Genomics Proteomics Bioinformatics 16 (2018) 397–404

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

REVIEW

Rice Genomics: over the Past Two Decades and into the Future

Shuhui Song [1,2,3,*,a], Dongmei Tian [1,b], Zhang Zhang [1,2,3,c], Songnian Hu [2,3,d] Jun Yu [2,3,*,e]

[1] *BIG Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[2] *CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China*
[3] *University of Chinese Academy of Sciences, Beijing 100049, China*

Received 19 December 2018; revised 14 January 2019; accepted 23 January 2019
Available online 13 February 2019

Handled by Xiangfeng Wang

KEYWORDS

Rice genome;
Genomic diversity;
Heterosis;
Domestication

Abstract Domestic rice (*Oryza sativa* L.) is one of the most important cereal crops, feeding a large number of worldwide populations. Along with various high-throughput genome sequencing projects, rice genomics has been making great headway toward direct field applications of basic research advances in understanding the molecular mechanisms of agronomical traits and utilizing diverse germplasm resources. Here, we briefly review its achievements over the past two decades and present the potential for its bright future.

Assembling and understanding the rice genomes using different sequencing approaches

Due to its limited genome size and diploidy, rice is an excellent choice among cereals for initiating genomic studies, serving as a model organism for plant biology and agricultural research. In 2002, the first two working draft genomes of the domestic rice (*Oryza sativa* L.) subspecies, *i.e.*, *japonica* (cultivar Nipponbare) and *indica* (cultivar 93-11), were successfully sequenced using whole-genome-shotgun (WGS) sequencing technology [1,2]. The International Rice Genome Sequencing Project (IRGSP) Consortium was launched in September 1997, comprising research groups from Japan, the United States, France, South Korea, India, and China, aimed at delineating the Nipponbare genome using a map-based clone-by-clone (CBC) strategy [3,4]. Meanwhile, there were two other efforts to sequence the same *japonica* cultivar from two private companies, Syngenta and Monsanto. Their sequencing data were publicly released in a controlled way and integrated into the IRGSP data. On the other hand, the *indica* genome project

* Corresponding authors.
E-mail: junyu@big.ac.cn (Yu J), songshh@big.ac.cn (Song S).
[a] ORCID: 0000-0003-2409-8770.
[b] ORCID: 0000-0003-0564-625X.
[c] ORCID: 0000-0001-6603-5060.
[d] ORCID: 0000-0003-3966-3111.
[e] ORCID: 0000-0002-2702-055X.

Peer review under responsibility of Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China.

https://doi.org/10.1016/j.gpb.2019.01.001

launched by the Chinese Superhybrid Rice Genome Project (CSRGP) went beyond genome sequencing and carried on with hybrid rice related studies [5]. In 2005, the complete genomes of both rice subspecies were released [6,7], covering 95% of the 389-Mb genome. Furthermore, 37,544 non-transposable-element (nTE)-related protein-coding genes were predicted, with TEs accounting for 34.79% of the genomes. The finished genome is providing essential information for positional cloning and molecular studies.

High-quality genome assemblies offer unprecedented information and insights into genomics, evolution, and biology of rice, even when there were only two species with genome sequences available for comparative analysis – *Arabidopsis thaliana* and *O. sativa*. The first dilemma that needs to be resolved is where to place TEs, inside or outside a gene, as both plant genomes have considerable amount of TE-related contents. Distinct rules apply concerning how genes and genomes are organized differentially between angiosperms and vertebrates, whose collinearity has been appreciated by comparative genomicists. It appears that the two major lineages of the animal and plant kingdoms divergently made their distinct choices much earlier in evolution, perhaps since the birth of their unicellular ancestors [8]. However, it has not yet been solved which lower lineages share the same genomic parameters (such as intron size limits and ratio of genic *vs*. intergenic spaces) with the higher lineages within the kingdom of unicellular eukaryotes. Furthermore, the members of such a kingdom are yet to be firmly classified phylogenetically, and the decisive cellular mechanisms – RNA splicing machineries – are rather diversified, with some even cryptic among those organisms [9]. Insertions and dynamics of TEs are not only relevant to the history of related genes and their functional regulation within narrow taxonomic groups but also play significant roles in genome evolution among higher taxonomic groups, such as the ratio of the long terminal repeat (LTR) elements, *Gypsy vs. Copia* [10]. The second dilemma is related to the accelerated mutation mechanisms; in this case, transcript-centric positive GC gradients become obvious in the genomes of Gramineae [11,12]. The gradients in GC content along the direction of transcription are not universal, which is shared by only the grass family of plants and warm-blooded vertebrates [11]. The third dilemma has to do with polyploidy and ancient whole genome duplication (WGD) events of plant genomes. Again, this dilemma is also shared between angiosperms and vertebrates, where polyploidy is prevalent among lower vertebrate and angiosperm lineages but abandoned by higher vertebrates, including reptiles, birds, and mammals [7,13,14].

One could certainly formulate more questions and dilemmas just at the genomic level. Moreover, it is well known that high-quality rice genome sequences and well-built references for its landraces, as well as subspecific and within-genus specific representatives, are essential for gene level comparative analysis. The revised reference genome assembly for Nipponbare (Os-Nipponbare-Reference-IRGSP-1.0) released in 2012 has been an excellent example, representing a product of concerted efforts using a variety of technologies, from optical mapping to CBC mapping, from clone-based assembly to heavy WGS coverage, and from long-read sequences of the Roche GS FLX to short-read sequences of the Illumina Genome Analyzer II platforms [15].

