



# Advances in genome-wide association studies of complex traits in rice

Qin Wang<sup>1</sup> · Jiali Tang<sup>1</sup> · Bin Han<sup>2</sup> · Xuehui Huang<sup>1</sup>

Received: 26 July 2019 / Accepted: 5 November 2019 / Published online: 12 November 2019  
© Springer-Verlag GmbH Germany, part of Springer Nature 2019

## Abstract

Genome-wide association studies (GWAS), genetic surveys of the whole genome to detect variants associated with a trait in natural populations, are a powerful approach for dissecting complex traits. This genetic mapping approach has been applied in rice over the last 10 years. During the last decade, GWAS was used to identify the loci underlying tens of rice traits, and several important genes were detected in GWAS and further confirmed in follow-up functional experiments. In this review, we present an overview of the whole process in a typical GWAS, including population design, genotyping, phenotyping and analysis methods. Recent advances in rice GWAS are also provided, including several examples of the functional characterization of candidate genes. The possible breakthroughs of rice GWAS in the next decade are discussed with regard to their application in breeding, the consideration of epistatic interactions and in-depth functional annotations of DNA elements and genetic variants throughout the rice genome.

## Introduction

Genome-wide association studies (GWAS) are a relatively new way for rice geneticists to investigate the genetic architecture of complex traits among diverse varieties and identify the causative loci (or even the causative genes) underlying these traits. GWAS uses statistical methods to search for associations between sequence polymorphisms in the rice genome and phenotypic variation in rice varieties. Compared with conventional biparental populations, GWAS has two advantages: (1) the rice varieties used in GWAS populations contain much more genetic diversity than the two parental lines used in segregation populations, and (2) most GWAS can result in a relatively high mapping resolution due to the existence of numerous historical recombination events (Takeda and Matsuoka 2008). Owing to the advent

of second-generation sequencing (especially the wide applications of the Illumina sequencing platform, Huang et al. 2009; Elshire et al. 2011) and the development of efficient statistical algorithms suitable for plant populations (mainly the mixed model, Yu et al. 2006; Kang et al. 2008; Zhang et al. 2010; Kang et al. 2010; Lippert et al. 2011; Zhou and Stephens 2012; Listgarten et al. 2012; Wang et al. 2014), the GWAS platform was developed in rice several years ago and has since been applied in the genetic dissection of many complex traits (Huang and Han 2014; McCouch et al. 2016). Despite many remaining challenges, tens of GWAS carried out in rice during the last 10 years have enabled the identification of hundreds of associated loci.

In this review, we explain the experiments and critical factors necessary for a successful GWAS, including the generation of a GWAS population, genotyping, phenotyping and a software pipeline. We also summarize the major findings of recent GWAS in rice and the functional evaluations of candidate genes from GWAS. These studies performed over recent years extended our genetic mapping ability and our understanding of the genetic control of many important traits in rice. The perspective for future GWAS, the technical improvements in GWAS follow-up studies and the potential applications in breeding are also discussed.

---

Communicated by Qifa Zhang.

✉ Xuehui Huang  
xhhuang@shnu.edu.cn

<sup>1</sup> Shanghai Key Laboratory of Plant Molecular Sciences, College of Life Sciences, Shanghai Normal University, Shanghai 200234, China

<sup>2</sup> National Center for Gene Research, CAS Center for Excellence of Molecular Plant Sciences, Institute of Plant Physiology and Ecology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China

## Experimental design

The first large-scale GWAS was developed for humans more than 10 years ago (Burton et al. 2007). When applied in medical genetics, GWAS compares genome-wide polymorphisms of people with a disease (e.g., diabetes mellitus) to those of similar people without the disease (healthy samples with similar genetic backgrounds are best). Because the people are first classified by their phenotypes and then added to the GWAS panel according to their phenotypic data, this approach (also known as the case–control design) is called the phenotype-first design. In contrast to the phenotype-first design, the GWAS in plants mostly adopts a genotype-first approach, in which the selection of samples for the GWAS panel is mainly based on their genotypic diversity, without the special consideration of one trait (Nordborg and Weigel 2008; Atwell et al. 2010). For a typical GWAS in crops, the population's genetic data are used for statistical analysis to screen the associations between genotypes and multiple phenotypes, for which the population is usually a germplasm resource covering a large geographic distribution and with diverse phylogenetic relationships. Consequently, population genetics is often involved in the selections of the GWAS panel, and in most cases, the main criteria for the selection are a high level of genetic diversity and a low level of population structure. A high level of genetic diversity means that more loci associated with more phenotypic diversity may be captured through a GWAS, while a low level of population structure results in fewer false positives in the associations. Moreover, seed purification (selfing for 1–2 generations) is a necessary process for generating a successful GWAS population because many accessions in the seed banks are not homozygous.

