, the genes that encoded for protein kinase domains. Predictions of various types of noncoding RNAs distributed across the rice pseudomolecules indicated a total of 763 transfer RNA genes, 158 miRNAs, 215 small nucleolar RNAs (snoRNA) and 93 spliceosomal RNA genes. rDNA loci could also be ascertained on a number of chromosomes, including the nucleolar organiser at the telomeric end of the short arm of chromosome 9.

To accommodate the original endosymbionts inside the cells as organelles, a continuous transfer of organellar DNA to the nucleus has took place in the course of evolution, diminishing the sizes of the organellar genomes to their present-day sizes. Based on the parameters used during homology searches, IRGSP detected 421–453 chloroplast insertions and 909–1191 mitochondrial insertions in the rice genome,

and on the basis of their distribution across the 12 chromosomes, it was suggested that mitochondrial and chloroplast transfers occurred independently (International Rice Genome Sequencing Project and Sasaki 2005). The rice genome was also found to be populated with transposable elements, belonging to Class I, Class II or neither of the two families. The transposon content of rice was predicted to be at least 35%, with elements belonging to all known transposon superfamilies. The Class I elements (e.g. long terminal repeats (LTRs) and non-LTR retrotransposons) were less in number compared to Class II elements (e.g. *hAT*, *CACTA*, *IS256/Mutator*, *IS5/Tourist* and *IS630/Tc1/mariner* superfamilies), and majority of Class I elements were found to be concentrated in heterochromatic regions such as the centromeric and pericentromeric regions. Class I SSRs which behave as hypervariable loci were also found in the rice genome. These repeats of >20 nucleotides in length have immense application in genetics and breeding and have been shown to be polymorphic in diverse cultivars of rice (McCouch et al. 2002). A total of 18,828 Class I SSRs (di-, tri- and tetra-nucleotide) were identified and annotated on the rice genome that represented 47 distinct motif families. Finally, intraspecific sequence polymorphisms were detected between the two cultivated rice subspecies, *japonica* and *indica*. Comparison of orthologous sequences between *O. sativa* ssp. *indica* cv. Kasalath and *O. sativa* ssp. *japonica* cv. Nipponbare identified a total contingent of 80,127 polymorphic sites, resulting in a high-resolution genetic map for rice. Insertions and deletions were also detected (International Rice Genome Sequencing Project and Sasaki 2005; Yu et al. 2002).

Genome-wide comparisons of the two published whole-genome shotgun assemblies of draft-quality rice sequences with the IRGSP pseudomolecules were also done. These assemblies predicted genome sizes of 433 Mb for *O. sativa* ssp. *japonica* cv. Nipponbare (6X coverage) and 466 Mb for *O. sativa* ssp. *indica* cv. 93-11 (6.28X coverage), which differed from the 389 Mb genome size reported by IRGSP. A detailed comparison of the sequences indicated that a substantial percentage of the contigs in both assemblies were misaligned or nonhomologous or provided duplicate coverage, suggesting the draft sequences to be inadequate for gene annotation or functional genomics.

### 3.2 Updates on the Current Rice Genome Data

The Michigan State University Rice Genome Annotation Project Database and Resource (MSU RAP-DB) is a project funded by the National Science Foundation (NSF). This provides sequence and annotation data for the rice genome, using the sequence of IRGSP, that of *Oryza sativa* ssp. *Japonica* cv. Nipponbare (International Rice Genome Sequencing Project and Sasaki 2005). The parallel and complementary annotation efforts of the MSU RGAP and IRGSP/RAP are unified on the set of pseudomolecules that represent the 12 chromosomes, and therefore the gene loci, gene models and associated annotations are comparable. The findings from the latest unified Os-Nipponbare-Reference-IRGSP-1.0 pseudomolecules and MSU Rice Genome Annotation Project Release 7 are summarised in Table 3 (Kawahara et al.

**Table 3** Summary of MSU RGAP Release 7 (2011)

|                    |         |
|--------------------|---------|
| Total loci         | 55,986  |
| Non-TE loci number | 39,045  |
| Gene models        | 49,066  |
| Gene size          | 2853 bp |
| Exons/gene         | 4.9     |
| Introns/gene       | 3.9     |
| TE loci number     | 16,941  |
| Gene models        | 17,272  |
| Gene size          | 3223 bp |
| Exons/gene         | 4.2     |
| Introns/gene       | 3.2     |

*TE* transposable element-related genes and gene models, *Non-TE* non-transposable element-related gene models. In Release 7, there were 373,245,519 bp of nonoverlapping rice genome sequence pertaining to the 12 rice chromosomes, derived from 3184 BAC/PAC clones. 55,986 genes (loci) had been identified, of which 6457 had 10,352 additional alternative splicing isoforms resulting in a total of 66,338 transcripts (or gene models) in the rice genome. Small gene models (<50 amino acids) were excluded. Two approaches were employed for identification of TE-related gene models: BLASTN searches against the MSU Oryza Repeat Database and by identifying gene models containing TE-related Pfam domains

2013). Continuous revision and revalidation of the minimal tiling paths were undertaken to update the sequence assembly. Advances in sequencing strategies, platforms and whole-genome sequencing allowed identification of sequencing errors in the revised assembly. Resequencing the genome of two different Nipponbare individuals using the Illumina Genome Analyzer II/IIx platform identified a total of 4886 sequencing errors in 321 Mb of the assembled genome. This indicated an error rate in the original IRGSP assembly to be only 0.15 per 10,000 nucleotides. Five insertions/deletions were also identified using the Roche 454 pyrosequencing platform. This resequencing data, generated from two different individuals, gave information regarding allelic differences between the original Nipponbare individual used in the IRGSP and the two individuals used in this resequencing effort.

Sequence annotation efforts have designated ~400 distinct gene families in rice. In the Pfam database (<http://pfam.xfam.org/>), which harbours a large collection of protein families, 42,365 domains are reported for 48,930 protein sequences from the rice proteome, while 5891 unique domain organisations or architectures are also listed. The kinase domain (Pkinase) is maximally represented in the proteome, with currently 517 sequences containing this unique domain. This is in agreement with several instances of tandem repeats, already known for the rice genome.

## 4 Systems Biology Resources for Rice Genomics

The major challenge after achieving completion of ‘gold standard’ rice genome sequence was to organise all the available primary and secondary data from diverse experimental sources. INE, discussed previously, served as a unified platform for

integrated map displays for each chromosome (Sakata et al. 2000). But advances in bioinformatics have led to development of species-specific or general databases, which allowed efficient genome browsing, extension of the genetic information to multiple plant species and other organisms as well as cross-referencing data with functional capabilities. Apart from various genome browsers with novel features that have been developed over the last decade, a plethora of new generation analysis tools were also enabled for functional studies, transcriptome analysis, gene coexpression analysis, protein-protein interaction networks, metabolic pathway analysis, orthology identification analysis and assessment of indexed rice mutant genes (Chandran and Jong 2014; Hong et al. 2019). All these tools have their own significance and greatly enhance the functional implication of rice genes.

## 4.1 Genome Browsers

The concurrent release of genome sequences for *Oryza sativa* enabled generation of huge volumes of functional genomics data, made easy by high-throughput analysis tools. Therefore an integrated genome browser to visualise, navigate, analyse and annotate the rice genome was becoming indispensable for researchers and biologists. To serve this purpose of data organisation and visualisation, browsers such as Rice-Map (<http://www.ricemap.org/>), Rice Genome Annotation Project (RGAP, <http://rice.plantbiology.msu.edu/>), Rice Annotation Project Database (RAP-DB, <https://rapdb.dna.affrc.go.jp/>), Rice Functional Genomic Express Database (RiceGE, <http://198.202.69.30/cgi-bin/RiceGE>) and Gramene (<http://gramene.org/>) were eventually developed, each built on datasets differing from each other (Wang et al. 2011, Kawahara et al. 2013, Sakai et al. 2013, Tello-Ruiz et al. 2018). Rice-Map delivers several precomputed *japonica* and *indica* subspecies annotations and also provides an interactive interface for users to browse the different genomic features at multiple levels (Wang et al. 2011). RGAP provides the fundamental rice genome data, supplemented with various analysis tools such as BLAST search, GO retrieval, domain and motif search, etc. (Kawahara et al. 2013). In contrast, Gramene allows for extension of data to other candidates within the plant kingdom, or subspecies within *Oryza* (Tello-Ruiz et al. 2018). These genome browsers allow bulk data retrieval, batch query searches and efficient identifier mapping between loci listed in discrete databases.

## 4.2 Orthology Identification Databases

These databases have significantly augmented genome-wide analysis of a group of genes and comparative genomics research. For transferring the knowledge of economically important QTLs from one crop species to another, it is crucial that the orthologues are correctly identified. This knowledge can be exploited for development of markers in agronomically important crops. Such analysis can also boost the expanse of the predicted protein-protein interaction networks: experimentally

proven interactions between a pair or group of orthologous protein interactors can be extrapolated to other species, thus facilitating functional studies. Phytozome (<https://phytozome.jgi.doe.gov/pz/portal.html>) provides a centralised comparative hub for plant genome including land plants and algae (Goodstein et al. 2012). Each plant gene's evolutionary history can be browsed starting right from its sequence, gene structure, gene family and organisation within the genome, thus providing a platform to compare genes across plant species. GreenPhylDB (<http://www.greenphyl.org/cgi-bin/index.cgi>) is another web resource containing a comprehensive catalogue of gene families, automatically clustered and manually annotated into orthologous groups (Conte et al. 2008). It is supplemented with a powerful tool, GreenPhyl Ortholog Search Tool (GOST), which can predict orthologous relationships between protein sequences across a broad taxonomy of green plants.

### 4.3 Transcriptome Databases

The ultimate goal in rice molecular biology is to correctly assign a functional role to all the predicted genes in the rice genome. Despite major developments in high-throughput analysis tools, the number of fully characterised genes is far more less than the number of predicted genes, still classified as 'hypothetical', 'unknown' or 'unannotated'. Large-scale gene expression profiling from different organ, tissue or cell types at various developmental stages greatly enhances the efforts put into functional characterisation of rice genes (Roychoudhury et al. 2011). The rapid accumulation of microarray data for rice in public repositories allowed hand-in-hand development of diverse informatics platforms and data analysis tools (Roychoudhury and Banerjee 2015). Rice Expression Profile Database (RiceXPro, <http://ricexpro.dna.affrc.go.jp/>) was built on the Agilent 44 K microarray platform, with probes derived from manually curated gene models in RAP-DB and full-length cDNA sequence information extracted from Knowledge-based Oryza Molecular biological Encyclopedia (KOME, [cdna01.dna.affrc.go.jp/cDNA/](http://cdna01.dna.affrc.go.jp/cDNA/)) database (Rice Full-Length cDNA Consortium 2003; Sato et al. 2011). The microarray analysis was done from organs/tissues of the rice plant at various growth stages under natural field conditions, rice seedlings treated with different phytohormones and specific cell types or tissues isolated by laser microdissection (LMD) techniques. The Rice Oligonucleotide Array Database (ROAD, <http://www.ricearray.org/>) integrates information from six rice microarray platforms, including the Affymetrix, Agilent 22 K and 44 K, BGI/Yale and the NSF 20 K and 45 K, thus making it very comprehensive, and provides a user-friendly web interface with various functional analysis tools (Cao et al. 2012). ROAD allowed meta-profile analysis for different anatomic tissues at various developmental stages, gene coexpression analysis and creation of coexpression networks and gene ontology (GO) and KEGG orthology (KO) analyses of query genes. With the emergence of next-generation sequencing technologies (NGS), RNA sequence-based profiling is rapidly gaining ground and is already replacing the databases built on microarray data, thus overcoming the limitations of microarray technique. Rice Expression Database (RED,



[expression.ic4r.org/](http://expression.ic4r.org/)) is a repository of gene expression profiles derived from curated and high-quality RNA-Seq data from tissues at various developmental stages or encompassing a wide variety of biotic and abiotic treatments (Xia et al. 2017). Similarly, AgriSeqDB (<https://expression.latrobe.edu.au/agriseqdb>) is an online RNA-Seq database with features for visualisation, analysis and interpretation of transcriptome data from various stages of development and tissue/cell types from several species, for major agricultural crops such as rice, wheat, maize, barley and tomato (Robinson et al. 2018). Transcriptome ENyclopedia Of Rice (TENOR, <http://tenor.dna.affrc.go.jp>) provides comprehensive large-scale mRNA-Seq data obtained from rice sampled from a variety of conditions: ten abiotic stress conditions such as high salt stress, osmotic stress, high and low phosphate or cadmium levels, drought, cold and flood and two plant hormone treatments (abscisic acid and jasmonic acid) (Kawahara et al. 2016).