Harnessing genome sequences to understand the biological basis of heterosis

Heterosis (hybrid vigor) is a phenomenon wherein F1 hybrids bear superiority for multiple agronomic traits attributable to the mix of genetic contributions of its parental inbred lines [16]. This is important in the use of F1 hybrid cultivars that are often elite crop varieties selected by breeders. To meet increasing food demands from population growth, scientists have cultivated hundreds of rice superhybrids in the past few decades. Remarkably, until 2018, a total of 131 rice cultivars had been officially approved as superhybrids with high-yield potential by the Ministry of Agriculture of China (http://www.ricedata.cn/variety/superice.htm). Among them, Liangyou-pei-jiu (LYP9) is one of the representatives developed using a two-line crossing between PA64S and 93-11. The featured high yield, fine grain quality, and strong biotic resistance (bacterial leaf blight and fungal blast diseases) of LYP9 are attributed to its intersubspecific heterosis [17]. In addition, another widely-planted hybrid Shan-you 63 (SY63), was generated using a method named the three-line hybrid system and bred from a cross between the male-sterile Zhen-shan 97A (ZS97A) and the restorer line Ming-hui 63 (MH63). SY63 features superior yield, multiple disease resistance, wide adaptability, and good eating quality, leading to large-scale plantation in southern and central China over the past three decades [18].

Over the past decades, valuable efforts have been devoted to understanding the biological basis of heterosis, including transcriptomic and epigenomic analyses [19–21]. Although several traditional models of heterosis (such as dominance, overdominance, and epistasis) have been suggested to explain the increased yield [22], we still do not understand the molecular mechanisms of heterosis. It is vital to have the high-quality genome sequences of the hybrid parents, which ultimately allows hybrid gene mapping free of sequence gaps and at a single-base resolution. Quality assembly of the parental genomes of hybrids (SY63 and LYP9) have been recently reported [23–25]. A map-based sequencing effort to assemble the parental genomes of SY63, *i.e.*, ZS97 and MH63, yielded 237 contigs for ZS97 and 181 contigs for MH63, covering 90.6% and 93.2% of their estimated genome sizes, respectively [24]. Similarly, with the support of CSRGP, the two parental genomes of LYP9, 93-11 and PA64S, have also been sequenced with high coverage [25]. Consequently, a series of variety-specific genes have been determined through comparative genome studies.

Oryza glaberrima is a domestic rice species in Africa that is reproductively isolated from Asian rice. In the early 2000s, aimed to generate 'New Rice for Africa' (NERICA), introgressions were carried out by crossing *O. sativa* and *O. glaberrima* cultivars, followed by recurrent back-crossing with the Asian rice parent. In 2017, the genomes of TOG5681 and CG14, parents of two NERICA generations, were sequenced [26]. The complete genome sequences would provide a rich resource, helping to tackle the issue of reproductive isolation for potential hybrid breeding from other distantly-related rice species.

To further reveal genetic elements of heterosis, low-coverage resequencing efforts have been reported in different

superhybrid populations. One such effort dissected the immortalized second filial (IMF2) populations derived from the SY63 hybrid, showing the varied contribution of genetic components to yield traits [18]. For instance, overdominance/pseudo-over dominance contributes to a variety of yield-related traits (*e.g.*, the vigor of the yield, number of grains per panicle, and grain weight). In particular, the dominance × dominance interaction is closely associated with tillers per plant and grain weight. To map the heterotic loci at a fine scale, diverse superhybrid varieties and their inbred parental lines were massively resequenced [27,28]. These include the recombinant inbred lines (RILs) of the super hybrid rice LYP9 [29] and the F_2 lines from 17 representative hybrid rice crosses [30,31]. The genetic architecture of yield traits and numerous superior alleles that contribute to heterosis were then proposed [27,28]. Collectively, the availability of these parental genome sequences and hybrid population data provide rich resources for deciphering the genetic basis and molecular mechanisms of rice heterosis.

Revealing rice domestication processes by exploiting additional genome sequences of *Oryza* species

More genome sequences of different *Oryza* cultivars have been constantly added to the rice knowledge base, including several strains of Asian cultivars and African cultivars (*O. glaberrima*) with the AA genome in the past few years (**Figure 1**A, **Table 1**). The deeply sequenced *O. sativa* genomes include IR64 (a conventional *indica* rice variety in China) [32], IR8 (also known as Miracle rice) [33], Swarna (an *indica* rice cultivar variety with low glycemic index) [34], Shu-hui498 (R498, an *indica* rice variety cultivar used as a restorer line in a three-line hybrid system) [35], DJ123, and N22 (*indica* rice with important disease resistance and abiotic tolerance alleles) [32,33]. The African cultivars (*O. glaberrima*) provide an excellent resource for varietal improvement of *O. sativa*, as they harbor multiple important agronomic traits, especially for biotic and abiotic resistance. In 2014, CG14 was sequenced and assembled into 12 pseudomolecules with a total size of 316 Mb, and 33,163 gene models were annotated [26,36]. Most of these cultivars were *de novo* sequenced with high coverage and often through integration of both short and long reads from the NGS platforms, which ensure high sequence coverage and moderate contiguity. For instance, comparative analyses of mutations in three orthologous genes *O. sativa Shattering 1* (*OsSh1*), *O. sativa Shattering 4* (*OsSh4*), and *qSh1* (LOC_Os01g62920) from African and Asian rice confirm independent domestication of genes controlling panicle shattering. However, to better identify and compare the orthologous loci, higher contiguity and quality rice genomes are still highly desired for future sequencing.