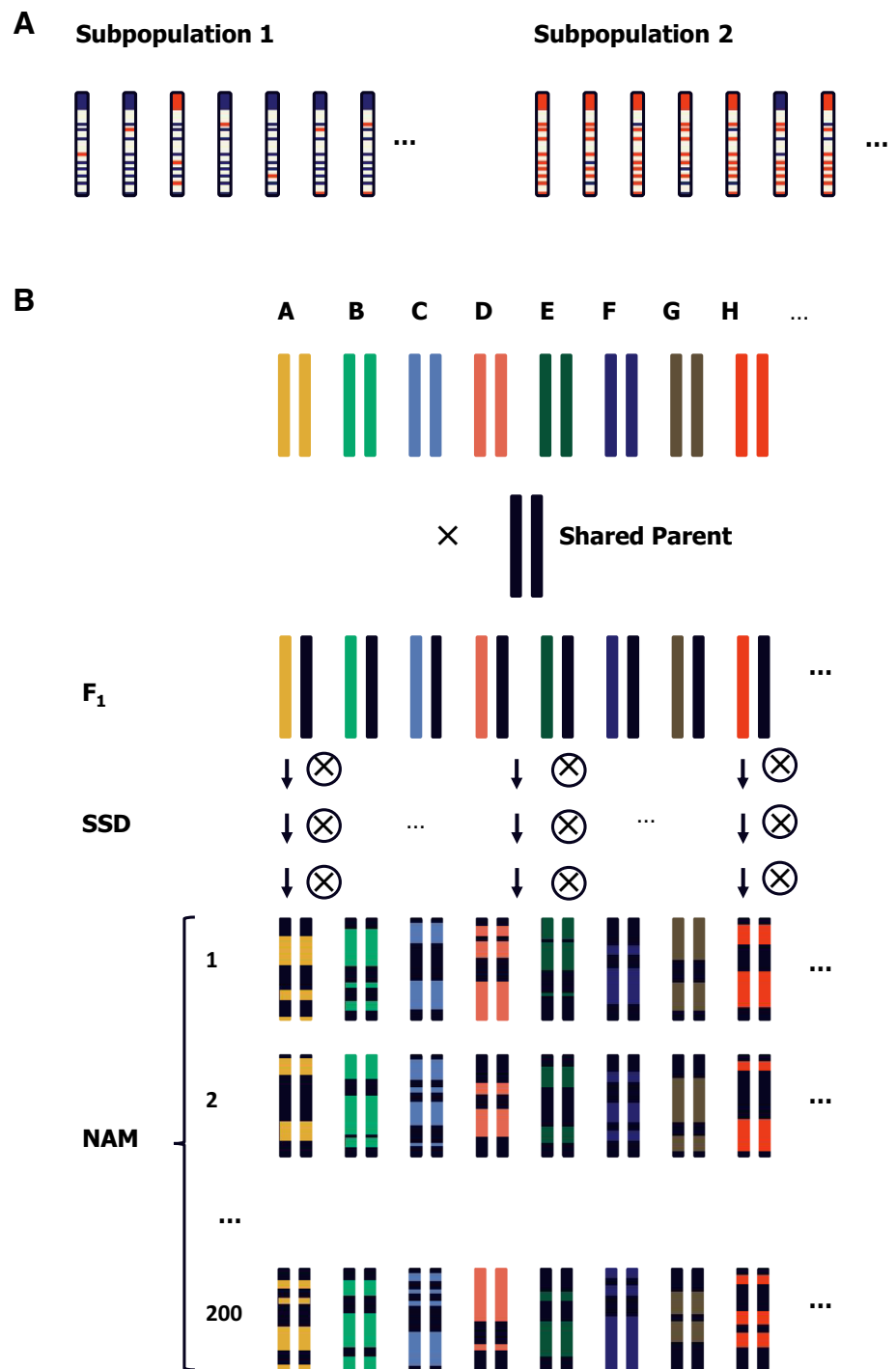
As one of the world's major crops, rice (*Oryza sativa*) has a very broad geographical distribution and is adapted to many ecological and agronomic conditions. Correspondingly, the genetic diversity in rice is very rich. According to an analysis of a pan-genomic dataset in rice (Zhao et al. 2018), each rice gene contains an average of ~16 coding variants in multiple haplotypes and many more variants in noncoding regions (including the promoter regions) with potential effects on the regulation of gene expression. This rich diversity is very beneficial for GWAS. However, as a self-pollinated crop, rice exhibits very strong population structure. The genetic differentiation between the *indica/Xian* rice subspecies and the *japonica/Geng* rice subspecies, using the  $F_{st}$  index as one measure, is very high (~0.55). When further subdivided, cultivated rice may include multiple subpopulations—*indica/Xian* (widely distributed in East Asia), temperate *japonica/Geng* (widely distributed in East Asia), *aus* rice

(mostly in South Asia), tropical *japonica/Geng* (mostly in Southeast Asia and the USA), and Basmati-type rice (mostly in South Asia). To avoid the influence of strong population structure, a GWAS must be performed separately for each subpopulation.