#### 4.4 Promoter Databases

Promoter databases provide information regarding the core promoter structures and regulatory elements, which have been experimentally verified or predicted from consensus. Various informative resources for promoter detection and analysis are available for different plant species like rice, *Arabidopsis*, poplar, etc. Plant Promoter Database (PPDB, <http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi>) is a web-based promoter database which comprises of sequence lists of bioinformatically identified promoter elements, extracted by local distribution of short sequence analysis (LDSS) (Kusunoki and Yamamoto 2017). Promoter recognition is accomplished by annotating genome sequence to these lists of TATA boxes, initiators, GA and CA elements, Y patches and regulatory element groups (REGs), supplemented with information on experimentally demonstrated transcription start sites (TSSs). Additionally, REGs are linked to the information in the Plant cis-acting regulatory DNA elements (PLACE, <http://www.dna.affrc.go.jp/htdocs/PLACE/>) database (Higo et al. 1998). PLACE is a database of nucleotide sequence motifs of regulatory elements extracted from published reports on genes in vascular plants and other plant species. It is appended with a Signal Scan programme that allows users to search for cis-regulatory elements in their query sequence. The outputs are assigned PLACE accession numbers and are hyperlinked to PubMed or GenBank identifiers. The Plant Promoter Analysis Navigator (PlantPAN, <http://plantpan2.itps.ncku.edu.tw/>) tool allows users to search for transcription factor binding sites (TFBSs), respective transcription factors (TFs) and several important cis- and trans-regulatory elements in their query promoter sequences or set of promoter sequences in plants (Chang et al. 2008). It also allows determining co-occurrence of TFs and their binding sites for the promoters of the input gene groups and construction of gene-regulatory networks using coexpression analysis. PlantPAN also provides cross species promoter search utilising paralogs and orthologues.

## 4.5 Databases for Rice Coexpression Analysis

Gene coexpression analysis is performed to associate genes of known or unknown function with biological pathways, by discerning the correlation patterns among them across microarray and other transcriptomic datasets. But with recent advances in next-generation sequencing, gene coexpression networks built on RNA-Seq data enable improving the definitions of gene function and associating noncoding genes and splice variants with some biological roles. RiceFRIEND (<http://ricefriend.dna.affrc.go.jp/>) is a gene coexpression database built on large volumes of rice microarray data, derived from various tissues at different developmental stages under natural field conditions or under the influence of some phytohormone treatments – the same Agilent 44 K array dataset available in the RiceXPro database (Sato et al. 2011, 2013). RiceFRIEND provides a platform with two modes of search: single guide gene and multiple guide genes mediated search for coexpressed genes or functionally related genes in various signalling or metabolic pathways. As already mentioned previously, PlantPAN also has options for gene coexpression analysis and for construction of regulatory networks based on co-occurrence TFs and protein-protein interactions (Chang et al. 2008). OryzaExpress is another integrated database, providing interactive user interface to compare gene expression networks in rice with that of other plants (Arabidopsis Gene Expression Network data from ATTED-II) (Hamada et al. 2011).

## 4.6 Phylogenomics Databases

Experimental determination of biological functions of genes within large families becomes problematic because redundancy limits exactly assigning the contribution of individual genes. As of 2019, 1072 genes have been reported in the collection of the Overview of functionally characterized Genes in Rice Online database (OGRO, <http://qtaro.abr.affrc.go.jp/ogro/>) (Yamamoto et al. 2012). Therefore, a combinatorial approach of incorporating phylogenetics into comparative genomics had been undertaken to predict the functions of candidates within large-gene families considering the similarity among gene products: a phylogenomics analysis. In rice, phylogenomics databases have been constructed for six gene families (<http://ricephylogenomics.ucdavis.edu/index.shtml>): kinases, glycosyltransferases (GTs), glycoside hydrolases (GHs), transcription factors (TFs), transporters and cytochrome p450 monooxygenases (P450s). The Rice Kinase Database (RKD) derives information from the NCBI Gene Expression Omnibus (GEO) and provides a platform to integrate functional genomics data into phylogenetic trees built for individual families (Dardick et al. 2007; Jung et al. 2010). The RKD includes an interactive chromosomal map depicting the position of kinase genes, protein-protein interaction maps and meta-expression data developed on microarray data analysis. Thus RKD facilitates effective estimation of functional redundancy or dominance among closely related kinases within subfamilies. The Rice GT database integrates functional genomics information of all putative rice glycosyltransferases on a

phylogenetic tree (Cao et al. 2008). All these putative GTs have been identified through similarity searches against annotated GTs in the Carbohydrate-Active enZymes (CAZy) database (<http://www.cazy.org/>) and subsequently classified into subclasses, based on domain composition and sequence similarities (Lombard et al. 2014). Similarly, the rice GH database lists all the GHs identified through sequence similarity searches in the CAZy database and shares a lot of features with the GT database (Sharma et al. 2013). It incorporates several features such as gene expression patterns, orthologous relationships, structural information and mutant availability for individual GH families in a phylogenomics context. These efforts have provided some significant insight into understanding of cell wall structure and biotic and abiotic stress tolerance. The Rice TF database (<http://ricephylogenomics.ucdavis.edu/tf/>) integrates information on putative and predicted rice TFs and transcriptional regulators, retrieved from the Plant Transcription Factor Database (PlnTFDB, <http://plntfdb.bio.uni-potsdam.de/v3.0/>) (Jin et al. 2017). Another database that provides comprehensive information regarding rice TF expression during drought and salinity stress and at various developmental stages is the Rice Stress-Responsive Transcription Factor Database (RiceSRTFDB, <http://www.nipgr.res.in/RiceSRTFDB.html>) (Priya and Jain 2013). This enables a holistic view of the TF of interest involved in stress response at a particular developmental stage, the cis-regulatory elements in the promoters, mutant availability and phenotype information. The Rice Transporter database (<http://ricephylogenomics.ucdavis.edu/transporter/>) contains all the potential rice membrane transporter genes, retrieved from the Transporter Protein Analysis Database (TransportDB; <http://www.membranetransport.org/>), while the Rice CytP450 Database (<http://ricephylogenomics.ucdavis.edu/p450/>) was developed to integrate functional genomics information for all putative rice P450s in a phylogenomics context (Elbourne et al. 2017). Both these databases are supplemented with similar features like orthologous relationships, gene expression patterns, phenotype information, etc., as discussed above for other databases.

## 4.7 Proteome Databases

Proteome databases in rice have been built on diverse platforms or data acquisition types, thereby each having their own significance. Rice Proteome Database (RPD, [http://gene64.dna.affrc.go.jp/RPD/main\\_en.html](http://gene64.dna.affrc.go.jp/RPD/main_en.html)) contains a catalogue of rice proteins, extracted from various tissues and subcellular compartments and separated and analysed on two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) (Komatsu 2005). The database contains an ensemble of reference maps based on 2D-PAGE, with sequences of individual proteins, as well as functional characterisation of major proteins. Plant Proteomics Database (PPDB, <http://ppdb.tc.cornell.edu/>) contains an assembly of all protein-encoding gene models in *Arabidopsis*, maize and rice, which are linked to each other via BLAST options (Sun et al. 2009). Each gene model is associated with some experimental information which has been derived from in-house proteome and mass spectrometry analysis

and is also supplemented with curated information regarding protein function, properties and subcellular localisation. Nowadays, with the advances in mass spectrometry, large-scale and quality-controlled peptide sequencing data are available. OryzaProteogenomics Database (OryzaPG-DB, <http://oryzapg.iab.keio.ac.jp/>) is a data repository of shotgun proteomics data, built from the outputs of 27 nanoLC-MS/MS runs, on a hybrid ion trap-orbitrap mass spectrometer, of tryptic digests from undifferentiated cultured rice cells (Helmy et al. 2011). Peptides were identified, when the product ion spectra were searched against protein, cDNA, transcript and genome databases and were subsequently mapped onto the rice genome. OryzaPG is the first proteogenomics-based database of the rice proteome, which associated peptide-based expression profiles with corresponding genomic origin.

#### 4.8 Protein-Protein Interaction Databases

These databases have flourished due to the recent developments in techniques employed to identify the potential interactors of a protein candidate. Understanding the protein interactomes from yeast two-hybrid (Y2H), split-ubiquitin (sUbc), bimolecular fluorescence complementation (BiFC), fluorescence resonance energy transfer (FRET), affinity purification (AP), co-immunoprecipitation (Co-IP) and mass spectrometry (MS) has reignited the possibilities of comparative analysis of protein-protein interactions (PPIs). Thus based on the assumption that protein complexes across different species share significant structural and functional similarity, evolutionary conserved proteins will not only retain their structure and function but also the interactions with other protein partners. Predicted Rice Interactome Network (PRIN, <http://bis.zju.edu.cn/prin/>), a well-annotated PPI database for rice, extended the current expanse of the PPI data by integrating information from experimentally verified interologs of six model organisms (*Saccharomyces cerevisiae* (yeast), *Caenorhabditis elegans* (worm), *Drosophila melanogaster* (fruit fly), *Homo sapiens* (human), *Escherichia coli* K12 and *Arabidopsis thaliana*), using computational approaches (Gu et al. 2011). Supplemented with GO assignments, subcellular localisation data and gene expression data, PRIN provides a user-friendly web interface for easy database search and effective network visualisation. STRING (Search Tool for the Retrieval of Interacting Genes/Proteins, <https://string-db.org/>) is another database of established and predicted protein-protein interactions, which are of direct (physical) or indirect (functional) nature (Szklarczyk et al. 2015). Information regarding interactions in STRING database stems from genomic context predictions, high-throughput lab experiments, coexpression data, automated text-mining, identifying interologs in other organisms and interactions reported in other primary databases.

## 4.9 Databases for Metabolome Analysis

Metabolomics addresses the global metabolic changes in biological systems and has immense application in disease diagnosis and functional genomics. Metabolic profiling of cells, tissues or organisms, biomarker identification and pathway analysis are an integral part of it and demand easy-to-use analysis tools. MetaboAnalyst (<https://www.metaboanalyst.ca/>) provides a web-based user interface with metabolomics data processing tools, options for data normalisation, statistical analysis, graphing, metabolite identification and characterisation as well as pathway mapping (Chong et al. 2018). MetPA (Metabolomics Pathway Analysis, <https://www.metaboanalyst.ca/>) is a web-based tool, which combines advanced pathway enrichment analysis and pathway topology analysis, to visualise and analyse metabolic pathways involved in a particular metabolomics study (Xia and Wishart 2010). MetPA currently allows visualisation and analysis of a total of 1600 pathways for 21 model organisms, including rice. EXPath (<http://expath.itps.ncku.edu.tw>) uses public large-scale microarray datasets, derived from samples under abiotic or biotic stress or under different hormone treatments, for gene coexpression analysis and identification of differentially expressed genes (DEGs) (Chien et al. 2015). Down the pipeline, this information is finally utilised for inferring enriched KEGG pathways and GO terms for three plant species, namely, *Arabidopsis*, rice and maize.