Wild rice, being adapted to diverse geographical environments and exhibiting tolerance to biotic and abiotic stresses, can be exploited as important genetic resources and gene pools for molecular breeding. To date, there are 22 wild species in the genus *Oryza* that are distributed throughout the world, which are classified into ten genome types (AA, BB, CC, BBCC, CCDD, EE, FF, GG, HHJJ, and HHKK). Strategies to harness beneficial traits for crop improvement have been exemplified by the introgression of bacterial blight resistance gene *Xa21* from the wild rice *Oryza longistaminata* [37]. In 2003, the International *Oryza* Map Alignment Project (IOMAP) was initiated, with the aim of providing high-quality wild rice genomic resources for the discovery and utilization of beneficial genes and traits. So far, around 10 new reference genomes have been generated for wild rice species, including *O. nivara* (AA) and *Oryza rufipogon* (AA) from Asia; *Oryza barthii* (AA), *O. longistaminata* (AA), and *Oryza brachyantha* (FF) from Africa [36,38]; *Oryza glumaepatula* (AA) from South America; *Oryza meridionalis* (AA) from Australia; as well as *Oryza punctata* (BB) from Africa and *Oryza granulata* (GG) from China [33,36,39]. In addition, two novel perennial wild rice species from tropical Australia with AA genomes were also sequenced (one is similar to *O. rufipogon* in plant morphology, and the other is similar to the annual *O. meridionalis*) [40]. The available genomes of wild progenitors and close relatives provide valuable resources for the identification of candidate genes and chromosomal regions selected during domestication [41]. To date, many genes with significantly lower diversity unique to cultivated rice have been identified, representing candidate regions for selective sweeps during domestication [42–44]. Comparative genomic analyses between the wild and the cultivated rice species are essential for mechanistic investigation of genome organization, gene family expansion, segmental duplication, *etc.*

Dissecting genetic components for complex agronomic traits using genome-wide association studies

Building a comprehensive collection of landraces in terms of morphology, genetic diversity, and geography is fundamental for following genetic studies, such as genome-wide association studies (GWAS) of genotype-to-phenotype relatedness. Totally 773,948 rice accessions are available in various gene banks worldwide, with high genetic diversity [45]. For instance, there are ~101,000 from the International Rice Genebank Collection (IRGC) at the International Rice Research Institute (IRRI), 61,470 from the China National Crop Gene Bank (CCGB) [46], and ~18,000 from the United States Department of Agriculture (USDA) Rice Genebank [47]. Such collections enable population-based genome-wide studies for a broad scope of genetic and biological purposes.

An excellent example that utilizes a large number of rice accessions for GWAS by taking advantage of low-cost sequencing [48,49] was shown by Han and his coworkers. They performed GWAS analyses and identified hundreds of known and new loci associated with 14 agronomic traits, covering two morphological characteristics (leaf angle and tiller number), four grain-related traits (grain width, grain length, grain weight, and spikelet number), three grain quality traits (gelatinization temperature, protein content, and amylose content), three coloration traits (apiculus color, pericarp color, and hull color), and physiological features (heading date, drought tolerance, and degree of seed shattering). Another comprehensive study that involved metabolic profiling and metabolic GWAS (mGWAS) identified hundreds of common variants that exert important influences on the production of secondary metabolites, accordingly revealing the biochemical relevance of gene–metabolite associations in rice [50]. Using the same sequenced materials, many genetic loci were revealed to be related to biochemical traits (*e.g.*, absolute content of chlorophyll), physiological features (*e.g.*, seed germination and degree of seed

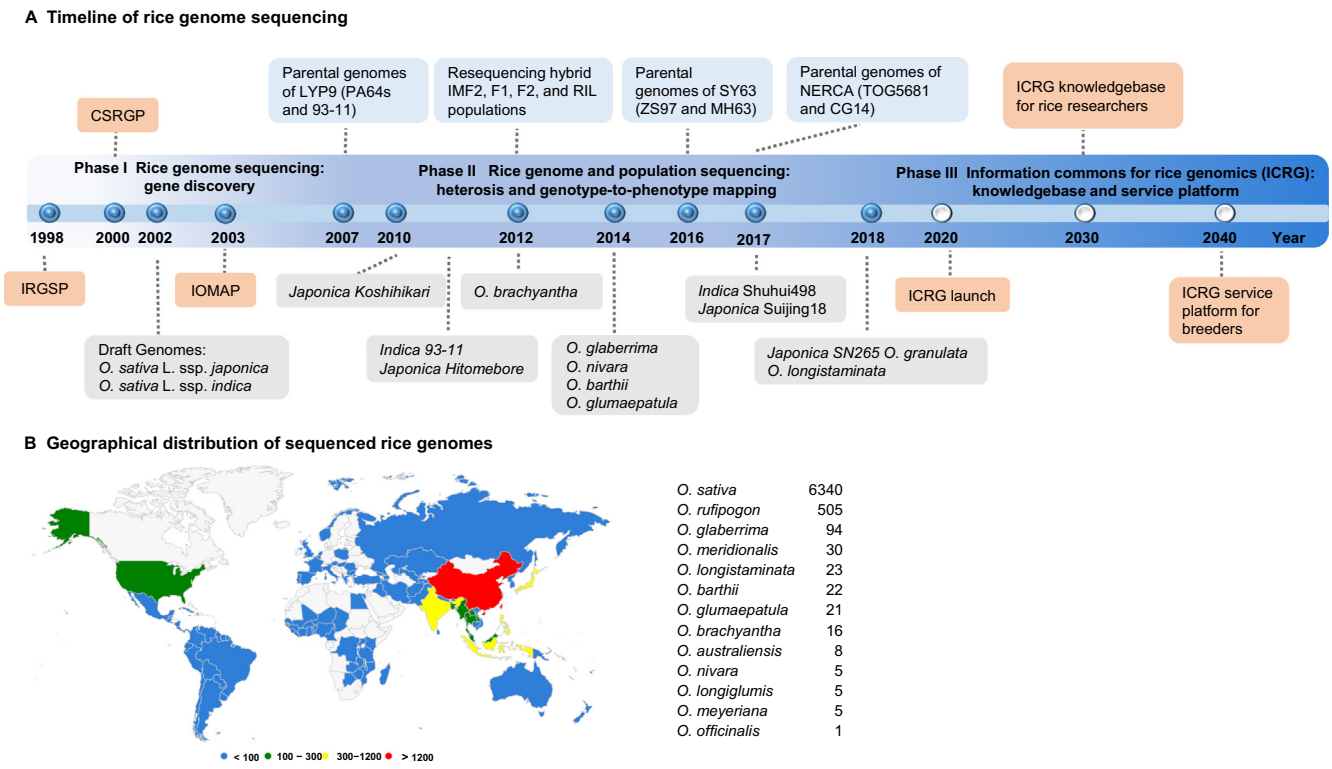

Figure 1 The timeline of rice genomics and geographical distribution of sequenced rice genomes