Designs using the existing rice accessions, however, lead to a problem—the nearly fixed alleles or rare alleles within one subpopulation (e.g., 99% “A” allele in one subpopulation and 99% “C” allele in another subpopulation) are difficult to identify in a GWAS (Fig. 1a), even when large sample sizes and statistical corrections for population structure are used (Marouli et al. 2017). One method for addressing these problems is to construct multiple recombinant populations, including nested association mapping (NAM) populations (Fig. 1b), multiparent advanced generation intercross (MAGIC) populations (Bandillo et al. 2013; Dell'Acqua et al. 2015; Huang et al. 2015a; Ogawa et al. 2018) and multiple chromosome segment substitution line (mCSSL) populations. NAM is a new approach for the genetic mapping of complex traits with the advantages of both linkage mapping and association mapping (McMullen et al. 2009). In maize, a total of 25 diverse varieties were crossed with the reference line B73, generating 25 sets of recombinant inbred line (RIL) populations, with ~200 lines for each set. A total of 5000 NAM RILs were combined for genotyping and phenotyping, resulting in high mapping power (Buckler et al. 2009). In these newly developed populations, both the allelic frequencies and the genotypic combinations were completely changed, making the detection of the QTLs with rare alleles within subpopulations or highly differentiated alleles between subpopulations easier. In rice, several sets of MAGIC populations have been developed using diverse accessions (Bandillo et al. 2013). For example, the *indica* MAGIC population was developed by multiple intercrosses among eight diverse *indica* lines, followed by several generations of selfing by single-seed descent (SSD). The *indica* MAGIC population finally comprised 1328 lines, and a GWAS was performed using these lines, which revealed several QTLs, including *Sub1* for submergence tolerance and *Xa4/Xa5* underlying disease resistance.

Another issue in conventional GWAS is how to accommodate multiple alleles of the same gene that are abundant within the rice germplasm population. For example, the waxy gene controlling rice grain starch synthesis contains at least seven alleles with different effects (the ancestral allele; four alleles, each with one missense mutation in the coding region; one 23-bp indel in exon 2 and one mutation in the exon1–intron1 junction site), causing the amylose content to range from 2.2 to 27.8% among diverse rice accessions (Zhang et al. 2019). In one GWAS using ~1000 worldwide accessions, only the variant with the largest phenotypic effect (a 23-bp indel causing a coding frameshift in the waxy gene) was found to be significantly associated with

**Fig. 1** Comparison between natural population using germplasm resources and nested association mapping population by multiple crosses in rice for GWAS. Diverse haplotypes are indicated by different colors



a large change in amylose content (sticky rice and nonsticky rice). In another GWAS using ~ 1500 rice hybrids (nearly all belonging to nonsticky rice), variation in the exon1–intron1 junction was found to be associated with a modest change in amylose content. Possible solutions for dealing with such genetic loci with multiple alleles may include: (1) the improvement of statistical methods used in GWAS (e.g., gene-based associations); (2) performing a GWAS with

different accessions (e.g., wild rice, landraces and modern cultivars) spanning more subpopulations (e.g., *aus*-type rice and *basmati*-type rice); (3) the generation of populations of NAM/MAGIC lines; and (4) phenotyping the GWAS population in multiple environments (a powerful way to mine more genes and a possible way to solve the issue of multiple alleles). These improvements may extend our understanding of the genetic controls of complex traits by multiple alleles.

Sample size (note: accessions with missing genotypes and phenotypes should be excluded) is very important for ensuring enough statistical power in GWAS. The mapping power may be too low when the sample size of one GWAS in rice is too small (e.g.,  $n < 200$ ), and the cost may be too high if a very large number accessions are collected (e.g.,  $n > 5000$ ). Consequently, the sample sizes that are able to yield significant associations in rice GWAS mostly range from 200 to 3000. There are several factors to be considered in determining the sample size. The sample size must be very large for GWAS of complex composite traits, that is, traits controlled by many genes, each with modest genetic effects (e.g., yield per plant), while only a small size is needed for GWAS of qualitative traits controlled by 1–2 major genes (e.g., leaf coloration). For GWAS of complex traits that are difficult to evaluate precisely (e.g., drought tolerance), both the sample size and the replicates within each sample need to be increased. Moreover, when the traits are probably controlled by genes with low-frequency alleles, the experimental design used for the GWAS must be improved based on the sample size and the sample diversity. For example, the frequency of the resistant allele for blast disease (e.g., *Pi2/Pi9*, *Pib*, and *Pigm*) is often very low (typically  $< 5\%$ ), and the GWAS population must be large unless the rice accessions with blast disease-resistant phenotypes are enriched in the collected populations.

## Genotyping of whole-genome variants

Once the rice accessions for a GWAS are available, the genotyping process begins. Rice has a relatively modest extent of linkage disequilibrium, with an average resolution of  $\sim 100$  kb in *indica* and  $\sim 200$  kb in *japonica*, meaning that at least thousands of segregating markers throughout the rice genome are needed. Before the wide application of high-throughput sequencing technology, a single nucleotide polymorphism (SNP) genotyping array is used to perform a GWAS in rice. By using a custom genotyping chip, 44,100 SNP variants were genotyped for 413 diverse rice accessions collected from 82 countries (Zhao et al. 2011). The high genotyping resolution using chips identified numerous common variants underlying 34 complex traits.

Second-generation sequencing is a new approach for high-throughput genotyping in rice GWAS (Huang et al. 2010; Xie et al. 2015; Wang et al. 2016, 2018b). Using this approach,  $\sim 500$  rice accessions were sequenced with  $\sim 1 \times$  coverage (Huang et al. 2010). For each rice accession, the raw sequence reads covered less than half of the rice genome, indicating a large amount of missing genotypic data, which is not suitable for a subsequent GWAS. Due to linkage disequilibrium among polymorphisms of local regions in rice, the missing data could be imputed through

statistical methods (Wang et al. 2018a). The K-nearest-neighbors algorithm was applied to perform the imputations of the sequence-based genotypic data, which worked very well for the rice accessions. During the past 10 years, the throughput of second-generation sequencing has increased by a large amount (from  $\sim 1$  Gb to 1 Tb per run), with the cost decreasing rapidly. At present, the cost for library construction and whole-genome sequencing is below \$30 for each rice accession, which makes the sequencing-based genotyping approach the routine process used in rice GWAS.