## 4.10 Rice Gene Indexed Mutant Databases

The rice research community has greatly benefited from the high quality of the rice genome and also the appreciably fair annotation in other cereal crop species. To elucidate the function of all the predicted coding and noncoding regions within the rice genome, the International Rice Research Institute (IRRI) in close association with the International Rice Functional Genomics Consortium (IRFGC) developed an enormous collection of indexed rice mutant genes. The members within a population of gene indexed mutants are distinguished on the following criteria: mutagenesis methods adopted (T-DNA insertion, transposons, chemical or physical mutagens), rice variety, mutant phenotypes, seed availability, reporter-gene expression patterns and mutated loci per genome. All this information facilitated further functional characterisation of the genome. Rice Mutant Database (RMD, <http://rmd.ncpgr.cn/>) encompasses the information regarding ~129,000 rice T-DNA insertion (enhancer trap) lines generated by an enhancer trap system and delivers three distinct functional classes: (1) novel gene identification, (2) identification of regulatory elements and (3) identification of patterns of ectopic expression of target gene at particular growth stage or tissue type (Zhang et al. 2006). Oryza Tag Line (<http://oryzatagline.cirad.fr/>) is a phenotypic mutant database for the French genomics initiative ‘Genoplante’ rice insertion line library (Larmande et al. 2008). Based on the molecular characterisation of the mutagen insertion sites, the sequence information of flanking sequence tags (FSTs) was retrieved. This database also allows forward genetic search through queries based on mutant phenotype or reporter-gene expression, coupled with other

categories such as organ, developmental stage or trait. OryGenesDB (<http://orygenesdb.cirad.fr/data.html>) displays sequence information on T-DNA and *Ds* FSTs in insertional rice mutants, produced in the framework of Genoplante and the EU consortium Cereal Gene Tags (Droc et al. 2006). This database is supplemented with tools for reverse genetics, allowing a molecular geneticist to readily find T-DNA, *Tos17* and/or *Ds* insertion lines in genes of interest and to retrieve all the annotations associated with these sequences, derived from external rice molecular resources (e.g. cDNA full length, gene, EST, markers, expression data, etc.).

#### 4.11 Other Databases

Apart from the above listed categories of databases, there are other web resources available, each built on diverse data sources and possessing unique features and tools compatible for functional genomic annotations. RiceVarMap (<http://ricevarmap.ncpgr.cn/v2/>) provides curated information of genomic variations—single nucleotide polymorphisms (SNPs) and small insertions/deletions (INDELS)—from sequencing data of ~4700 rice accessions (Zhao et al. 2015). The variations are associated with comprehensive resources of functional annotations, chromatin accessibility data, risks associated with gene expression, phenotype data and images, agronomic and metabolic traits, geographical details, etc. DroughtDB (<http://pgsb.helmholtz-muenchen.de/droughtdb>) is a valuable tool for researchers working on drought stress (Alter et al. 2015). This database is a manually and expert-curated compilation of drought stress-responsive genes in plants, which have been molecularly characterised. Each drought stress-responsive gene, already characterised either in *Arabidopsis* or *Oryza sativa*, is supplemented with information on its sequence, physiological or molecular function, mutant phenotypes and their homologs in nine model crop species, like maize, barley, etc. With new insights into the role of microRNAs in plant growth and development, repositories of microRNA data are also being developed. The plant microRNA database (PMRD, <http://bioinformatics.cau.edu.cn/PMRD/>) congregates available information on plant microRNA data, for example, sequence of microRNAs and their target genes, secondary dimension structure, expression profiling, genomic origins, etc. (Zhang et al. 2010). Gene duplications are widespread in plant species. For example, receptor-like kinase families in *Arabidopsis* and rice have nearly 600 and 1000 members, respectively. This implies gene duplication and subsequent mutations have generated new genes with diverse functions. In similar lines, recent developments in gene coexpression networks have implied that not only genes but pathways can also undergo multiplication and diversification to perform related functions in various parts of an organism. Such pathways are called modules, which may give us a lead to understand the biological processes and functions well. FamNet (<http://aranet.mpimp-golm.mpg.de/famnet.html>) is an interactive platform for exploration and visualisation of these multiplied modules in gene coexpression networks of eight plant species (Ruprecht et al. 2016). With this tool, multiplied modules involved in tip growth in pollen tubes

and root hairs or in secondary metabolite synthesis were identified and functionally characterised in separate plant species.

Recent advances in development and refinement of new and existing high-throughput analysis tools have significantly impacted the functional annotation of the rice genome. From browsing the genome to deciphering phylogenomic relationships, from mapping orthologues to identifying SNPs and INDELS, from identification of gene coexpression networks to metabolic pathways, these tools have leveraged the quality of meaningful predictions. Table 4 lists the diverse omics tools—browsers, databases, mapping tools, etc.—that have not been touched upon in the above section, but are used at large for functional genomics.

---

## 5 Rice Genomics Propelled Research on Important Agronomic Traits

The availability of the rice genome sequence data and development of new tool sets to analyse the data has greatly revolutionised the research in rice genetics and breeding. A search in the NCBI PubMed database with the keywords ‘map-based cloning’, ‘rice’ and ‘*Oryza*’ has retrieved nearly over 4700 publications corresponding to rice research, from the past three decades (see Fig. 3). Following the establishment of the Rice Genome Research Program in Japan, there was a major boost in rice research, with 40–60 publications per year. This was nearly 2–3 times the publication volume before RGP was established. But following the public release of the rice genome sequence by IRGSP in 2005, which incorporated the draft sequences from Monsanto and Syngenta, there were, on an average, more than 200 publications per year pertaining to rice research. It is evident from the volume of publications in the last 10 years that the numbers are on the rise, which has been propelled with the genomes of several wild and domesticated rice varieties and other cereals being sequenced. This rough analysis, which does not encompass any manual curation of the retrieved citations, simply indicates the enormous advancement in the rice research area with the completed rice genome sequence, with a multitude of genes being associated with metabolic, developmental or stress-related signalling pathways (Wing et al. 2018; International Rice Genome Sequencing Project and Sasaki 2005). Moreover, the reference genome of rice allowed exploration into the genomes across the entire *Oryza* genus as well as studying of genetic variations among domesticated rice species and their wild relatives (Wing et al. 2018; Civián et al. 2015; Wang et al. 2014; Meyer et al. 2016). Sequence-based analysis of the variations in different species allowed breeders to exploit such variations for rice improvement. But none of this was possible without the significant increase in the number of molecular markers, and access to the knowledge about their physical order in chromosomes and proximity to annotated genes, all of which was utilised to predict gene-trait associations. Another noticeable impact of the rice genome sequence was an opportunity to gain an understanding of the molecular and genetic basis of the traits such as efficient mineral utilisation, resistance towards biotic and abiotic stress, physical features and so on. All this insight helped rice

**Table 4** List of genome browsers, databases and diverse omics tools which facilitate high-throughput analysis of the rice genome

| Databases  | Resource links  | References                    |
|--|---|-------------------------------|
| <i>Genome browsers</i>                           |   |                               |
| TIGR Rice Genome Annotation project              | <a href="http://blast.jcvi.org/euk-blast/index.cgi?project=osa1">http://blast.jcvi.org/euk-blast/index.cgi?project=osa1</a> | Yuan et al. (2003)            |
| OsGDB  | <a href="http://www.plantgdb.org/OsGDB/">http://www.plantgdb.org/OsGDB/</a>   | Dong et al. (2005)            |
| Rice FPC Genome Browser                          | <a href="https://www.genome.arizona.edu/fpc/rice/gbrowse/">https://www.genome.arizona.edu/fpc/rice/gbrowse/</a>             | Pampanwar et al. (2005)       |
| Oryzabase  | <a href="https://shigen.nig.ac.jp/rice/oryzabase/">https://shigen.nig.ac.jp/rice/oryzabase/</a>                             | Kurata and Yamazaki (2006)    |
| Rice-Map   | <a href="http://www.ricemap.org/">http://www.ricemap.org/</a>   | Wang et al. (2011)            |
| MSU Rice Genome Annotation Project (RGAP)        | <a href="http://rice.plantbiology.msu.edu/">http://rice.plantbiology.msu.edu/</a>   | Kawahara et al. (2013)        |
| Rice Annotation Project Database (RAP-DB)        | <a href="https://rapdb.dna.affrc.go.jp/">https://rapdb.dna.affrc.go.jp/</a>   | Sakai et al. (2013)           |
| Gramene  | <a href="http://www.gramene.org/">http://www.gramene.org/</a>   | Tello-Ruiz et al. (2018)      |
| <i>Resources for rice orthology analysis</i>     |   |                               |
| GreenPhylDB                                      | <a href="http://www.greenphy1.org/cgi-bin/index.cgi">http://www.greenphy1.org/cgi-bin/index.cgi</a>                         | Conte et al. (2008)           |
| Phytozome  | <a href="https://phytozome.jgi.doe.gov/pz/portal.html">https://phytozome.jgi.doe.gov/pz/portal.html</a>                     | Goodstein et al. (2012)       |
| Putative Orthologous Groups (POGs) DB            | <a href="http://cas-pogs.uoregon.edu/#/">http://cas-pogs.uoregon.edu/#/</a>   | Tomcal et al. (2013)          |
| InParanoid                                       | <a href="http://inparanoid.sbc.su.se/cgi-bin/index.cgi">http://inparanoid.sbc.su.se/cgi-bin/index.cgi</a>                   | Sonnhammer and Östlund (2015) |
| Plaza  | <a href="https://bioinformatics.psb.ugent.be/plaza/">https://bioinformatics.psb.ugent.be/plaza/</a>                         | Van Bel et al. (2018)         |
| <i>Resources for rice transcriptome analysis</i> |   |                               |
| RiceXPro   | <a href="http://ricexpro.dna.affrc.go.jp/">http://ricexpro.dna.affrc.go.jp/</a>   | Sato et al. (2011)            |
| Rice Oligonucleotide Array Database (ROAD)       | <a href="http://www.ricearray.org/">http://www.ricearray.org/</a>   | Cao et al. (2012)             |
| PLEXdb   | <a href="http://www.plexdb.org/plex.php?database=Rice">http://www.plexdb.org/plex.php?database=Rice</a>                     | Dash et al. (2012)            |
| Transcriptome Encyclopaedia Of Rice (TENOR)      | <a href="https://tenor.dna.affrc.go.jp/">https://tenor.dna.affrc.go.jp/</a>   | Kawahara et al. (2016)        |
| Rice Expression Database (RED)                   | <a href="http://expression.ic4r.org/">http://expression.ic4r.org/</a>   | Xia et al. (2017)             |
| AgriSeqDB  | <a href="https://expression.latrobe.edu.au/agriseqdb">https://expression.latrobe.edu.au/agriseqdb</a>                       | Robinson et al. (2018)        |
| <i>Resources for rice promoter analysis</i>      |   |                               |
| Plant Promoter Analysis Navigator (PlantPAN)     | <a href="http://plantpan2.itps.ncku.edu.tw/">http://plantpan2.itps.ncku.edu.tw/</a>   | Chang et al. (2008)           |
| Osiris   | <a href="http://www.bioinformatics2.wsu.edu/Osiris">http://www.bioinformatics2.wsu.edu/Osiris</a>                           | Morris et al. (2008)          |
| GrassPROMDB (Grassius)                           | <a href="https://grassius.org/grasspromdb.php">https://grassius.org/grasspromdb.php</a>                                     | Yilmaz et al. (2009)          |

