3000 Genome Project: A Brief Insight

5

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

Abstract

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

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

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

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

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

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

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

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

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

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

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

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

Keywords

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

Abbreviations

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

5.1 Introduction

For majority of the world's poor, rice (*Oryza sativa* L.) offers most of the regular calories in their diet. However, due to the continuous growth of the human population, there is a continuous demand for food crops, including rice [1]. The global rice production is estimated to upsurge by 25% or more by the next decade (Seck et al.,

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

5.2 Germplasms and Sequencing of 3GRP

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

generated that have an average sequencing depth of 14X, average mapping rates, and genome coverage of 94.0% and 92.5%, respectively. Raw sequencing data are available from DDBJ (accession ERP005654), GigaDB (http://gigadb.org/dataset/200001), and EBi & NCBI (accession PRJEB6180).

5.3 Genome Size, Population Structure, and Genetic Diversity

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

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

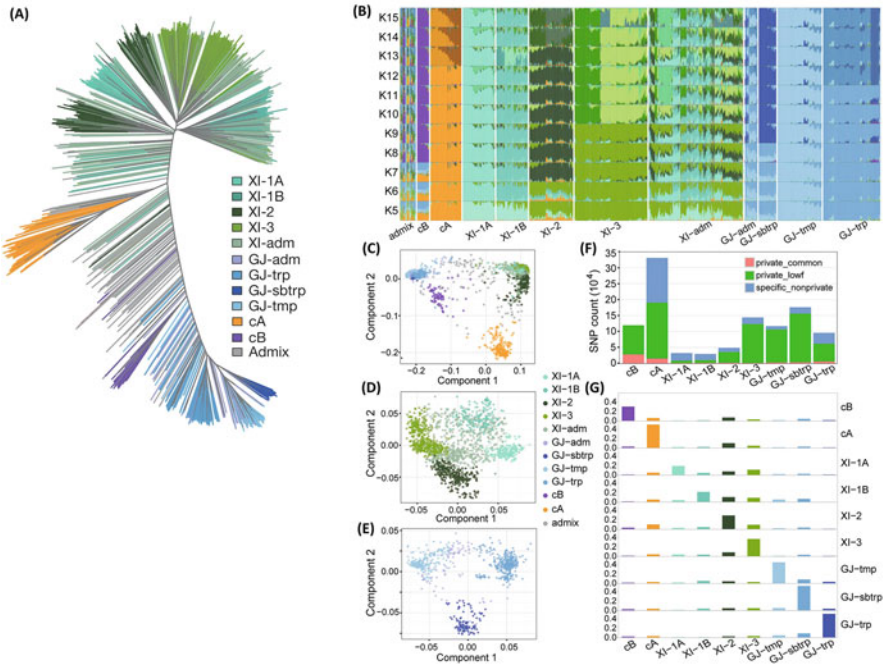

Fig. 5.1 (**a**) Unweighted neighbor-joining tree based on 3010 samples and computed on a simple matching distance matrix for filtered SNPs. (**b**) ADMIXTURE analyses for $k = 5$ to $k = 15$. (**c**–**d**) Multidimensional scaling plots for all ($n = 3010$) (**c**), XI ($n = 1786$) (**d**) and GJ ($n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