Because Asian cultivated rice is a self-pollinated crop, most accessions of rice are inbred lines. The homozygous genotypes greatly facilitate both genotype calling and missing data imputation. However, hybrid rice and natural wild/weedy rice accessions contain a large proportion and a modest proportion of heterozygous genotypes, respectively. The genotype of hybrid rice accessions can be determined precisely by sequencing their inbred parental lines, and wild/weedy rice accessions from natural conditions can be “purified” by several generations of selfing. Without such information or experiments, the processes would be complicated because, even with high coverage (e.g., tenfold), the raw genotypes may not be the real genotypes. For example, for one heterozygous site (e.g., an A/T polymorphism), most likely only one allele (e.g., “A”) is covered by multiple reads, while the other allele (e.g., “A”) is not sequenced. Hence, the uncertainty regarding heterozygous genotypes, coupled with sequencing errors and alignment errors (typically for paralogs), must be corrected with imputation processes. The genotype calling for these heterozygous genomes from hybrid or wild rice (Huang et al. 2012a, 2015b) can be improved with greater sequencing depths and imputed by hidden Markov model-based imputation methods (e.g., that in the Beagle pipeline, Browning and Browning 2009). In particular, when reference haplotype maps are provided (that is, haplotype-based information for the rice population), the performance can be improved.

## High-throughput phenotyping

The agronomic traits that have received much attention in rice molecular genetic studies include grain yield, grain quality and stress resistance, and a high-quality phenotype dataset is crucial in rice GWAS. Phenotyping of these agronomic traits in thousands of rice accessions with many replicates is very laborious and time-consuming and is thus usually performed by several farmers or researchers for several months or even 2–3 years, making it much slower than the genotyping steps. Recently, rapid progress in remote sensing and robotic technologies (e.g., small unmanned aerial vehicles, Reynolds et al. 2019a, b) has made high-throughput phenotyping possible in rice. By capturing relevant images

of rice plants in greenhouses or in the field, software pipelines of image analyses have been developed to estimate various phenotypes for each plant line (Yang et al. 2014). In a recent GWAS of drought resistance in rice, a mechanized phenotyping platform was used to measure the morphological changes before and after drought stresses (Guo et al. 2018). The nondestructive image-based techniques monitored the detailed dynamic response of 507 diverse rice varieties, and 51 image-based traits were extracted from quantitative analyses of objects in binary images, illustrating the power of the mechanized phenotyping methods. The hardware for high-throughput experiments is becoming increasingly cost-efficient and miniaturized, and the image recognition by artificial intelligence algorithms of plant phenotypes is becoming increasingly precise. Hence, it is very likely that more phenotyping studies for rice GWAS will be accomplished with automated platforms requiring very little manual intervention.

Metabolic profiling, gene expression profiling and methylation profiling can be considered complex quantitative traits in a broad sense. There has been an increasing interest over the past few years in investigating the genetic variation in metabolite content, gene expression levels and gene methylation levels among rice accessions through GWAS. For example, metabolic GWAS of rice leaves and grains was performed, resulting in the identification and annotation of many candidate genes involved in metabolic pathways (Chen et al. 2014, 2016). The levels of metabolites were found to be of potential physiological and nutritional importance. A GWAS using the data from gene expression profiling as the traits, called an eGWAS, has been performed in maize (Kremling et al. 2018). With transcriptomic data from seven tissues in 255 maize lines, expression quantitative trait loci (including *trans*- and *cis*-variants) were characterized in the maize genome. In *Arabidopsis thaliana* (Kawakatsu et al. 2016), transcriptome and methylome data were collected from > 1000 accessions. In addition to the genes known to be involved in epigenetics, the GWAS also identified many novel loci controlling DNA methylation levels in *Arabidopsis* genomes. In the future, similar studies in rice will provide important resources and genetic insights into natural variation in gene regulation.

## Analysis methods and significance thresholds for GWAS

Due to the problem of strong population structure, which often leads to many spurious associations, many GWAS methods (e.g., simple linear association and correction by principal component analysis and genomic control in Plink software for human GWAS) do not work well in rice. The linear mixed model approach was developed to account for

multiple levels of relatedness by using a kinship matrix, which greatly improved the performance of GWAS by reducing both the false-positive rate and the false-negative rate (Yu et al. 2006). One of the limitations of the original analysis method using a linear mixed model is the computational burden—it takes a long time or many CPUs to handle a large dataset. The method was improved in terms of computation time by clustering individuals into groups (Zhang et al. 2010), by using the same variance components estimated from the null hypothesis throughout the genome (Kang et al. 2010) and other methods that greatly improve computation speed (such as FaST-LMM, Lippert et al. 2011). When the linear mixed model became capable of handling the computations of thousands of lines genotyped at millions of SNPs, coupled with the imputation from low-coverage sequencing becoming applicable for generating high-quality genotypic data, rice GWAS began to be widely used in the genetic mapping of various traits. Additional gene-based and multi-omics data analyses, which have been developed very fast in the past decade, may further contribute to finding the causative genes and understanding the underlying mechanisms in rice (Chen et al. 2014, 2016; Yano et al. 2016; Si et al. 2016).