(continued)



**Table 4** (continued)

| Databases   | Resource links  | References                                   |
|---|---|--|
| Plant Promoter Database (PPDB)                                    | <a href="http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi">http://ppdb.agr.gifu-u.ac.jp/ppdb/cgi-bin/index.cgi</a>                         | Kusunoki and Yamamoto (2017)                 |
| <i>Resources for rice coexpression analysis</i>                   |   |  |
| OryzaExpress  | <a href="http://plantomics.mind.meiji.ac.jp/OryzaExpress/">http://plantomics.mind.meiji.ac.jp/OryzaExpress/</a>                               | Hamada et al. (2011)                         |
| RiceFREND   | <a href="http://ricefrend.dna.affrc.go.jp/">http://ricefrend.dna.affrc.go.jp/</a>   | Sato et al. (2013)                           |
| PLANt co-EXpression database (PLANEX)                             | <a href="http://planex.plantbioinformatics.org/">http://planex.plantbioinformatics.org/</a>   | Yim et al. (2013)                            |
| RECoN: Rice Environment Coexpression Network                      | <a href="https://plantstress-pereira.uark.edu/RECoN/">https://plantstress-pereira.uark.edu/RECoN/</a>   | Krishnan et al. (2017)                       |
| <i>Resources for rice phylogenomics</i>                           |   |  |
| Rice Kinase Database (RKD)  | <a href="http://ricephylogenomics.ucdavis.edu/kinase/">http://ricephylogenomics.ucdavis.edu/kinase/</a>                                       | Dardick et al. (2007) and Jung et al. (2010) |
| Rice GlycosylTransferase Database (RGTD)                          | <a href="http://ricephylogenomics.ucdavis.edu/index.shtml">http://ricephylogenomics.ucdavis.edu/index.shtml</a>                               | Cao et al. (2008)                            |
| Stress-responsive Transcription Factor Database (STIFDB)          | <a href="http://caps.ncbs.res.in/stifdb2/">http://caps.ncbs.res.in/stifdb2/</a>   | Shameer et al. (2009)                        |
| Rice glycoside hydrolase database                                 | <a href="http://ricephylogenomics.ucdavis.edu/cellwalls/gh/genInfo.shtml">http://ricephylogenomics.ucdavis.edu/cellwalls/gh/genInfo.shtml</a> | Sharma et al. (2013)                         |
| Rice Stress-Responsive Transcription Factor Database (RiceSRTFDB) | <a href="http://www.nipgr.res.in/RiceSRTFDB.html">http://www.nipgr.res.in/RiceSRTFDB.html</a>   | Priya and Jain (2013)                        |
| Rice Transporter Database   | <a href="http://ricephylogenomics.ucdavis.edu/transporter/">http://ricephylogenomics.ucdavis.edu/transporter/</a>                             | Jung et al. (2015)                           |
| Rice CytP450 Database   | <a href="http://ricephylogenomics.ucdavis.edu/p450/">http://ricephylogenomics.ucdavis.edu/p450/</a>   | Jung et al. (2015)                           |
| Rice TF Database  | <a href="http://ricephylogenomics.ucdavis.edu/tf/">http://ricephylogenomics.ucdavis.edu/tf/</a>   | Jung et al. (2015)                           |
| <i>Resources for rice proteogenomics</i>                          |   |  |
| Rice Proteome Database (RPD)                                      | <a href="http://gene64.dna.affrc.go.jp/RPD/main_en.html">http://gene64.dna.affrc.go.jp/RPD/main_en.html</a>                                   | Komatsu (2005)                               |
| Plant Proteomics Database (PPDB)                                  | <a href="http://ppdb.tc.cornell.edu/">http://ppdb.tc.cornell.edu/</a>   | Sun et al. (2009)                            |
| OryzaPG-DB  | <a href="http://oryzapg.iab.keio.ac.jp/">http://oryzapg.iab.keio.ac.jp/</a>   | Helmy et al. (2011)                          |
| <i>Resources for rice protein-protein interaction analysis</i>    |   |  |
| Predicted Rice Interactome Network (PRIN)                         | <a href="http://bis.zju.edu.cn/prin/">http://bis.zju.edu.cn/prin/</a>   | Gu et al. (2011)                             |
| Database of interacting proteins in <i>Oryza sativa</i> (DIPOS)   | <a href="http://comp-sysbio.org/dipos/?id=5">http://comp-sysbio.org/dipos/?id=5</a>   | Sapkota et al. (2011)                        |
| STRING  | <a href="https://string-db.org/">https://string-db.org/</a>   | Szklarczyk et al. (2015)                     |
| <i>Resources for rice metabolic pathway analysis</i>              |   |  |
| KEGG  | <a href="https://www.genome.jp/kegg/pathway.html">https://www.genome.jp/kegg/pathway.html</a>   | Kanehisa and Goto (2000)                     |
| MapMan  | <a href="http://mapman.gabipd.org/home">http://mapman.gabipd.org/home</a>   | Thimm et al. (2004)                          |

(continued)

**Table 4** (continued)

| Databases  | Resource links  | References                                     |
|--|---|--|
| MetaboAnalyst/MetPA  | <a href="https://www.metaboanalyst.ca/">https://www.metaboanalyst.ca/</a>   | Xia and Wishart (2010) and Chong et al. (2018) |
| RiceNetDB  | <a href="http://bis.zju.edu.cn/ricenetdb/">http://bis.zju.edu.cn/ricenetdb/</a>                                     | Liu et al. (2013)                              |
| RiceCyc  | <a href="http://archive.gramene.org/pathway/ricecyc.html">http://archive.gramene.org/pathway/ricecyc.html</a>       | Dharmawardhana et al. (2013)                   |
| EXPath   | <a href="http://expath.itps.ncku.edu.tw">http://expath.itps.ncku.edu.tw</a>   | Chien et al. (2015)                            |
| PANTHER  | <a href="http://www.pantherdb.org/pathway/">http://www.pantherdb.org/pathway/</a>                                   | Mi et al. (2017)                               |
| Plant Reactome   | <a href="http://plantreactome.gramene.org/index.php?lang=en">http://plantreactome.gramene.org/index.php?lang=en</a> | Naithani et al. (2017)                         |
| <i>Resources for rice gene indexed mutants</i>                             |   |  |
| Rice Mutant Database (RMD)   | <a href="http://rmd.ncpgr.cn">http://rmd.ncpgr.cn</a>   | Zhang et al. (2006)                            |
| OryGenesDB   | <a href="http://orygenesdb.cirad.fr/data.html">http://orygenesdb.cirad.fr/data.html</a>                             | Droc et al. (2006)                             |
| Taiwan Rice Insertional Mutants (TRIM)                                     | <a href="http://trim.sinica.edu.tw/">http://trim.sinica.edu.tw/</a>   | Chern et al. (2007)                            |
| Oryza Tag Line   | <a href="http://oryzatagline.cirad.fr/">http://oryzatagline.cirad.fr/</a>   | Larmande et al. (2008)                         |
| <i>Other database resources</i>  |   |  |
| Plant MicroRNA Database (PMRD)   | <a href="http://bioinformatics.cau.edu.cn/PMRD/">http://bioinformatics.cau.edu.cn/PMRD/</a>                         | Zhang et al. (2010)                            |
| NIASGBdb (National Institute of Agrobiological Sciences plantfdb database) | <a href="http://www.gene.affrc.go.jp/databases_en.php">http://www.gene.affrc.go.jp/databases_en.php</a>             | Takeya et al. (2011)                           |
| RiceVarMap   | <a href="http://ricevarmap.ncpgr.cn/v2/">http://ricevarmap.ncpgr.cn/v2/</a>   | Zhao et al. (2015)                             |
| DroughtDB  | <a href="http://pgsb.helmholtz-muenchen.de/droughtdb">http://pgsb.helmholtz-muenchen.de/droughtdb</a>               | Alter et al. (2015)                            |
| FamNet   | <a href="http://www.gene2function.de/famnet.html">http://www.gene2function.de/famnet.html</a>                       | Ruprecht et al. (2016)                         |

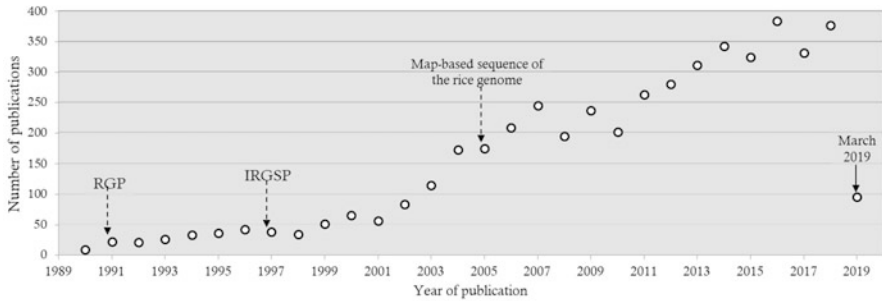
Resource links and references of corresponding publications are also provided

researchers to engineer ‘Green Super Rice’, a new generation of sustainable crops that will tackle the food demands of the growing population.

Limitations in arable land and water resources in rice-producing countries and the issues related to the increasing trend of global population growth can be addressed to a great extent by increasing per unit area yield of rice. This has been largely possible in the last decade because of correctly identifying the molecular markers determining the agronomically desirable traits and using them effectively in rice production.

## 5.1 Improving Plant Architecture and Yield

The semidwarf trait has been introduced into a number of rice cultivars growing worldwide, as it improves light-interception properties, harvest index, nitrogen



**Fig. 3** A timeline for the period 1990–2019 depicts nearly 4783 publications related to research on gene cloning in rice, retrieved from the NCBI PubMed database. For the year 2019, the records listed are till the month of March. The dotted arrows indicate the years when RGP and IRGSP were established and the map-based genome sequence of rice was released

responsiveness and lodging resistance, without having non-detrimental effects on the grain yield. The International Rice Research Institute developed a semidwarf variety of rice IR8, from a cross between Peta, a tall *indica* variety, and Dee-geo-woo-gen (DGWG), a Taiwanese indigenous semidwarf rice variety. This new variety led to a dramatic improvement in rice yields, thereby bringing about the Green Revolution to tropical Asia (Hargrove and Cabanilla 1979). IR8 had a culm length of 90–100 cm, compared to its parent line, Peta, characterised by longer culm (150–180 cm) and long hanging leaves. Sequence data from the rice genome and supporting mapping studies facilitated identification of genes responsible for the semidwarfism trait in rice (Monna et al. 2002; Spielmeier et al. 2002). The incomplete recessive gene, *sd1*, located on the long arm of chromosome 1, encodes for GA20-oxidase with loss-of-function mutations (OsGA20ox2). This gene product regulates synthesis of gibberellins, which regulates the GA biosynthesis pathway. Semidwarfism is one of the traits that got positively selected during domestication of *japonica* rice. In addition to affecting plant height, it has positive effect on the tiller number and enhances erection of the blade.