A. Timeline of rice genomics. The solid circles indicate past events of rice genomics, including Phase I for rice genome sequencing and Phase II for rice genome and population sequencing. The open circles indicate projected future events for Phase III. In Phase III, the Information Commons for Rice Genomics (ICRG) is expected to be built around the year 2020, covering all rice genome assemblies of high quality and free of sequence gaps, becomes an open-access knowledgebase for all rice researchers around the year 2030, and becomes a service platform for rice breeders to design their new crops around the year 2040. **B.** Geographical distribution of sequenced rice genomes. The numbers of the genomes sequenced per country were color coded (blue, < 100; green, 100–300; yellow, 300–1200; red, > 1200), and the total numbers of sequenced genomes from different rice species are listed on the right. IRGSP, International Rice Genome Sequencing Project; CSRGP, Chinese Superhybrid Rice Genome Project; IOMAP, International *Oryza* Map Alignment Project; IMF2, immortalized second filial; RIL, recombinant inbred lines; NERICA, New Rice for Africa; ICRG, Information Commons for Rice Genomics.

shattering), and content of mineral elements [51–55]. A further study reported a mapping effort for major-effect loci at the level of respectively causal SNPs, amylose content, seed length, and pericarp color by combining the diversity of the rice collection with low-coverage sequencing [56]. In short, such methods, by combining low-coverage genome-wide NGS-sequencing and GWAS, represent a complementary strategy for dissecting complex traits. However, there are still many genetic characteristics of important agronomic traits that have not yet been characterized. More studies are certainly required to reveal the genetic mechanisms by combining more phenotypic and genotypic data in natural populations in the near future.

Analyzing genomic diversity through comparative genomic studies and data integration

More rice genomes with high quality sequences have provided unlimited opportunities for identifying genetic and other molecular markers, *e.g.*, SNPs, or simple sequence repeats (SSRs), which greatly facilitates population-based studies and marker-assisted breeding. To build an open-access information commons for rice genomics (ICRG) would be desirable to host genome assemblies and genome variations in the future and hopefully to integrate other large-scale genome annotations, including information from other omics-level collections. There are several databases available to be integrated, such as Ensembl Genome [57], Gramene [58], RAP-DB [59,60], RGAP [15], dbSNP at NCBI [61], HapRice [62], SNP-Seek [63,64], IC4R [65] in BIG Data Center [66], RiceBase [67], and RiceVarMap [68]. Notably, since 2017, the largest collection of rice SNPs we have organized is deposited in the GVM database (http://bigd.big.ac.cn/gvm/) by collecting and systemically analyzing sequence data of 5152 rice accessions (Figure 1B), in which a total of 18,616,579 SNPs and 9122 manually curated genotype-to-phenotype entries were integrated [69]. Furthermore, more than 10,000 novel, full-length, protein-coding genes and a high number of presence-absence variations (PAVs) were identified by resequencing a core collection of over 3000 Asian cultivated rice accessions from 89 countries, representing another component of species genetic diversity [70–72]. These sequence variations and resources are useful for population structure and diversity analysis. One such example is utilizing a large number of genomic variations for population-based phylogenomic analyses,