The association significance thresholds used in rice GWAS are crucial. In most reports on human GWAS, the genome-wide significance threshold is  $P$  value  $< 5 \times 10^{-8}$  (0.05/1 million markers, The Wellcome Trust Case Control Consortium, 2007), sometimes followed by independent replication in a new dataset. There are no fixed and conventional thresholds for rice GWAS, especially when different populations and different numbers of markers are used. Permutation tests can be used to estimate the GWAS  $P$  value threshold by reshuffling the phenotypic data and performing a GWAS with the reshuffled phenotypes. In most cases of rice GWAS performed with a linear mixed model, the threshold  $10^{-7}$  is reasonable. Moreover, because most rice GWAS uses whole-genome sequencing data and there are enough markers in each linkage disequilibrium block, a single SNP passing the threshold (not multiple SNPs around one locus showing associations) in a Manhattan plot (the plot of  $-\log$  GWAS  $P$  values across the rice genome) is often due to some errors.

We noticed that nearly all analyses used in rice genetics currently involve linear models. In recent years, algorithms, especially deep learning, have been shown to be very powerful in many areas, including image recognition and voice recognition. Some exploratory studies in human genetics assessed the performance of deep learning with linear models in solving genetics-based problems and found that deep learning did not significantly outperform linear models (Bellot et al. 2018). Up to date, few pioneer tests using the new algorithms for GWAS have been reported in plants. We believe that novel methodologies (e.g., deep learning) are

much needed in the genetic modeling of complex traits in rice, but more in-depth research is needed to adapt them.

## Recent GWAS in rice and characterization of candidate genes

Since becoming a considerably mature and verified method (Huang et al. 2012b; Huang and Han 2014), GWAS has been used to investigate the genetic basis of natural variation in biological traits in rice, especially in the last 5 years (Table 1). The traits examined in these GWAS include grain size traits (Duan et al. 2017; Gong et al. 2017), panicle traits (Crowell et al. 2016), callus induction (Zhang et al. 2018), mesocotyl length (Sun et al. 2018), chlorophyll content (Wang et al. 2015), stigma exertion (Zhou et al. 2017), cold tolerance (Xiao et al. 2018), drought tolerance (Guo et al. 2018), and metabolism in rice leaves/

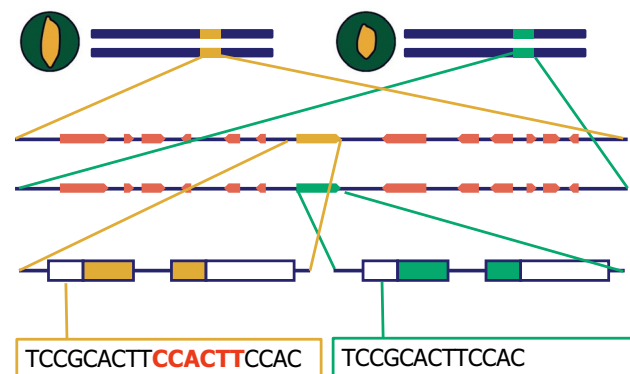
grains (Chen et al. 2014, 2016). In Table 1, we provide a summary of several major GWAS in rice. In addition to *O. sativa* (Asian cultivated rice), a GWAS was also applied in the close relative *O. glaberrima* (African cultivated rice, Meyer et al. 2016). Six salt tolerance-related traits were phenotyped in 93 African rice landraces, and 11 significantly associated loci were identified, including the candidate genes *OsHAK5* and *OsHAK6*. Moreover, the activation of retrotransposons was treated as a phenotype, and its genetic basis was detected. Using genomic data for 3000 rice accessions, 53,262 transposon insertion polymorphisms from 32 retrotransposon families were identified (Carpentier et al. 2019). The underlying genetic factors were examined by looking for associations between each SNP and the number of copies of retrotransposons, and significant association peaks were detected. Nearly all associations overlapped with genomic loci enriched in retrotransposon insertions, suggesting that the presence