Apart from plant architecture, other component traits that determine yield of a rice plant are tiller number, number of grains per panicle, grain weight and grain filling rate. The *MONOCULM 1* (*MOC1*) gene is the first gene identified related to controlling rice tiller number (Li et al. 2003). *MOC1* is a member of the plant-specific GRAS transcription factor family. *MOC1* is required for growth of axillary meristems at both vegetative and reproductive stages. Consequently, loss of function of *MOC1* affects formation of both tiller and panicle. *Ghd7* encodes a CCT domain containing transcriptional activator that simultaneously controls number of grains per panicle, plant height and heading date (Xue et al. 2008). Another QTL with similar effects was *Ghd8*, which encodes the OsHAP3 subunit of a heterotrimeric heme activator protein (HAP) (Yan et al. 2011). This gene manipulated flowering time in rice and upregulated *MOC1*, thereby increasing the number of tillers and grain yield. Along with *Ghd7.1*, which encodes a PSEUDO-RESPONSE REGULATOR (PRR), *Ghd7* and *Ghd8* render pleiotropic effects

(delays heading, increases plant height and grain yield) under long-day conditions. Map-based cloning facilitated identification of *Hd1* gene, which largely affected the interaction between *Ghd7* and *Ghd8* (Zhang et al. 2015). Recently it has been demonstrated that combinatorial loss-of-function alleles of *Ghd7*, *Ghd8* and *Hd1* allow expansion of rice cultivars to higher altitudes, therefore defining the ecogeographical adaptation and yield potential in rice cultivars. High-yielding varieties can also be associated with high grain numbers per ear. Several genes, such as the *DENSE AND ERECT PANICLE* genes, *DEP1*, *DEP2* and *DEP3*, have been identified that affect this trait (Xu et al. 2016). *Gn1* gene encoding a cytokinin oxidase *OsCKX2* has been identified by map-based cloning. Downregulation of *Gn1* expression leads to an increment in the quantity of glumous flowers, thereby grain number per ear (Yeh et al. 2015). *IPAI* (*Ideal Plant Architecture 1*) encodes the transcription factor *OsSPL14*, which regulates the plant architecture through *DEP1* (Zhang et al. 2017). Fewer but more productive tillers are produced upon achieving optimal expression levels of *IPAI*, thereby regulating plant height and ear length.

## 5.2 Improving Grain Quality

This demand of enhanced grain quality is primarily determined by consumers and is thereby important to farmers and millers. The quality of rice grain depends on a number of features, which vary according to the preferences of consumers across the world. The grain size, shape and its translucent appearance are important features determining grain quality and are directly related to yield. A complement of genes has been indicated to control these traits. *GS3*, *GW2* and *GW5* are negatively correlated with grain length and width, grain weight and grain fullness (Zheng et al. 2015). Upregulation of *GS5* expression is involved in seed yield, size, thousand seed weight and seed setting rate as well as grain width and weight. *GW8* expression promotes cell division and regulates grain size, thousand seed weight and filling speed, thereby increasing yield and affecting milling quality. Milling quality is largely determined by the chalkiness and intact nature of the milled rice. Grain chalkiness is highly undesirable and a major QTL, *Chalk5*, influences this trait. Elevated expression of *Chalk5* increases chalkiness of the endosperm, thereby posing a major problem in milling and post processing (Li et al. 2014). A perfect combination of amylose content and gelling temperature influences the cooking quality and palatability of rice. The *WAXY* gene codes for the starch synthase enzyme, which is responsible for synthesis of amylose in the endosperm. Based on the amylose content (AC), rice is commercially classified into five categories: high (25–33%), intermediate (20–25%), low (12–20%), very low (5–12%) and waxy (0–5%) (Juliano 1992; Yamanaka et al. 2004). Grains with high AC result in dry and well-separated rice after cooking but eventually turn hard upon cooling. Cooking of rice varieties with intermediate AC leads to a soft, nonsticky texture in rice, while low and very low AC results in a soft and sticky texture. Waxy rice (also called, sweet rice or glutinous rice) becomes sticky when cooked. Using map-based cloning strategy, *ALK* was identified to be the key gene controlling gelatinisation

temperature and therefore the gel consistency in rice (Gao et al. 2003). Although rice is a major carbohydrate and even a potential protein source for rice-eating populations, it is a poor source of essential micronutrients. For countries where rice is the staple food, the primary cause of micronutrient malnutrition, including iron, zinc and vitamin A deficiencies, is a rice-based diet. All these deficiencies can be related with reduced working capacity, decreased mental capacity, blindness, stunting and elevated morbidity and mortality. To address this issue, food-based approaches have been adopted that involve biofortification of genotypes resulting in increased levels of vitamins and minerals. Golden Rice, expressing two genes encoding phytoene synthase and carotene desaturase, produced the provitamin A carotenoid  $\beta$ -carotene in rice endosperm and is one of the earliest examples of biofortification (Schaub et al. 2005). Increasing the seed iron content by overexpression of the iron storage protein ferritin in rice grains and controlling the mineral status by overexpressing ion transporters are other strategies adopted. Ongoing efforts aim at improvising the rice grain with other micronutrients, specifically high-quality protein and vitamin E. Other approaches include screening germplasm collections for detecting variation in mineral content (Gregorio et al. 2000) and initiating rice breeding programmes to develop mineral-rich genotypes, using high iron/zinc germplasms and crossing it with plants containing other micronutrient traits (e.g. vitamin E, high protein,  $\beta$ -carotene). A different approach of rice biofortification can be adopted to regulate/reduce the levels of compounds such as phytate and tannins, both of which complex with minerals and prevent their absorption by the body during digestion. This approach simply aims to enhance mineral bioavailability (Welch and Graham 2004).

### 5.3 Enhanced Efficiency of Nutrient Use

Reducing dependency on fertiliser application can be achieved by gaining a deep understanding of the molecular mechanisms underlying nitrogen and phosphorus usage. Overexpression of *OsNRT1.1A* (*OsNPF6.3*), a member of the *Oryza sativa* nitrate transporter 1/peptide transporter family, in rice significantly improved nitrogen utilisation and grain yield; additionally, the maturation time was also shortened (Wang et al. 2018b). Overexpression of *OsNPF7.2*, a low-affinity nitrate transporter, significantly enhanced nitrate influx and promoted tillering, thereby improving grain yield (Wang et al. 2018a). Likewise, overexpression of *OsNRT2.1*, a high-affinity nitrate transporter, increases yield and manganese accumulation, while high expression of *OsNRT2.3b*, a pH-sensitive nitrate transporter, enhances the pH-buffering capacity of the plant, promoting plant adaptation, and also facilitates increased uptake of nitrogen, iron and phosphorus (Fan et al. 2016; Luo et al. 2018). The phosphate transporter gene *OsPht1;8* (*OsPT8*) increases Pi uptake and is involved in phosphate homeostasis in rice (Jia et al. 2011). Members of the sulphate transporter family in rice (*OsSULTR*) play an important role in regulating sulphur demand by the plant (Sasaki et al. 2016). *OsHAK5*, a high-affinity K transporter, controls potassium acquisition and distribution in the rice plant (Chen et al. 2017). Therefore,

identification and manipulation of potentially useful genes have to be done with an aim to maximise uptake, allocation and utilisation.

#### 5.4 Increased Resistance to Abiotic and Biotic Stress

Developing rice varieties that can withstand frequent and harsh changes in the climatic conditions is an important goal of rice research worldwide. These abiotic stresses include drought, flood, temperature (heat and cold), salinity and diverse soil problems like iron toxicity. Tremendous efforts are being put to identify genes or genomic regions responsible for conferring resistance or susceptibility to these conditions (Roychoudhury and Paul 2012; Gollmack et al. 2014). Breeders have adopted a breeding method called marker-assisted breeding, to accurately incorporate specific desirable traits into novel varieties at a faster rate. Through the use of advanced technology and contribution from national institutions, the International Rice Research Institute has developed 'climate change-ready rice' that are able to tolerate harsh climatic conditions (<https://www.iri.org/climate-change-ready-rice>). Several QTLs have been identified, which are being extensively studied to gain a better insight into their molecular and physiological processes. OsNAC10 and OsAHL1 confer drought tolerance and drought avoidance by regulating root development under stress conditions (Jeong et al. 2010; Zhou et al. 2016). The SUB1A gene that was derived from an Indian rice variety was found to conserve energy till floodwater reduces, thereby conferring resistance to submergence. Two ethylene response factors SNORKEL1 and SNORKEL2 enabled adaptation of rice to deep-water (Hattori et al. 2009). QTLs, e.g. *Ctb1* and *COLD1*, were associated with chilling tolerance, while *TT1* and *OsRab7* are linked to thermotolerance (Gardener and Kumar 2015). A major region of the rice genome, named as Saltol, has been identified by scientists that confers tolerance to salinity. Saltol is being exploited at large to develop crop varieties that can tolerate exposure to salinity stress at both seedling and reproductive stages (Thomson et al. 2010). This approach, combined with submergence tolerance trait, is targeted to rice varieties cultivated in coastal areas, where during rainy season, salinity and submergence are major problems. Additionally several QTLs have been identified from genetic maps that could be associated with iron or aluminium toxicity tolerance and mineral homeostasis. Several of these genes have also been integral parts of breeding programmes of climate change-ready rice.

The increase in the number of cloned disease resistance genes has been greatly accelerated by the map-based sequence of the rice genome. A molecular and functional understanding of these genes gave insights into devising strategies to prevent yield loss during bacterial blight and fungal blast infections, whose causal pathogens are *Xanthomonas oryzae* pv. *oryzae* and *Magnaporthe oryzae*, respectively. Additionally many genes have been fine-mapped, paving the path to identify functional markers. Marker-assisted selection was employed to introgress such genes into parent varieties to develop elite hybrids. The R gene in rice *Xa3* (also known as *Xa26*) displayed enhanced resistance level and enlarged resistance

spectrum (i.e. resistance to multiple *Xoo* races) when expressed under native promoter in *japonica* rice (Cao et al. 2007). The atypical protein with armadillo repeats, encoded by the *Ptr* gene, confers broad-spectrum blast resistance (Zhao et al. 2018). Involvement in conferring resistance to bacterial and fungal pathogens was also demonstrated for *OsGAP1* encoding a GTPase-activating protein (GAP), receptor-like kinases like OsBRR1 and XA21 and transcription factors like WRKY45 (Singh et al. 2018). Although marker-assisted selection provides an effective strategy for pyramiding of disease resistance genes, it can be combined with the transgenic approach using cloned genes for developing disease-resistant rice cultivars.

---

## 6 Conclusion and Future Perspectives

### 6.1 Rice Genome Sequence: A Milestone in an Expanse of Untrodden Information

Access to the sea of genomic sequence information of a cereal crop plant was a revolutionary achievement in the twenty-first century. This is solely because rice is a quintessential component of the diet and livelihoods of several million people globally. There is always a befitting demand to increase the yield, productivity, health and acceptability of the plant, which has been largely addressed across generations through traditional crop-breeding programmes. The nature of demands has undergone a noticeable change in the present global scenario, considering uncontrollable population growth, climate change, loss of cultivable land, greenhouse gas emissions and excessive use of pesticides and fertilisers. The rice genome sequence at hand has enabled map-based cloning of QTLs, marker-based characterisation of beneficial genes, functional analysis by developing functional knockouts with mutation or gene tagging and exploration of possibilities of introgression of these genes or genomic regions to develop new agronomically competent hybrids. Moreover, it also facilitated a better understanding of the successful ecogeographical adaptation of rice, genetic variation between wild and domesticated species and also the molecular components underlying agronomically beneficial phenotypes.