Table 1 Genome sequence resources of *Oryza* species

Rice	Genome type	Genome size (Mb)	Sequencing depth (×)	Assembly level	Database	Weblink	PMID
O. sativa L. ssp *japonica* (cv. Nipponbare) IRGSP1.0	AA	374.4	NA	Chr	RAP-DB	https://rapdb.dna.affrc.go.jp/	23299411 18089549
	AA	374.4	NA	Chr	RGAP	http://rice.plantbiology.msu.edu/	24280374
	AA	374.4	NA	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCF_001433935.1	24280374
O. sativa L. ssp. *japonica* (cv. Nipponbare)	AA	391.1	NA	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000149285.1	15685292
O. sativa L. ssp. *japonica* (cv. Nipponbare)	AA	379.6	117	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_003865235.1	30448535
O. sativa L. ssp. *japonica* (cv. Nipponbare)	AA	355.6	110	Scaffold	CSHL	http://schatzlab.cshl.edu/data/rice	25468217
O. sativa L. ssp. *indica* (cv. 93-11)	AA	395.4	116	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_003865215.1	30448535
O. sativa L. ssp. *indica* (cv. 93-11)	AA	426.3	NA	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000004655.2	11935017
O. sativa L. ssp. *indica* (cv. HR-12)	AA	389.8	93	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000725085.2	26984283
O. sativa L. ssp. *indica* (cv. RP Bio-226)	AA	352.1	20	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001305255.1	NA
O. sativa L. ssp. *indica* (cv. Shuhui498)	AA	391.0	120	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_002151415.1	28469237
O. sativa L. ssp. *indica* (cv. Minghui 63)	AA	398.8	228	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001618785.1	27535938
	AA	387.4	120	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001623365.2	27622467
O. sativa L. ssp. *indica* (cv. Zhenshan 97)	AA	386.5	253	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001618795.1	27535938
	AA	387.3	120	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001623345.2	27622467
O. sativa L. ssp. *indica* (cv. IR8)	AA	389.1	70	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001889745.1	29358651
O. sativa L. ssp. *indica* (cv. IR64)	AA	345.2	110	Scaffold	CSHL	http://schatzlab.cshl.edu/data/rice	25468217
O. sativa L. ssp. *indica* (cv. Swarna)	AA	NA	10	NA	NA	NA	26068787
O. sativa L. ssp. *japonica* (cv. HEG4)	AA	342.0	220	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000817615.1	29158416
O. sativa L. ssp. *japonica* (cv. Hitomebore)	AA	382.6	179	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000321445.1	NA
O. sativa L. ssp. *japonica* (cv. A123)	AA	337.7	250	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000817635.1	29158416
O. sativa L. ssp. *japonica* (cv. Koshihikari)	AA	382.2	15.7	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_000164945.1	20423466
O. sativa L. ssp. *indica* (cv. N 22)	AA	362.3	65	Chr	NCBI	https://www.ncbi.nlm.nih.gov/assembly/GCA_001952365.1	29358651
O. sativa L. ssp. *indica* (cv. DJ123)	AA	345.9	110	Scaffold	CSHL	http://schatzlab.cshl.edu/data/rice	25468217
O. rufipogon (cv.W1943)	AA	339.2	130	Scaffold	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000817225.1	29358651
O. barthii	AA	308.3	110	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000182155.2	29358651
O. glaberrima (CG14)	AA	303.3	30	Scaffold	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000147395.2	29358651 25368197
O. glumaepatula	AA	372.9	135	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000576495.1	29358651
O. meridionalis	AA	335.7	166	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000338895.2	29358651
O. punctata	BB	393.8	130	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000573905.1	29358651
O. brachyantha	FF	259.9	104	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000231095.2	29358651
O. nivara	AA	338.0	102	Chr	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000576065.1	29358651
O. longistaminata	AA	326.4	52.5	Scaffold	NCBI, Ensembl	https://www.ncbi.nlm.nih.gov/assembly/GCA_000789195.1/	NA
O. granulate	GG	777.0	117	Contig	GWH	http://bigd.big.ac.cn/gwh/Assembly/116/show	30271965

Note: ssp, subspecies; cv, cultivar; Chr, chromosome; NA, not applicable; RAP-DB, Rice Annotation Project database; RGAP, Rice Genome Annotation Project; GWH, Genome Warehouse.

providing evidence for the variety of the *O. sativa* gene pool in 5 major different groups – *indica, aus/boro, basmati/sadri, tropical japonica, and temperate japonica*, and in some unknown subpopulations related to geographic location [70].

SSR markers are another class of molecular marker that are widely used in gene mapping and breeding practice. They are also the primary choice for genotyping due to their high density, codominant inheritance, high allelic diversity, and highly reproducible methodology for detection. An excellent example for marker-assisted backcrossing breeding using SSR markers is to integrate rice blast resistance genes into a number of popular rice varieties to improve the blast disease resistance [73,74]. The current genome coverage by both SSR and SNP markers is abundant for marker-assisted selection (MAS) and QTL (quantitative trait loci) mapping.

Furthering function-centric and trait-centric gene cloning

The advances in rice genome sequencing projects have greatly boosted functional genomic studies. These studies aimed at exploring genes and regulatory networks of agronomically important traits and their application in varietal improvement, which include but are not limited to, yield, grain quality, reproductive development, and resistance to disease, pests, or abiotic stress. Over the past decades, scientists have used various platforms successfully for functional genomics, such as large-scale mutant libraries [75,76], core germplasm collections, high-density gene expression arrays, and transcriptome sequencing [77]. In doing so, they have defined a number of trait-related genes with agronomic importance. Collectively, a total of 2358 rice functional genes (http://www.ricedata.cn/) were successfully cloned using map-based cloning techniques, including genes related to grain yield, grain size/weight, and grain quality [76,78]. For instance, *GW5*, which regulates cell division during seed development, affects grain width [79,80]; whereas the recently identified plant-specific transcription factor 13 (*OsSPL13*) appears to increase grain length [81].

Another typical example is genes involved in regulating plant architecture by controlling tillering and promoting panicle branching. For instance, *OsSPL14* (also known as ideal plant architecture, *1IPA1*), one of the *OsmiR156* targets that interacts with TEOSINTE BRANCHED1, negatively regulates tiller bud outgrowth [82–84]. In addition, *OsSPL14* also regulates the length and grain numbers of panicles by directly interfering with dense and erect panicle 1 (*DEP1*), a key protein determining panicle architecture [84]. The introduction of the *OsSPL14ipa1* allele into Xiushui 11 (XS11) results in approximately an increase of 11% in grain yield [85]. Additional details on achievements of rice functional genomics have been well described in a recent review [86].

Future perspectives

As the global human population is projected to reach 9 billion by 2050, rice researchers and breeders, together with those working on the other two major cereal crops – wheat and corn – are pressured to forge ahead to make decisive contributions to the prevention of potential food crises along the way. To fulfill such a challenging achievement, rice genomics and information integration must be conducted continuously, and an all-in effort from the rice research community would be needed to build the ICRG as a platform for the curation and annotation of sharable resources, which are not limited to data and knowledge but also experimental materials (Figure 1).