**Table 1** List of GWAS in rice during recent few years

Traits	Population size	Associated loci	Genes with functional validations	References
Metabolic features	529 accessions	36	<i>Os02g57760</i> , <i>Os07g32060</i>	Chen et al. (2014)
13 traits (e.g., plant compactness)	529 accessions	141		Yang et al. (2014)
Chlorophyll content	529 accessions	46		Wang et al. (2015)
38 agronomic traits	1495 hybrids	130		Huang et al. (2015b)
Metabolic features	502 accessions	32	<i>Os11g42370</i> , <i>Os11g25454</i> , <i>Os04g11970</i>	Chen et al. (2016)
49 panicle phenotypes	242 accessions	489		Crowell et al. (2016)
Grain length, grain width, amylose content, pericarp color	203 varieties	4		Wang et al. (2016)
Salt tolerance traits	93 landraces	11		Meyer et al. (2016)
7 traits (e.g., awn length)	176 varieties	Not reported	<i>Os01g62780</i> , <i>Os11g08410</i> , <i>Os04g52479</i> , <i>Hd1</i> , <i>Os08g37890</i>	Yano et al. (2016)
Grain length and grain weight	381 varieties	Based on previous GWAS	<i>OsSPL13</i>	Si et al. (2016)
Broad-spectrum blast resistance	67 from a large collection	Not reported	<i>Os03g32230</i>	Li et al. (2017)
Stigma exertion and related floral traits	533 accessions	23		Zhou et al. (2017)
Grain width	102 varieties	3	<i>GSE5</i>	Duan et al. (2017)
Grain length-to-width ratio, chalky grain rate	10,074 F <sub>2</sub> lines	23		Gong et al. (2017)
Flag leaf angle	529 accessions	40 and 32 in environments	<i>OsbHLH153</i> , <i>OsbHLH173</i> , <i>OsbHLH174</i>	Dong et al. (2018)
Drought responses	507 accessions	470	<i>OsPPI5</i>	Guo et al. (2018)
Callus induction traits	510 accessions	21	<i>OsIAA10</i>	Zhang et al. (2018)
Cold stress adaption	641 cultivars	Not a typical GWAS	<i>bZIP73</i>	Liu et al. (2018)
Mesocotyl length	510 accessions	One major locus	<i>OsGSK2</i>	Sun et al. (2018)
Cold tolerance	1033 accessions	13	<i>Os10g34840</i>	Xiao et al. (2018)
Grain length and grain width	270 accessions	9	<i>OsSNB</i>	Ma et al. (2019)
Cadmium accumulation	127 cultivars	12	<i>OsCd1</i>	Yan et al. (2019)
10 traits (e.g., seed setting rate)	100 lines	34		Chen et al. (2019)

of the active copy of retrotransposons is one of the most important causes of transposition activation in rice.

The functions of several important genes within QTLs in rice have been successfully characterized by the GWAS approach (Si et al. 2016; Yano et al. 2016; Li et al. 2017; Dong et al. 2018; Liu et al. 2018; Sun et al. 2018; Ma et al. 2019). Due to the modest rate of linkage disequilibrium decay in rice, the associated loci usually contain several genes (~ 10, on average), which is different from the pattern in maize, which has rapid linkage disequilibrium decay (Li et al. 2013; Navarro et al. 2017). When integrated with other functional evidence (e.g., homolog information, expression profiling, and sequencing variation in genic regions), the GWAS approach often results in a few strong candidate genes for follow-up validation experiments, usually by genetic complementary and genome-editing methods. Using this strategy, one major QTL controlling grain size, *GLW7*, was identified, which encoded the plant-specific transcription factor *OsSPL13* (Si et al. 2016). According to the genetic transformation of different constructs, a tandem repeat in the 5' UTR of *OsSPL13*, which affected the expression level of this gene, was proven to be the causative variant (Fig. 2). Another GWAS identified four new genes underlying awn length, heading date, plant height and panicle number in a *japonica* population (Yano et al. 2016). Candidate gene analysis was performed through in-depth annotations based on the estimated effect of sequencing variation and validated all four genes using transgenic approaches. A GWAS of broad-spectrum resistance to blast disease in 534 sequenced rice accessions detected six associated SNPs in the *cis*-elements of the promoter regions of six genes (Li et al. 2017). Linkage mapping using 3685 recombinant inbred lines indicated that one of them, the C2H2-type transcription factor “Bsr-d1” gene, was tightly linked to the resistance phenotype, which was further confirmed by silencing, overexpression and CRISPR-mediated knock-out of the Bsr-d1 gene. In addition, a GWAS for mesocotyl



**Fig. 2** Illustration of fine mapping in GWAS, using the identification of *OsSPL13* (underlying grain size in rice) as an example

length among 510 diverse accessions identified one candidate gene, *OsGSK2*, a conserved kinase involved in brassinosteroid signaling (Sun et al. 2018). The candidate gene was verified using transgenic rice lines and was also found to be an important gene under domestication selection in rice.

## Application in breeding

Conventional breeding in rice usually requires a long period of time, largely owing to the personal experiences and feelings of the breeders. Recently, it became possible to use molecular breeding strategies to design new rice varieties (Zhang 2007; Zeng et al. 2017; Ouyang 2019). GWAS may be an important and effective approach for rice breeding by guiding the genetic improvement of inbred lines and helping in the design of hybrid crosses. The sequence variants that are detected as being associated with important agronomic traits, especially those for disease resistance and the accumulation of heavy-metal elements, can be designed for use as molecular markers in rice breeding directly. GWAS data are also useful in genomic selection for highly complex traits (e.g., grain yield and combining ability in hybrid rice, Huang et al. 2015b; Chen et al. 2019). However, these applications are not yet widely used in rice breeding because there are still gaps between genomic studies and breeding. One of the gaps is that most breeders, even those with molecular biology training, are not familiar with genomic data. More useful and friendly software systems, similar to mobile apps in smartphones, are needed to fill this gap. Another gap is that the effectiveness of molecular breeding based on GWAS data needs to be further improved, mainly for highly complex traits. For many traits in rice, the gene loci detected from current GWAS designs can explain only a small proportion of the phenotypic variation (much smaller than the heritability of the traits), and more gene loci useful in breeding are not detected. Moreover, there are complex QTL–QTL interactions and G×E interactions in rice, which have not been fully addressed in genetic studies.