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

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

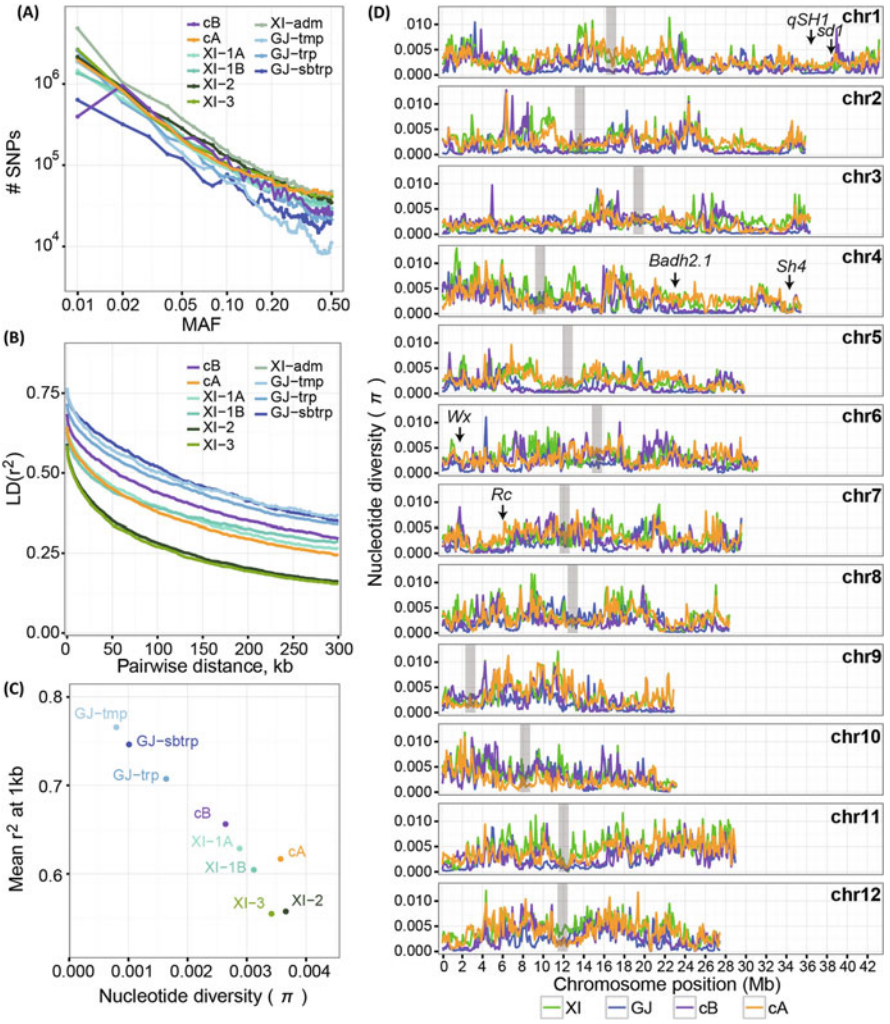

Fig. 5.2 (**a**) MAF histogram. (**b**) Genome-wide linkage disequilibrium. (**c**) Nucleotide diversity versus linkage disequilibrium. (**d**) Diversity scans (π) for all chromosomes for major groups (XI, GJ, cA, and cB) using 100-kb windows in which centromeric regions are highlighted in gray (Adapted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

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

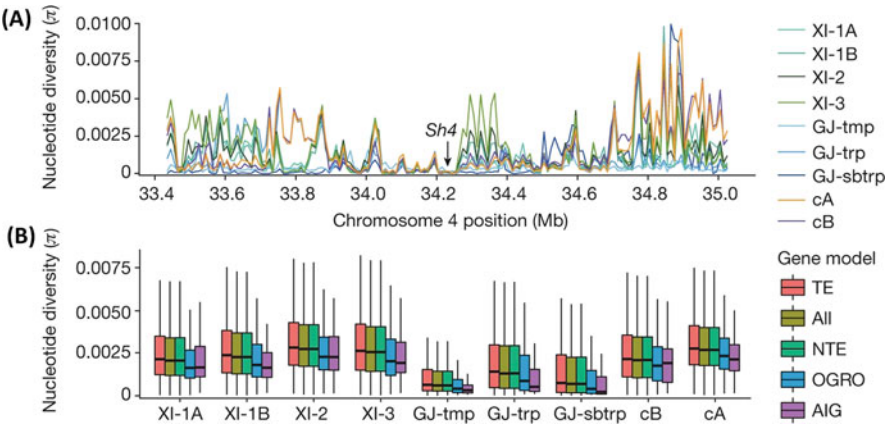

Fig. 5.3 (**a**) Differential nucleotide diversity between subpopulations at the *Sh4* locus on chromosome 4 using 10-kb sliding windows. (**b**) Box plots of the distribution of π in 100-kb regions surrounding gene models across the genome. Box plots are shown for $k = 9$ subpopulations for all 100-kb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

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

5.4 Structural Variations

Structural variants (SVs) identification and characterization have revolutionized the perception of the landscape of genetic variance in numerous organisms. A structural variant is usually characterized as a genome alteration (with respect to a reference