First, together with the introduction and application of the third-generation sequencers, we envisage that ICRG will contain more high-quality, gap-free genome sequences acquired systematically in the next decade or so from the existing germplasms, which may be expanded to other cereal crops and their wild counterparts. An international effort is in principle the best choice to unite rice scientists around the globe to build a platform upon which to collect data, to exchange information, and to share knowledge. As data accumulate, this platform must be organized to incorporate information from multiple omics levels, such as epigenomics, ribogenomics, proteomics, and metabolomics. Second, a significant effort must focus on gene-level genome annotations based on intensive comparative analyses among cultivars, wild counterparts, and elite hybrids. Most difficulties are expected to come from three basic components: defining all functional genes and their variants, annotating all TEs, and distinguishing orthologous and paralogous genes and their functional distinctions. Specialized databases have to be built and curated by dedicated scientists, in which genome polyploidy and chromosomal level regulatory principles and mechanisms are most likely involved. Third, ICRG dedicated to rice biotechnology must be built by rice genomicists and bioinformaticians for end users, such as rice biologists and crop breeders. Because the most likely tools for genetically modified crops now appear to be genome or gene editing in addition to the conventional tools of genetic engineering and hybridization [87], it is a necessity that genome assemblies be of ultimate quality and contiguity. Both are not effortless when working with the current state-of-art technological toolboxes. All possible future milestones are marked in the timeline of rice genomics (Figure 1A).

Competing interests

The authors declare no competing interests.

Acknowledgments

The authors acknowledge support from the Youth Innovation Promotion Association of the Chinese Academy of Sciences, China (Grant No. 2017141) awarded to SS, as well as the Strategic Priority Research Program (Grant No. XDA08010304), Key Research Program of Frontier Sciences (Grant No. QYZDY-SSW-SMC017), and R&D Projects of Scientific Research Equipment Programs (Grant Nos. YZ201568 and YZ201402) of the Chinese Academy of Sciences, China awarded to JY.

References

[1] Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *japonica*). Science 2002;296:92–100.

[2] Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, et al. A draft sequence of the rice genome (*Oryza sativa* L. ssp. *indica*). Science 2002;296:79–92.
[3] Eckardt NA. Sequencing the rice genome. Plant Cell 2000;12:2011–7.
[4] Sasaki T, Burr B. International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. Curr Opin Plant Biol 2000;3:138–41.
[5] Yu J, Wong GK, Liu S, Wang J, Yang H. A comprehensive crop genome research project: the Superhybrid Rice Genome Project in China. Philos Trans R Soc Lond B Biol Sci 2007;362:1023–34.
[6] International Rice Genome Sequencing Project. The map-based sequence of the rice genome. Nature 2005;436:793–800.
[7] Yu J, Wang J, Lin W, Li S, Li H, Zhou J, et al. The genomes of *Oryza sativa*: a history of duplications. PLoS Biol 2005;3:e38.
[8] Yu J, Wong GKS, Wang J, Yang H. Shotgun sequencing (SGS). In: Meyer RA, editor. Encyclopedia of molecular cell biology and molecular medicine. Germany: Wiley-VCH; 2006. p. 71–114.
[9] Wu J, Xiao J, Wang L, Zhong J, Yin H, Wu S, et al. Systematic analysis of intron size and abundance parameters in diverse lineages. Sci China Life Sci 2013;56:968–74.
[10] Al-Mssallem IS, Hu S, Zhang X, Lin Q, Liu W, Tan J, et al. Genome sequence of the date palm *Phoenix dactylifera* L. Nat Commun 2013;4:2274.
[11] Wong GK, Wang J, Tao L, Tan J, Zhang J, Passey DA, et al. Compositional gradients in *Gramineae* genes. Genome Res 2002;12:851–6.
[12] Wang J, Zhang J, Li R, Zheng H, Li J, Zhang Y, et al. Evolutionary transients in the rice transcriptome. Genomics Proteomics Bioinformatics 2010;8:211–28.
[13] Rodriguez F, Arkhipova IR. Transposable elements and polyploid evolution in animals. Curr Opin Genet Dev 2018;49:115–23.
[14] MacKintosh C, Ferrier DEK. Recent advances in understanding the roles of whole genome duplications in evolution. F1000Res, 6. p. 1623.
[15] 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 (N Y) 2013;6:4.
[16] Birchler JA. Heterosis: the genetic basis of hybrid vigour. Nat Plants 2015;1:15020.
[17] Cheng SH, Zhuang JY, Fan YY, Du JH, Cao LY. Progress in research and development on hybrid rice: a super-domesticate in China. Ann Bot 2007;100:959–66.