In addition to molecular breeding through several generations of crossing and selection, genome-editing technology provides another powerful tool for rapid and effective genetic improvement. GWAS in rice provides many gene targets for breeding (Hickey et al. 2019), and the CRISPR-Cas9 system is an innovation that speeds up selective breeding by allowing several QTL genes to be edited precisely and simultaneously or even novel alleles to be created (Shen et al. 2018; Liu et al. 2019c). Genetic studies in tomato, rice and maize (Krieger et al. 2010; Huang et al. 2016; Liu et al. 2019a) revealed that the overdominance effects of a few genes in hybrid crops (e.g., *IPAI* and *hd3a*) are probably due to dosage effects—the heterozygous state of the genes exhibited ideal expression levels for grain yield

performance. CRISPR-Cas9 editing of the promoter regions of these genes has the potential to create a new allele (with homozygous genotypes) with an effect that is equivalent to the genetic effects of the heterozygous state in hybrid rice (Liu et al. 2019b). Hence, the CRISPR-Cas9 system will play an important role in the era of “GWAS plus” in rice (Fig. 3).

## Conclusions and perspectives

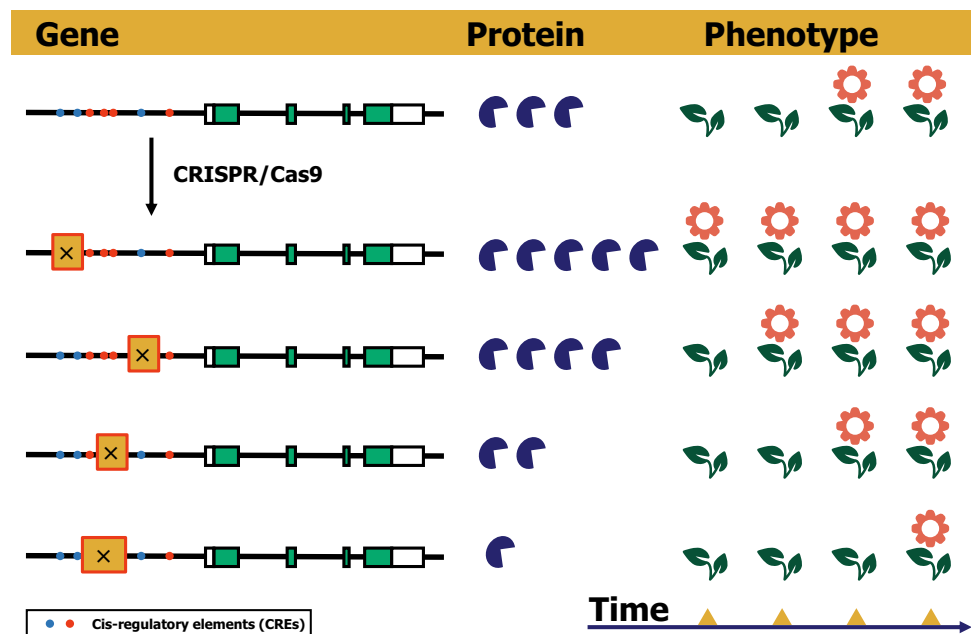
In human genetics, GWAS has laid the foundation for uncovering the molecular basis of complex medical traits. The idea of GWAS was introduced to the plant research community very early, with many preliminary and pioneer studies on genetic diversity, linkage disequilibrium, population structure and association mapping methods using *Arabidopsis*, maize and rice as model species (Caicedo et al. 2007; Zhu et al. 2007; Mather et al. 2007; Clark et al. 2007; Nordborg and Weigel 2008; Zhang et al. 2009; McNally et al. 2009; Tian et al. 2009; Myles et al. 2009). Largely owing to the advent of high-throughput sequencing technology and the development of linear mixed models, GWAS has become a feasible and popular method for the genetic dissection of complex traits in rice and many other plants. In contrast to GWAS of human populations, GWAS of rice has unique advantages. Because rice is a self-pollinated species, in principle, a GWAS population can be genotyped only once but phenotyped multiple times for different kinds of traits and under various environmental conditions, and the genetic architecture of tens of agronomic traits has been investigated

via GWAS and follow-up studies using several major sets of rice collections.

Nearly 10 years have passed since the first attempt to perform a GWAS in rice was published. Despite many successes, GWAS faces new challenges (Zhou and Huang 2019). As stated above, genetic interactions (also called epistatic interactions) and  $G \times E$  interactions are mostly ignored in rice GWAS, although these interactions have been shown to be particularly important for quantitative traits (Manolio et al. 2009; Forsberg et al. 2017). New statistical analysis methods and the corresponding experimental designs for these interactions need to be well addressed in the future. Moreover, to date, most GWAS populations in rice have included temperate japonica and indica rice (accounting for > 80%). The use of exotic subpopulations [e.g., basmati, tropical japonica, aus and wild rice (*O. rufipogon*)] is relatively limited in number. This situation is similar to that in humans, in which most GWAS data are from European populations, while large-scale genomic studies of African, Asian, multiethnic and admixed populations are relatively few in number (Wojcik et al. 2019). In fact, many critical variants and associations present in non-European populations were missed. Based on these advancements in humans, genetic studies on exotic germplasm accessions in rice, including GWAS and subsequent works, need to be enhanced in the future.