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

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

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

5.5 Conclusion and Future Perspective

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

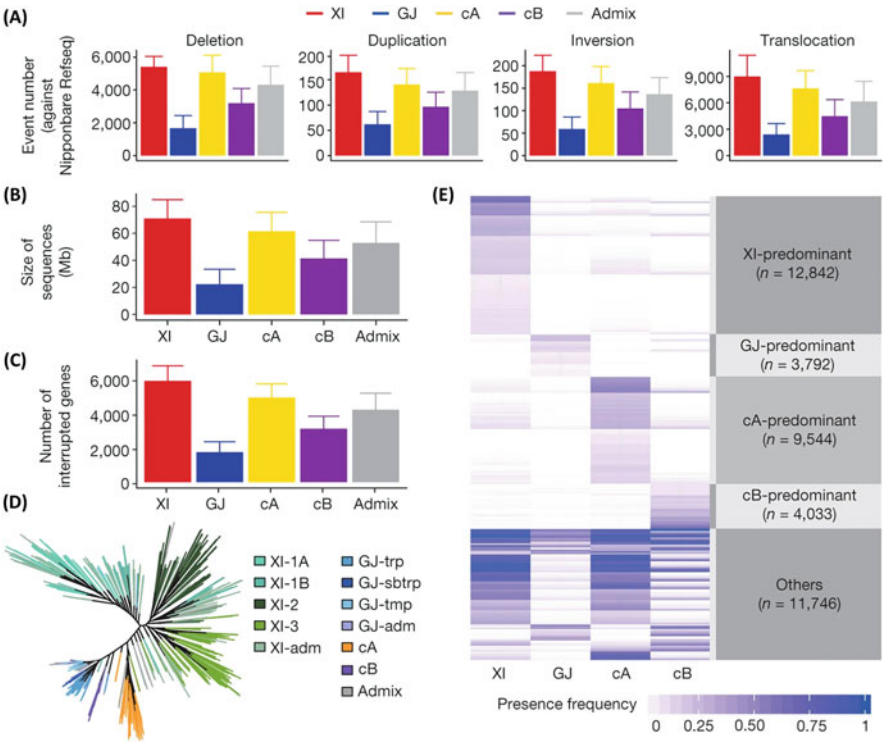

Fig. 5.4 (**a**) Number of deletions, duplications, inversions, and translocations. (**b**) Genome sizes affected by SVs. (**c**) Numbers of genes affected (included or interrupted) by the SVs. (**d**) Phylogenetic relationship of 453 rice accessions built from 10,000 randomly selected SVs. (**e**) Characterization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

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

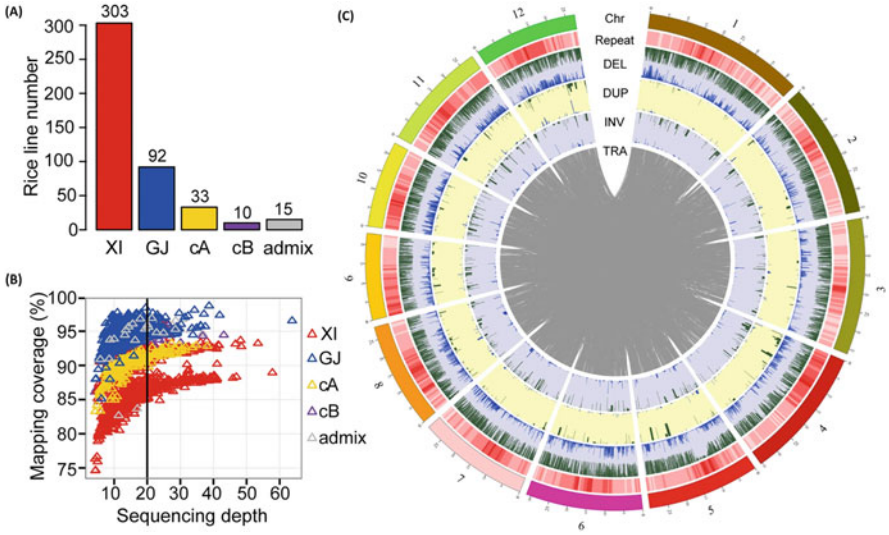

Fig. 5.5 (**a**) Number of accessions with sequencing depths ≥20× and mapping depth ≥15×. (**b**) Mapping coverage of the 3010 rice genomes to the Nipponbare RefSeq as a function of sequence depth. (**c**) Circular presentation of different types of structural variation detected in 453 high-coverage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

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

Conflict of Interest None.

References

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

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