[18] Zhou G, Chen Y, Yao W, Zhang C, Xie W, Hua J, et al. Genetic composition of yield heterosis in an elite rice hybrid. Proc Natl Acad Sci U S A 2012;109:15847–52.
[19] Song S, Qu H, Chen C, Hu S, Yu J. Differential gene expression in an elite hybrid rice cultivar (*Oryza sativa*, L) and its parental lines based on SAGE data. BMC Plant Biol 2007;7:49.
[20] Ge X, Chen W, Song S, Wang W, Hu S, Yu J. Transcriptomic profiling of mature embryo from an elite super-hybrid rice LYP9 and its parental lines. BMC Plant Biol 2008;8:114.
[21] Wei G, Tao Y, Liu GZ, Chen C, Luo RY, Xia HA, et al. A transcriptomic analysis of superhybrid rice LYP9 and its parents. Proc Natl Acad Sci U S A 2009;106:7695–701.
[22] Goff SA, Zhang Q. Heterosis in elite hybrid rice: speculation on the genetic and biochemical mechanisms. Curr Opin Plant Biol 2013;16:221–7.
[23] Zhang J, Chen LL, Sun S, Kudrna D, Copetti D, Li W, et al. Building two indica rice reference genomes with PacBio long-read and Illumina paired-end sequencing data. Sci Data 2016;3:160076.
[24] Zhang J, Chen LL, Xing F, Kudrna DA, Yao W, Copetti D, et al. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. Proc Natl Acad Sci U S A 2016;113:E5163–71.
[25] Wang D, Xia Y, Li X, Hou L, Yu J. The Rice Genome Knowledgebase (RGKbase): an annotation database for rice comparative genomics and evolutionary biology. Nucleic Acids Res 2013;41:D1199–205.
[26] Wang M, Yu Y, Haberer G, Marri PR, Fan C, Goicoechea JL, et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. Nat Genet 2014;46:982–8.
[27] Huang XH, Yang SH, Gong JY, Zhao Y, Feng Q, Gong H, et al. Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nat Commun 2015;6:6258.
[28] Zhen G, Qin P, Liu KY, Nie DY, Yang YZ, Deng XW, et al. Genome-wide dissection of heterosis for yield traits in two-line hybrid rice populations. Sci Rep 2017;7:7635.
[29] Gao ZY, Zhao SC, He WM, Guo LB, Peng YL, Wang JJ, et al. Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. Proc Natl Acad Sci U S A 2013;110:14492–7.
[30] Huang X, Yang S, Gong J, Zhao Q, Feng Q, Zhan Q, et al. Genomic architecture of heterosis for yield traits in rice. Nature 2016;537:629–33.
[31] Gong JY, Miao JS, Zhao Y, Zhao Q, Feng Q, Zhan QL, et al. Dissecting the genetic basis of grain shape and chalkiness traits in hybrid rice using multiple collaborative populations. Mol Plant 2017;10:1353–6.
[32] Schatz MC, Maron LG, Stein JC, Hernandez Wences A, Gurtowski J, Biggers E, et al. Whole genome de novo assemblies of three divergent strains of rice, *Oryza sativa*, document novel gene space of *aus* and *indica*. Genome Biol 2014;15:506.
[33] Stein JC, Yu Y, Copetti D, Zwickl DJ, Zhang L, Zhang C, et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. Nat Genet 2018;50:285–96.
[34] Rathinasabapathi P, Purushothaman N, Ramprasad VL, Parani M. Whole genome sequencing and analysis of Swarna, a widely cultivated indica rice variety with low glycemic index. Sci Rep 2015;5:11303.
[35] Du H, Yu Y, Ma Y, Gao Q, Cao Y, Chen Z, et al. Sequencing and *de novo* assembly of a near complete indica rice genome. Nat Commun 2017;8:15324.
[36] Zhang QJ, Zhu T, Xia EH, Shi C, Liu YL, Zhang Y, et al. Rapid diversification of five *Oryza* AA genomes associated with rice adaptation. Proc Natl Acad Sci U S A 2014;111:E4954–62.
[37] Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, et al. A receptor kinase-like protein encoded by the rice disease resistance gene, Xa21. Science 1995;270:1804–6.
[38] Chen J, Huang Q, Gao D, Wang J, Lang Y, Liu T, et al. Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. Nat Commun 2013;4:1595.
[39] Wu Z, Fang D, Yang R, Gao F, An X, Zhuo X, et al. *De novo* genome assembly of *Oryza granulata* reveals rapid genome expansion and adaptive evolution. Commun Biol 2018;1:84.
[40] Brozynska M, Copetti D, Furtado A, Wing RA, Crayn D, Fox G, et al. Sequencing of Australian wild rice genomes reveals ancestral relationships with domesticated rice. Plant Biotechnol J 2017;15:765–74.
[41] Wang L, Hao L, Li X, Hu S, Ge S, Yu J. SNP deserts of Asian cultivated rice: genomic regions under domestication. J Evol Biol 2009;22:751–61.
[42] Xu X, Liu X, Ge S, Jensen JD, Hu F, Li X, et al. Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nat Biotechnol 2011;30:105–11.
[43] He Z, Zhai W, Wen H, Tang T, Wang Y, Lu X, et al. Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genet 2011;7:e1002100.
[44] Huang X, Kurata N, Wei X, Wang ZX, Wang A, Zhao Q, et al. A map of rice genome variation reveals the origin of cultivated rice. Nature 2012;490:497–501.