The opportunity to map genes to genome provides useful information regarding features of micro- and macro-level synteny of the rice genome to other cereal crops. Such comprehensive molecular linkage maps are the primary components of the toolkit required to breed improvised crops. Following functional assignments of genes, a large amount of information can be extended to other plant species, simply by the virtue of extensive colinearity in the conservation of gene order and content within members of the dicot and monocot families. Comparative mapping of grass genomes, using rice as the reference genome, revealed considerable macro-colinearity, represented by a graphical consensus map, 'Circle Diagram' (Moore et al. 1995). Further refinements of such maps gave the general impression that despite significant variations in DNA content or chromosome number, all the cereal crops examined tend to maintain similar gene order, thus maintaining significant

micro-colinearity at the mega-base level. Deletions, insertions, small-scale rearrangements or even gene amplification, duplication and translocations can account for the deviations observed. This extensive genome colinearity is exploited for fine mapping and map-based cloning, as several crop plants have large genomes and the extra information derived from the closely related model crop, rice, greatly facilitates cloning of genes in other crops.

Recently, an international resequencing effort of 3000 rice germplasm accessions in the Illumina-based next-generation sequencing (NGS) platform had been undertaken to understand the total genetic diversity within the *Oryza sativa* gene pool (3000 rice genomes project 2014). The International Rice Research Institute (IRRI) contributed 2466 accessions from its International Rice Genebank Collection (IRGC), while the remaining 534 accessions were obtained from the China National Crop Gene Bank (CNCGB) in the Institute of Crop Sciences, Chinese Academy of Agricultural Sciences (CAAS). The temperate *O. sativa* spp. *japonica* Nipponbare (Os-Nipponbare-Reference-IRGSP-1.0) was used as the reference genome for mapping of the selected reads, simply indicating that the significance, versatility, expanse and quality of this genome sequence. When aligned with the reference genome, innumerable single nucleotide polymorphisms could be mapped, depending on which the *O. sativa* gene pool could be differentiated into five varietal types of diverse origins: *indica*, *aus/boro*, *basmati/sadri*, tropical *japonica* and temperate *japonica*. This effort not only gave access to enormous volumes of SNP data but also a knowledge-based tool to discover allelic variants and associate important rice traits with diverse alleles. Additionally, a number of high-quality reference genomes also became accessible for rice genome research.

Another revolutionary development that happened alongside was the emergence of diverse omics web-based tools. Genome browsers and databases were developed and were gradually getting equipped with high-throughput analysis tools for comparative and functional genomics, transcriptomics, metabolomics, phylogenomics, interactomics, etc. Integration of information from different experiments—microarray, RNA-Seq, promoter analysis, yeast two-hybrid, mass spectrometry, co-immunoprecipitation, bimolecular fluorescence complementation and fluorescent protein localisation, to name a few—has now enabled researchers to understand the temporal and spatial expression patterns of a target gene, identify other targets that coexpress with the gene of interest, predict the functionality based on sequence and structural similarities, predict potential interactors and, thereby, possibly identify the signalling or metabolic pathway the candidate gene is involved in. In fact, several computational systems biology studies have been undertaken to clearly understand the mechanisms of salt tolerance or the regulatory crosstalks in the hormone biosynthesis pathways for modulating responses to biotic and abiotic stress (Wang et al. 2013; Deb et al. 2016). This advancement in the diversity, user accessibility and versatility of the omics-driven tools, software and programmes is gradually clearing up the ambiguity related with the rice genome.

However, this information explosion is meaningless and stagnated until a rice researcher is able to associate functional, structural or phenotypic implications with the corresponding genes. Despite the total number of genes worked out from the rice



whole-genome sequence, a total of ~3100 gene loci have been analysed till date for biological functions using various approaches (<http://funricegenes.ncpgr.cn/>). This disparity between analysed genes and total number of genes is due to the restrictions imposed by functional redundancy between the genes and the limited data available on gene function. Additionally several QTLs have been associated with different traits, but very few have been fine-mapped, identified or even explored for proper functions. At this stage of rice research, where there is a need for delivering a sustainable crop plant, more initiatives have to be undertaken to understand the functionality of the genome in entirety, to delineate the effect of allelic differences on gene functions, to dissect the molecular basis of ecogeographical adaptation of rice varieties and to explore the feasibility of manoeuvring these traits to our benefits.

Development of new improved varieties of cereal crops was a significant achievement of Green Revolution, which was possible through the efforts of Norman Borlaug, the father of the Green Revolution. Nearly half a century ago, the International Rice Research Institute (IRRI) developed IR8 variety of rice, also known as 'miracle rice' (Hargrove and Cabanilla 1979). This variety yielded more grains than the traditional ones, when grown with proper irrigation and nitrogen-rich fertilisers. But this miracle also cost a high price: excessive and indiscriminate use of chemical fertilisers and pesticides affected the environment, commercial fertilisers posed an additional cost to the rice farmers, and irrigation demands adversely affected the global water supply, which also faces additional pressure due to climate change, pollution, population growth, industrial requirements, etc. So the demand of an environment-friendly solution was addressed in a collaborative project between IRRI and the Chinese Academy of Agricultural Sciences (CAAS). Funded by the Bill & Melinda Gates Foundation (BMGF), the project aims to develop Green Super Rice (GSR) varieties that can ensure a stable and sustainable yield potential even when supplemented with fewer inputs or even if grown in unfavourable environmental conditions. In contrast to the breeding approaches undertaken during Green Revolution, where all other traits took a backseat compared to high yield, the GSR researchers adopted a different methodology. In this, a large number of backcrossed second-generation lines (BC2F2) and their succeeding generations were subjected to a range of biotic and abiotic stresses, following which weak lines could be eliminated and promising transgressive high-performing segregants could be identified. Several GSR varieties that are drought-tolerant; salinity-tolerant; submergence-tolerant; weed-tolerant; suitable for rainfed lowlands; multiply resistant to rice blast, rice planthoppers and gall midge; zero-input (no fertilisers and no pesticides); and high-yielding are now ready for seed exchange and germplasm distribution. GSR varieties that combine several of these traits are also in the pipeline, such that they can perform well in the toughest conditions, providing a stable and sustainable solution to the poorest farmers who cultivate rice. In recent times, the focus of GSR research has considerably changed as the additional challenge of balancing food security with preservation of natural resources and protection of the environment has come to light. The achievable solution is development of improved GSR varieties and development and diffusion of conventional agricultural practices that are environment-friendly as well.

**Acknowledgements** Financial support from the DBT-RA Program in Biotechnology and Life Sciences is gratefully acknowledged.

---

## References

- 3000 rice genomes project (2014) The 3,000 rice genomes project. *Gigascience* 1:7
- Adams MD et al (2000) The genome sequence of *Drosophila melanogaster*. *Science* 287:2185–2195
- Alter S et al (2015) Drought DB: an expert-curated compilation of plant drought stress genes and their homologs in nine species. *Database (Oxford)* 2015:bav046
- Altschul SF et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402
- Antonio BA et al (1996) Physical mapping of rice chromosomes 8 and 9 with YAC clones. *DNA Res* 3:393–400
- Ashikari M et al (1999) Rice gibberellin-insensitive dwarf mutant gene Dwarf 1 encodes the alpha-subunit of GTP-binding protein. *Proc Natl Acad Sci USA* 96:10284–10289
- Baba T et al (2000) Construction and characterization of rice genomic libraries: PAC library of Japonica variety, Nipponbare and BAC library of Indica variety, Kasalath. *Bull Natl Inst Agrobiol Resour* 14:41–52
- Barry GF (2001) The use of the Monsanto draft rice genome sequence in research. *Plant Physiol* 125:1164–1165
- Budiman MA et al (1999) Construction and characterization of rice Nipponbare BAC library. [http://www.genome.clemson.edu/rice\\_frame.html](http://www.genome.clemson.edu/rice_frame.html)
- Burge C, Karlin S (1997) Prediction of complete gene structures in human genomic DNA. *J Mol Biol* 268:78–94
- Camon EB et al (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. *BMC Bioinform* 6(1):17
- Cao Y et al (2007) The expression pattern of a rice disease resistance gene Xa3/Xa26 is differentially regulated by the genetic backgrounds and developmental stages that influence its function. *Genetics* 177(1):523–533
- Cao P et al (2008) Construction of a rice glycosyltransferase phylogenomic database and identification of rice-diverged glycosyltransferases. *Mol Plant* 1(5):858–877
- Cao P et al (2012) The rice oligonucleotide array database: an atlas of rice gene expression. *Rice* 5:17
- Chandran AKN, Jong K-H (2014) Resources for systems biology in rice. *J Plant Biol* 57:80–92
- Chang W-C et al (2008) Plant PAN: plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics* 9:561
- Chen M et al (2002) An integrated physical and genetic map of the rice genome. *Plant Cell* 14:537–545
- Chen G et al (2017) OsHAK1, a high-affinity potassium transporter, positively regulates responses to drought stress in rice. *Front Plant Sci* 8:1885
- Cheng Z et al (2001a) Toward a cytological characterization of the rice genome. *Genome Res* 11:2133–2141
- Cheng Z et al (2001b) High-resolution pachytene chromosome mapping of bacterial artificial chromosomes anchored by genetic markers reveals the centromere location and the distribution of genetic recombination along chromosome 10 of rice. *Genetics* 157:1749–1757
- Cheng Z et al (2002) Functional rice centromeres are marked by a satellite repeat and a centromere-specific retrotransposon. *Plant Cell* 14:1691–1704
- Chern CG et al (2007) A rice phenomics study—phenotype scoring and seed propagation of a T-DNA insertion-induced rice mutant population. *Plant Mol Biol* 65(4):427–438
- Chien CH et al (2015) EXPath: a database of comparative expression analysis inferring metabolic pathways for plants. *BMC Genomics* 16(Suppl 2):S6

- Chong J et al (2018) MetaboAnalyst 4.0: towards more transparent and integrative metabolomics analysis. *Nucleic Acids Res* 46(W1):W486–W494
- Civián P et al (2015) Three geographically separate domestications of Asian rice. *Nat Plants* 1:15164
- Conte MG et al (2008) GreenPhylDB: a database for plant comparative genomics. *Nucleic Acids Res* 36(Database issue):D991–D998
- Dardick C et al (2007) The rice kinase database. A phylogenomic database for the rice kinome. *Plant Physiol* 143(2):579–586
- Dash S et al (2012) PLEXdb: gene expression resources for plants and plant pathogens. *Nucleic Acids Res* 40(D1):D1194–D1201
- Deb A et al (2016) Regulatory cross-talks and cascades in rice hormone biosynthesis pathways contribute to stress signaling. *Front Plant Sci* 7:1303
- Delcher AL et al (1999) Alignment of whole genomes. *Nucleic Acids Res* 27:2369–2376
- Dharmawardhana P et al (2013) A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. *Rice (N Y)* 6(1):15
- Dong F et al (1998) Rice (*Oryza sativa*) centromeric regions consist of highly complex DNA. *Proc Natl Acad Sci USA* 95:8135–8140
- Dong Q et al (2005) Comparative plant genomics resources at PlantGDB. *Plant Physiol* 139(2):610–618
- Droc G et al (2006) OryGenesDB: a database for rice reverse genetics. *Nucleic Acids Res* 34(Database issue):D736–D740
- Eckardt NA (2000) Sequencing the rice genome. *Plant Cell* 12:2011–2017
- Eckardt NA (2004) Journey to the Center of the Genome: complete sequence of the Rice chromosome 8 centromere. *Plant Cell* 16(4):789–791
- Elbourne LDH et al (2017) TransportDB 2.0: a database for exploring membrane transporters in sequenced genomes from all domains of life. *Nucleic Acids Res* 45(Database issue):D320–D324
- Ewing B, Green P (1998) Base-calling of automated sequencer traces using PHRED. II. Error probabilities. *Genome Res* 8:186–194
- Ewing B et al (1998) Base-calling of automated sequencer traces using PHRED. I. Accuracy assessment. *Genome Res* 8:175–185
- Fan X et al (2016) Overexpression of a pH-sensitive nitrate transporter in rice increases crop yields. *Proc Natl Acad Sci USA* 113(26):7118–7123
- Freeling M (2001) Grasses as a single genetic system. Reassessment 2001. *Plant Physiol* 125(3):1191–1197
- Fukui K et al (1997) Identification of rice D-genome chromosomes by genomic *in situ* hybridization. *Theor Appl Genet* 95:1239–1245
- Gao Z et al (2003) Map-based cloning of the ALK gene, which controls the gelatinization temperature of rice. *Sci China Life Sci* 46:661–668
- Gardener C, Kumar SV (2015) Hot n' cold: molecular signatures of domestication bring fresh insights into environmental adaptation. *Mol Plant* 8(10):1439–1441
- Goff SA et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). *Science* 296:92–100
- Gollack D et al (2014) Tolerance to drought and salt stress in plants: unraveling the signaling networks. *Front Plant Sci* 5:151
- Goodstein DM et al (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40(Database issue):D1178–D1186
- Gordon D et al (1998) CONSED: a graphical tool for sequence finishing. *Genome Res* 8:195–202
- Gregorio G et al (2000) Breeding for trace mineral density in rice. *Food Nutr Bull* 21:382–386
- Gu H et al (2011) PRIN: a predicted rice interactome network. *BMC Bioinform* 12:161
- Hamada K et al (2011) OryzaExpress: an integrated database of gene expression networks and omics annotations in Rice. *Plant Cell Physiol* 52(2):220–229