[45] Allender C. The second report on the state of the world's plant genetic resources for food and agriculture. 2010, p. 370. http://www.fao.org/docrep/013/i1500e/i1500e00.htm.
[46] Zhang H, Zhang D, Wang M, Sun J, Qi Y, Li J, et al. A core collection and mini core collection of *Oryza sativa* L. in China. Theor Appl Genet 2011;122:49–61.
[47] Agrama HA, Yan WG, Lee F, Fjellstrom R, Chen MH, Jia M, et al. Genetic assessment of a mini–core subset developed from the USDA rice genebank. Crop Sci 2009;49:1336–46.
[48] Huang X, Wei X, Sang T, Zhao Q, Feng Q, Zhao Y, et al. Genome-wide association studies of 14 agronomic traits in rice landraces. Nat Genet 2010;42:961–7.
[49] Huang X, Zhao Y, Wei X, Li C, Wang A, Zhao Q, et al. Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. Nat Genet 2011;44:32–9.
[50] Chen W, Gao Y, Xie W, Gong L, Lu K, Wang W, et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. Nat Genet 2014;46:714–21.
[51] Matsuda F, Nakabayashi R, Yang Z, Okazaki Y, Yonemaru J, Ebana K, et al. Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J 2015;81:13–23.
[52] Wang Q, Xie W, Xing H, Yan J, Meng X, Li X, et al. Genetic architecture of natural variation in rice chlorophyll content revealed by a genome-wide association study. Mol Plant 2015;8:946–57.
[53] Wang Q, Zhao H, Jiang J, Xu J, Xie W, Fu X, et al. Genetic architecture of natural variation in rice nonphotochemical quenching capacity revealed by genome-wide association study. Front Plant Sci 2017;8:1773.
[54] Magwa RA, Zhao H, Xing Y. Genome-wide association mapping revealed a diverse genetic basis of seed dormancy across subpopulations in rice (*Oryza sativa* L.). BMC Genet 2016;17:28.
[55] Yang M, Lu K, Zhao FJ, Xie W, Ramakrishna P, Wang G, et al. Genome-wide association studies reveal the genetic basis of ionomic variation in rice. Plant Cell 2018;30:2720–40.
[56] Wang H, Xu X, Vieira FG, Xiao Y, Li Z, Wang J, et al. The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. Mol Plant 2016;9:975–85.
[57] Kersey PJ, Allen JE, Allot A, Barba M, Boddu S, Bolt BJ, et al. Ensembl genomes 2018: an integrated omics infrastructure for non-vertebrate species. Nucleic Acids Res 2018;46:D802–8.
[58] Tello-Ruiz MK, Naithani S, Stein JC, Gupta P, Campbell M, Olson A, et al. Gramene 2018: unifying comparative genomics and pathway resources for plant research. Nucleic Acids Res 2018;46:D1181–9.
[59] Rice Annotation Project. The Rice Annotation Project Database (RAP-DB): 2008 update. Nucleic Acids Res 2008;36:D1028–33.
[60] Sakai H, Lee SS, Tanaka T, Numa H, Kim J, Kawahara Y, et al. Rice Annotation Project Database (RAP-DB): an integrative and interactive database for rice genomics. Plant Cell Physiol 2013;54:e6.
[61] Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, et al. dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 2001;29:308–11.
[62] Yonemaru J, Ebana K, Yano M. HapRice, an SNP haplotype database and a web tool for rice. Plant Cell Physiol 2014;55:e9.
[63] Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res 2015;43:D1023–7.
[64] Mansueto L, Fuentes RR, Borja FN, Detras J, Abriol-Santos JM, Chebotarov D, et al. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res 2017;45:D1075–81.
[65] IC4R Project Consortium. Information commons for rice (IC4R). Nucleic Acids Res 2016;44:D1172–80.
[66] BIG Data Center Members. Database resources of the BIG data center in 2019. Nucleic Acids Res 2019;47:D8–14.
[67] Edwards JD, Baldo AM, Mueller LA. Ricebase: a breeding and genetics platform for rice, integrating individual molecular markers, pedigrees and whole-genome-based data. Database (Oxford) 2016;2016:baw107.
[68] Zhao H, Yao W, Ouyang Y, Yang W, Wang G, Lian X, et al. RiceVarMap: a comprehensive database of rice genomic variations. Nucleic Acids Res 2015;43:D1018–22.
[69] Song S, Tian D, Li C, Tang B, Dong L, Xiao J, et al. Genome variation map: a data repository of genome variations in BIG Data Center. Nucleic Acids Res 2018;46:D944–9.
[70] 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.
[71] Li JY, Wang J, Zeigler RS. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience 2014;3:8.
[72] The 3000 rice genomes project.. The 3,000 rice genomes project. GigaScience 2014;3:7.
[73] Kaur S, Panesar PS, Bera MB, Kaur V. Simple sequence repeat markers in genetic divergence and marker-assisted selection of rice cultivars: a review. Crit Rev Food Sci Nutr 2015;55:41–9.
[74] Miah G, Rafii MY, Ismail MR, Puteh AB, Rahim HA, Islam KhN, et al. A review of microsatellite markers and their applications in rice breeding programs to improve blast disease resistance. Int J Mol Sci 2013;14:22499–528.
[75] Li G, Chern M, Jain R, Martin JA, Schackwitz WS, Jiang L, et al. Genome-wide sequencing of 41 rice (*Oryza sativa* L.) mutated lines reveals diverse mutations induced by fast-neutron irradiation. Mol Plant 2016;9:1078–81.
[76] Meng X, Yu H, Zhang Y, Zhuang F, Song X, Gao S, et al. Construction of a genome-wide mutant library in rice using CRISPR/Cas9. Mol Plant 2017;10:1238–41.
[77] Jung KH, Ronald PC. Transcriptome profiling analysis using rice oligonucleotide microarrays. Methods Mol Biol 2013;956:95–107.
[78] Lu Y, Ye X, Guo R, Huang J, Wang W, Tang J, et al. Genome-wide targeted mutagenesis in rice using the CRISPR/Cas9 system. Mol Plant 2017;10:1242–5.
[79] Shomura A, Izawa T, Ebana K, Ebitani T, Kanegae H, Konishi S, et al. Deletion in a gene associated with grain size increased yields during rice domestication. Nat Genet 2008;40:1023–8.
[80] Weng J, Gu S, Wan X, Gao H, Guo T, Su N, et al. Isolation and initial characterization of *GW5*, a major QTL associated with rice grain width and weight. Cell Res 2008;18:1199–209.
[81] Si L, Chen J, Huang X, Gong H, Luo J, Hou Q, et al. *OsSPL13* controls grain size in cultivated rice. Nat Genet 2016;48:447–56.
[82] Jiao Y, Wang Y, Xue D, Wang J, Yan M, Liu G, et al. Regulation of *OsSPL14* by *OsmiR156* defines ideal plant architecture in rice. Nat Genet 2010;42:541–4.
[83] Miura K, Ikeda M, Matsubara A, Song XJ, Ito M, Asano K, et al. *OsSPL14* promotes panicle branching and higher grain productivity in rice. Nat Genet 2010;42:545–9.
[84] Lu Z, Yu H, Xiong G, Wang J, Jiao Y, Liu G, et al. Genome-wide binding analysis of the transcription activator ideal plant architecture1 reveals a complex network regulating rice plant architecture. Plant Cell 2013;25:3743–59.
[85] Zhang L, Yu H, Ma B, Liu G, Wang J, Wang J, et al. A natural tandem array alleviates epigenetic repression of IPA1 and leads to superior yielding rice. Nat Commun 2017;8:14789.
[86] Li Y, Xiao J, Chen L, Huang X, Cheng Z, Han B, et al. Rice functional genomics research: past decade and future. Mol Plant 2018;11:359–80.
[87] Zhang Y, Massel K, Godwin ID, Gao C. Applications and potential of genome editing in crop improvement. Genome Biol 2018;19:210.