- Hargrove TR, Cabanilla VL (1979) The impact of semi-dwarf varieties on Asian rice-breeding programs. *Bioscience* 29(12):731–735
- Harushima Y et al (1998) A high-density rice genetic linkage map with 2275 markers using a single F2 population. *Genetics* 148(1):479–494
- Hattori Y et al (2009) The ethylene response factors SNORKEL1 and SNORKEL2 allow rice to adapt to deep water. *Nature* 460(7258):1026–1030
- Hebsgaard SM et al (1996) Splice site prediction in Arabidopsis thaliana DNA by combining local and global sequence information. *Nucleic Acids Res* 24:3439–3452
- Helmy M et al (2011) OryzaPG-DB: Rice proteome database based on shotgun proteogenomics. *BMC Plant Biol* 11:63
- Heng HH et al (1997) FISH technology in chromosome and genome research. *Bioessays* 19:75–84
- Higo K et al (1998) PLACE: a database of plant cis-acting regulatory DNA elements. *Nucleic Acids Res* 26(1):358–359
- Hirochika H et al (1996) Retrotransposons of rice involved in mutations induced by tissue culture. *Proc Natl Acad Sci USA* 93:7783–7788
- Hong W-J et al (2019) Infrastructures of systems biology that facilitate functional genomic study in rice. *Rice* 12:15
- Hoque MS et al (2006) Over-expression of the rice OsAMT1-1 gene increases ammonium uptake and content, but impairs growth and development of plants under high ammonium nutrition. *Funct Plant Biol* 33:153–163
- International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409:860–921
- International Rice Genome Sequencing Project, Sasaki T (2005) The map-based sequence of the rice genome. *Nature* 436:793–800
- Jeong JS et al (2010) Root-specific expression of OsNAC10 improves drought tolerance and grain yield in rice under field drought conditions. *Plant Physiol* 153(1):185–197
- Jia H et al (2011) The phosphate transporter gene OsPht1;8 is involved in phosphate homeostasis in rice. *Plant Physiol* 156(3):1164–1175
- Jiang J et al (1995) Metaphase and interphase fluorescence *in situ* hybridization mapping of the rice genome with bacterial artificial chromosomes. *Proc Natl Acad Sci USA* 92:4487–4491
- Jin JP et al (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res* 45(D1):D1040–D1045
- Juliano BO (1992) Structure and function of the rice grain and its fractions. *Cereal Foods World* 7:772–774
- Jung K-H et al (2010) The Rice kinase Phylogenomics database: a guide for systematic analysis of the rice kinase super-family. *Trends Plant Sci* 15(11):595–599
- Jung K-H et al (2015) Phylogenomics databases for facilitating functional genomics in rice. *Rice (N Y)* 8:26
- Juretic N et al (2004) Transposable element annotation of the rice genome. *Bioinformatics* 20:155–160
- Kanehisa M, Goto S (2000) KEGG: kyotoencyclopedia of genes and genomes. *Nucleic Acids Res* 28(1):27–30
- Kawahara Y et al (2013) Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* 6:4
- Kawahara Y et al (2016) TENOR: database for comprehensive mRNA-Seq experiments in rice. *Plant Cell Physiol* 57(1):e7
- Kazama T, Toriyama KA (2003) Pentatricopeptide repeat containing gene that promotes the processing of aberrant atp6 RNA of cytoplasmic male-sterile rice. *FEBS Lett* 544:99–102
- Khus GS, Brar DS (2001) Rice genetics from mendel to functional genomics. In: *Rice genetics IV. Proceedings of the fourth international rice genetics symposium 2000*, 22–27, pp 3–25
- Kleffe J et al (1996) Logit linear models for the prediction of splice sites in plant pre-mrna sequences. *Nucleic Acids Res* 24:4709–4718

- Koike K et al (1997) Physical mapping of rice chromosomes 4 and 7 using YAC clones. *DNA Res* 4:27–33
- Komatsu S (2005) Rice proteome database: a step toward functional analysis of the rice genome. *Plant Mol Biol* 59(1):179–190
- Krishnan A et al (2017) RECoN: Rice environment Coexpression network for systems level analysis of abiotic-stress response. *Front Plant Sci* 8:1640
- Kurata N, Yamazaki Y (2006) Oryzabase. An integrated biological and genome information database for rice. *Plant Physiol* 140(1):12–17
- Kusunoki K, Yamamoto YY (2017) Plant promoter database (PPDB). *Methods Mol Biol* 1533:299–314
- Lai Z et al (1999) A shotgun optical map of the entire *Plasmodium falciparum* genome. *Nat Genet* 23:309–313
- Larmande P et al (2008) Oryza tag line, a phenotypic mutant database for the Genoplante rice insertion line library. *Nucleic Acids Res* 36(database issue):D1022–D1027
- Li X et al (2003) Control of tillering in rice. *Nature* 422(6932):618–621
- Li C et al (2006) Rice domestication by reducing shattering. *Science* 311:1936–1939
- Li Y et al (2014) Chalk5 encodes a vacuolar H(+)-translocating pyrophosphatase influencing grain chalkiness in rice. *Nat Genet* 46:398–404
- Lin J et al (1999) Whole genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285:1558–1562
- Liu L et al (2013) An integrative bioinformatics framework for genome-scale multiple level network reconstruction of Rice. *J Integr Bioinform* 10(2):223
- Lombard V et al (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 42:D490–D495
- Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25:955–964
- Lukashin AV, Borodovsky M (1998) GeneMark.Hmm: new solutions for gene finding. *Nucleic Acids Res* 26:1107–1115
- Luo B et al (2018) Overexpression of a high-affinity nitrate transporter OsNRT2.1 increases yield and manganese accumulation in Rice under alternating wet and dry condition. *Front Plant Sci* 9:1192
- Mao L et al (2000) Rice transposable elements: a survey of 73,000 sequence-tagged-connectors. *Genome Res* 10:982–990
- McCouch SR et al (2002) Development and mapping of 2240 new SSR markers for rice (*Oryza sativa* L.). *DNA Res* 9:257–279
- Mewes HW et al (1997) Overview of the yeast genome. *Nature* 387:7–8
- Meyer RS et al (2016) Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat Genet* 48:1083–1088
- Mi H et al (2017) PANTHER version 11: expanded annotation data from gene ontology and Reactome pathways, and data analysis tool enhancements. *Nucleic Acids Res* 45(Database issue):D183–D189
- Mohan M et al (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. *Mol Breed* 3:87–103
- Monna L et al (2002) Positional cloning of rice semidwarfing gene, sd-1: rice “green revolution gene” encodes a mutant enzyme involved in gibberellin synthesis. *DNA Res* 9(1):11–17
- Moore G et al (1995) Cereal genome evolution: grasses, line up and form a circle. *Curr Biol* 5:737–739
- Morris RT et al (2008) Osiris: an integrated promoter database for *Oryza sativa* L. *Bioinformatics* 24(24):2915–2917
- Nagaki K et al (2004) Sequencing of a rice centromere uncovers active genes. *Nat Genet* 36:138–145
- Naithani S et al (2017) Plant Reactome: a resource for plant pathways and comparative analysis. *Nucleic Acids Res* 45(D1):D1029–D1039

- Olsen KM et al (2006) Selection under domestication: evidence for a sweep in the rice waxy genomic region. *Genetics* 173:975–983
- Pampanwar V et al (2005) FPC web tools for Rice, maize, and distribution. *Plant Physiol* 138 (1):116–126
- Parsons JD (1995) Miropeats: graphical DNA sequence comparisons. *Comput Appl Biosci* 11:615–619
- Pertea M, Salzberg SL (2002) Using GlimmerM to find genes in eukaryotic genomes. *Curr Protocol Bioinformatics* 4(4.4)
- Priya P, Jain M (2013) RiceSRTFDB: a database of rice transcription factors containing comprehensive expression, cis-regulatory element and mutant information to facilitate gene function analysis. *Database (Oxford)* 2013:bat027
- Rice Full-Length cDNA Consortium (2003) Collection, mapping, and annotation of over 28,000 cDNA clones from *japonica* rice. *Science* 301(5631):376–379
- Robinson AJ et al (2018) AgriSeqDB: an online RNA-Seq database for functional studies of agriculturally relevant plant species. *BMC Plant Biol* 18:200
- Roychoudhury A, Banerjee A (2015) Transcriptome analysis of abiotic stress response in plants. *Transcriptomics* 3:e115
- Roychoudhury A, Paul A (2012) Abscisic acid-inducible genes during salinity and drought stress. In: Berhardt LV (ed) *Advances in medicine and biology*, vol 51. Nova Science, New York, pp 1–78
- Roychoudhury A, Datta K, Datta SK (2011) Abiotic stress in plants: from genomics to metabolomics. In: Tuteja N, Gill SS, Tuteja R (eds) *Omics and plant abiotic stress tolerance*. Bentham Science, Sharjah, pp 91–120
- Roychoudhury A, Paul S, Basu S (2013) Cross-talk between abscisic acid-dependent and abscisic acid-independent pathways during abiotic stress. *Plant Cell Rep* 32:985–1006
- Ruan SL et al (2011) Proteomic identification of OsCYP2, a rice cyclophilin that confers salt tolerance in rice (*Oryza sativa* L.) seedlings when overexpressed. *BMC Plant Biol* 11:34
- Ruprecht C et al (2016) FamNet: a framework to identify multiplied modules driving pathway expansion in plants. *Plant Physiol* 170(3):1878–1894
- Saji S et al (1996) Construction of YAC contigs on rice chromosome 5. *DNA Res* 3:297–302
- Saji S et al (2001) A physical map with yeast artificial chromosome (YAC) clones covering 63% of the 12 rice chromosomes. *Genome* 44:32–37
- Sakai H et al (2013) Rice annotation project database (RAP-DB): an integrative and interactive database for rice genomics. *Plant Cell Physiol* 54(2):e6
- Sakata K et al (1999) A computer program for prediction of gene domain on rice genome sequence. In *The 2nd Georgia Tech International Conference on Bioinformatics*, Abstracts, 78
- Sakata K et al (2000) INE: a rice genome database with an integrated map view. *Nucleic Acids Res* 28(1):97–101
- Sakata K et al (2002) RiceGAAS: an automated annotation system and database for rice genome sequence. *Nucleic Acids Res* 30:98–102
- Salamov AA, Solovyev VV (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Res* 10:516–522
- Sapkota A et al (2011) DIPOS: database of interacting proteins in *Oryza sativa*. *Mol Biosyst* 7 (9):2615–2621
- Sasaki T (1998) The rice genome project in Japan. *Proc Natl Acad Sci USA* 95(5):2027–2028
- Sasaki T, Burr B (2000) International Rice genome sequencing project: the effort to completely sequence the rice genome. *Curr Opin Plant Biol* 3(2):138–141
- Sasaki A et al (2016) Transporters involved in mineral nutrient uptake in rice. *J Exp Bot* 67 (12):3645–3653
- Sato Y et al (2011) RiceXPro: a platform for monitoring gene expression in *japonica* rice grown under natural field conditions. *Nucleic Acids Res* 39(Database issue):D1141–D1148
- Sato Y et al (2013) RiceFRIEND: a platform for retrieving coexpressed gene networks in rice. *Nucleic Acids Res* 41(Database issue):D1214–D1221

- Schaub P et al (2005) Why is Golden Rice Golden (yellow) instead of red? *Plant Physiol* 138 (1):441–450
- Shameer K et al (2009) STIFDB—Arabidopsis stress responsive transcription factor DataBase. *Int J Plant Genom* 2009:583429
- Sharma R et al (2013) Construction of a rice glycoside hydrolase phylogenomic database and identification of targets for biofuel research. *Front Plant Sci* 4:330
- Shimokawa T et al (1996) Assignment of YAC clones spanning rice chromosomes 10 and 12. *DNA Res* 3:401–406
- Singh K et al (1996) Centromere mapping and orientation of the molecular linkage map of rice (*Oryza sativa* L.). *Proc Natl Acad Sci USA* 93(12):6163–6168
- Singh PK et al (2018) Prospects of understanding the molecular biology of disease resistance in rice. *Int J Mol Sci* 19(4):1141
- Smit AFA et al (1996–2010) RepeatMasker Open-3.0. <http://www.repeatmasker.org>
- Soderlund C et al (2000) Contigs built with fingerprints, markers and FPC V4.7. *Genome. Research* 10(11):1772–1787
- Solovyev VV, Salamov AA (1997) The gene-finder computer tools for analysis of human and model organisms genome sequences. In: *Proceedings of the fifth international conference on intelligent systems for molecular biology*, pp 294–302
- Sonnhammer ELL, Östlund G (2015) In paranoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Res* 43:D234–D239
- Spielmeier W et al (2002) Semidwarf (sd-1), “green revolution” rice, contains a defective gibberellin 20-oxidase gene. *Proc Natl Acad Sci USA* 99:9043–9048
- Sun X et al (2004) Xa26, a gene conferring resistance to *Xanthomonas oryzae* pv. *Oryzae* in rice, encodes an LRR receptor kinase-like protein. *Plant J* 37:517–527
- Sun Q et al (2009) PPDB, the plant proteomics database at Cornell. *Nucleic Acids Res* 37(Database issue):D969–D974
- Sutton G et al (1995) TIGR assembler: a new tool for assembling large shotgun sequencing projects. *Genome Sci Technol* 1(1):9
- Sweeney MT et al (2007) Global dissemination of a single mutation conferring white pericarp in rice. *PLoS Genet* 3:e133
- Szklarczyk D et al (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* 43(database issue):D447–D452
- Takeya M et al (2011) NIASGBdb: NIAS Genebank databases for genetic resources and plant disease information. *Nucleic Acids Res* 39(Database issue):D1108–D1113
- Tan L et al (2008) Control of a key transition from prostrate to erect growth in rice domestication. *Nat Genet* 40:1360–1364
- Tanoue H et al (1997) Ordered YAC clone contigs assigned to rice chromosomes 3 and 11. *DNA Res* 4:133–140
- Tao Q et al (2002) One large-insert plant transformation-competent BIBAC library and three BAC libraries of *japonica* rice for genome research in rice and other grasses. *Theor Appl Genet* 105:1058–1066
- Tello-Ruiz MK et al (2018) Gramene 2018: unifying comparative genomics and pathway resources for plant research. *Nucleic Acids Res* 46(D1):D1181–D1189
- Temykh S et al (2001) Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): frequency, length variation, transposon associations, and genetic marker potential. *Genome Res* 11:1441–1452
- The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408:796–815
- The C. elegans Sequencing Consortium (1998) Sequence and analysis of the genome of *C. elegans*. *Science* 282:2012–2018
- Thimm O et al (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J* 37(6):914–939

- Thomson MJ et al (2010) Characterizing the Saltol quantitative trait locus for salinity tolerance in rice. *Rice* 3(2–3):148–160
- Tomcal M et al (2013) POGs2: a web portal to facilitate cross-species inferences about protein architecture and function in plants. *PLoS One* 8(12):e82569
- Tyagi AK, Mohanty A (2000) Rice transformation for crop improvement and functional genomics. *Plant Sci* 158:1–18
- Tyagi AK et al (2004) Structural and functional analysis of rice genome. *J Genet* 83(1):79–99
- Umehara Y et al (1995) Construction and characterization of a rice YAC library for physical mapping. *Mol Breed* 1(1):79–89
- Umehara Y et al (1996) An ordered yeast artificial chromosome library covering over half of rice chromosome 6. *Genome Res* 6:935–942
- Umehara Y et al (1997) Yeast artificial chromosome clones of rice chromosome 2 ordered using DNA markers. *DNA Res* 4:127–131
- United Nations Department of Economic and Social Affairs/Population Division (2017) 2017 Revision of World Population Prospects. <https://population.un.org/wpp/>
- Van Bel M et al (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res* 46(D1):D1190–D1196
- Venter JC et al (2001) The sequence of the human genome. *Science* 291:1304–1351
- Wang ZX et al (1996) Physical mapping of rice chromosome 1 with yeast artificial chromosomes (YACs). *DNA Res* 3:291–296
- Wang ZX et al (1999) The *Pib* gene for rice blast resistance belongs to the nucleotide binding and leucine-rich repeat class of plant disease resistance genes. *Plant J* 19:55–64
- Wang J et al (2011) Rice-map: a new-generation rice genome browser. *BMC Genomics* 12:165
- Wang J et al (2013) A computational systems biology study for understanding salt tolerance mechanism in rice. *PLoS One* 8(6):e64929
- Wang M et al (2014) The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat Genet* 46:982–988
- Wang J et al (2015) Artificial selection of *Gn1a* plays an important role in improving rice yields across different ecological regions. *Rice* 8:37
- Wang J et al (2018a) Rice nitrate transporter OsNPF7.2 positively regulates tiller number and grain yield. *Rice (N Y)* 11:12
- Wang W et al (2018b) Expression of the nitrate transporter gene OsNRT1.1A/OsNPF6.3 confers high yield and early maturation in rice. *Plant Cell* 30(3):638–651
- Welch RM, Graham RD (2004) Breeding for micronutrients in staple food crops from a human nutrition perspective. *J Exp Bot* 55:353–364
- Wing RA et al (2018) The rice genome revolution: from an ancient grain to Green super Rice. *Nat Rev Genet* 19:505–517
- Wu J et al (2002) A comprehensive rice transcript map containing 6591 expressed sequence tag sites. *Plant Cell* 14:525–535
- Wu J et al (2004) Composition and structure of the centromeric region of rice chromosome 8. *Plant Cell* 16:967–976
- Xia J, Wishart DS (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 26:2342–2344
- Xia L et al (2017) Rice expression database (RED): an integrated RNA-Seq-derived gene expression database for rice. *J Genet Genomics* 44(5):235–241
- Xu H et al (2016) The DENSE AND ERECT PANICLE 1 (DEP1) gene offering the potential in the breeding of high-yielding rice. *Breed Sci* 66(5):659–667
- Xue WY et al (2008) Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat Genet* 40:761–767
- Yamamoto K, Sasaki T (1997) Large-scale EST sequencing in rice. *Plant Mol Biol* 35:135–144
- Yamamoto E et al (2012) OGRO: the overview of functionally characterized genes in Rice online database. *Rice* 5:26



- Yamanaka S et al (2004) Identification of SNPs in the waxy gene among glutinous rice cultivars and their evolutionary significance during the domestication process of rice. *Theor Appl Genet* 108 (7):1200–1204
- Yan W-H et al (2011) A major QTL, *Ghd8*, plays pleiotropic roles in regulating grain productivity, plant height, and heading date in rice. *Mol Plant* 4(2):319–330
- Yang T et al (2014) The role of a potassium transporter *OsHAK5* in potassium acquisition and transport from roots to shoots in rice at low potassium supply levels. *Plant Physiol* 166:945–959
- Yano M et al (1997) Identification of quantitative trait loci controlling heading date in rice using a high-density linkage map. *Theor Appl Genet* 95(7):1025–1032
- Yeh S-Y et al (2015) Down-regulation of cytokinin oxidase 2 expression increases tiller number and improves rice yield. *Rice (N Y)* 8:36
- Yilmaz A et al (2009) GRASSIUS: a platform for comparative regulatory genomics across the grasses. *Plant Physiol* 149(1):171–180
- Yim WC et al (2013) PLANEX: the plant co-expression database. *BMC Plant Biol* 13:83
- Yoshimura S et al (1996) Identification of a YAC clone carrying the *Xa-1* allele, a bacterial blight resistance gene in rice. *Theor Appl Genet* 93:117–122
- Yu J et al (2002) A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). *Science* 296:79–92
- Yuan Q et al (2003) The TIGR rice genome annotation resource: annotating the rice genome and creating resources for plant biologists. *Nucleic Acids Res* 31:229–233
- Zang M (1997) Identification of protein coding region in the human genome based on quadratic discriminant analysis. *Proc Natl Acad Sci USA* 94:565–568
- Zdobnov EM, Apweiler R (2001) InterProScan—an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* 17:847–848
- Zhang Q (2007) Strategies for developing Green super rice. *Proc Natl Acad Sci USA* 104 (42):16402–16409
- Zhang Y et al (2004) Structural features of the rice chromosome 4 centromere. *Nucleic Acids Res* 32:2023–2030
- Zhang J et al (2006) RMD: a rice mutant database for functional analysis of the rice genome. *Nucleic Acids Res* 34(database issue):D745–D748
- Zhang Z et al (2010) PMRD: plant microRNA database. *Nucleic Acids Res* 38(database issue): D806–D813
- Zhang ZH et al (2012) Pleiotropism of the photoperiod insensitive allele of *Hd1* on heading date, plant height and yield traits in rice. *PLoS One* 7:e52538
- Zhang J et al (2015) Combinations of the *Ghd7*, *Ghd8* and *Hd1* genes largely define the ecogeographical adaptation and yield potential of cultivated rice. *New Phytol* 208 (4):1056–1066
- Zhang L et al (2017) A natural tandem array alleviates epigenetic repression of *IPA1* and leads to superior yielding rice. *Nat Commun* 8:14789
- Zhao H et al (2015) RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Res* 43(Database issue):D1018–D1022
- Zhao H et al (2018) The rice blast resistance gene *Ptr* encodes an atypical protein required for broad-spectrum disease resistance. *Nat Commun* 9:2039
- Zheng J et al (2015) Molecular functions of genes related to grain shape in rice. *Breed Sci* 65 (2):120–126
- Zhou L et al (2016) A novel gene *OsAHL1* improves both drought avoidance and drought tolerance in rice. *Sci Rep* 6:30264