With the development of molecular genetic studies, the rice research community has fine-mapped many quantitative trait loci and identified the causative genes at these loci (Xing and Zhang 2010; Zuo and Li 2014). To further enrich our understanding of the underlying molecular mechanisms, not only causative genes but also other causative sequence

**Fig. 3** Creation of new alleles through genome-editing technology for the genes identified from GWAS





variants need to be identified (Fig. 2). One remaining difficulty in rice GWAS is the lack of in-depth annotation of genetic variants. We still have limited knowledge of the potential effects of sequence variants on promoter regions, coding regions and UTRs. Learning from the ENCODE Project, which seeks to interpret the human genome sequence and precisely annotate the DNA elements (Dunham et al. 2012), a “user’s guide” to the rice genome (e.g., including large-scale, well-designed experiments based on data from Chip-Seq, Hi-C and coupled in-depth annotation platforms) is greatly needed, which will provide new insights into rice genetics. We believe that these efforts may also facilitate molecular breeding in rice.

**Acknowledgements** The research activities at our laboratory have been supported mainly by the National Key Research and Development Program of China (2016YFD0100902), the National Natural Science Foundation of China (31825015), Program of Shanghai Academic Research Leader (18XD1402900), Innovation Program of Shanghai Municipal Education Commission (2017-01-07-00-02-E00039) for supporting our research. We apologize to any authors whose work may not have been addressed owing to length restrictions.

**Author contribution** QW, JT, BH and XH conceived, designed and wrote this review manuscript and prepared the figures.

## Compliance with ethical standards

**Conflict of interest** The authors declare that they have no conflict of interest.

## References

- Atwell S, Huang YS, Vilhjalmsón BJ et al (2010) Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. *Nature* 465:627–631
- Bandillo N, Raghavan C, Muyco PA et al (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. *Rice* 6:11
- Bellot P, de los Campos G, Perez-Enciso M (2018) Can deep learning improve genomic prediction of complex human traits? *Genetics* 210:809–819
- Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. *Am J Hum Genet* 84:210–223
- Buckler ES, Holland JB, Bradbury PJ et al (2009) The genetic architecture of maize flowering time. *Science* 325:714–718
- Burton PR, Clayton DG, Cardon LR et al (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447:661–678
- Caicedo AL, Williamson SH, Hernandez RD et al (2007) Genome-wide patterns of nucleotide polymorphism in domesticated rice. *PLoS Genet* 3:1745–1756
- Carpentier MC, Manfroi E, Wei FJ et al (2019) Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat Commun* 10:24
- Chen W, Gao Y, Xie W et al (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat Genet* 46:714–721
- Chen W, Wang W, Peng M et al (2016) Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat Commun* 7:12767
- Chen J, Zhou H, Xie W et al (2019) Genome-wide association analyses reveal the genetic basis of combining ability in rice. *Plant Biotechnol Journal*. <https://doi.org/10.1111/pbi.13134>
- Clark RM, Schweikert G, Toomajian C et al (2007) Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*. *Science* 317:338–342
- Crowell S, Korniliev P, Falcao A et al (2016) Genome-wide association and high-resolution phenotyping link *Oryza sativa* panicle traits to numerous trait-specific QTL clusters. *Nat Commun* 7:10527
- Dell’Acqua M, Gatti DM, Pea G et al (2015) Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in *Zea mays*. *Genome Biol* 16:167
- Dong H, Zhao H, Li S et al (2018) Genome-wide association studies reveal that members of bHLH subfamily 16 share a conserved function in regulating flag leaf angle in rice (*Oryza sativa*). *PLoS Genet* 14:e1007323
- Duan P, Xu J, Zeng D et al (2017) Natural variation in the promoter of GSE5 contributes to grain size diversity in rice. *Mol Plant* 10:685–694
- Dunham I, Kundaje A, Aldred SF et al (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74
- Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* 6:e19379
- Forsberg SK, Bloom JS, Sadhu MJ et al (2017) Accounting for genetic interactions improves modeling of individual quantitative trait phenotypes in yeast. *Nat Genet* 49:497–503
- Gong J, Miao J, Zhao Y et al (2017) Dissecting the genetic basis of grain shape and chalkiness traits in hybrid rice using multiple collaborative populations. *Mol Plant* 10:1353–1356
- Guo Z, Yang W, Chang Y et al (2018) Genome-wide association studies of image traits reveal genetic architecture of drought resistance in rice. *Mol Plant* 11:789–805
- Hickey LT, Hafeez AN, Robinson H et al (2019) Breeding crops to feed 10 billion. *Nat Biotechnol*. <https://doi.org/10.1038/s41587-019-0152-9>
- Huang X, Han B (2014) Natural variations and genome-wide association studies in crop plants. *Annu Rev Plant Biol* 65:531–551
- Huang X, Feng Q, Qian Q et al (2009) High-throughput genotyping by whole-genome resequencing. *Genome Res* 19:1068–1076
- Huang X, Wei X, Sang T et al (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42:961–976
- Huang X, Kurata N, Wei X et al (2012a) A map of rice genome variation reveals the origin of cultivated rice. *Nature* 490:497–502
- Huang X, Zhao Y, Wei X et al (2012b) Genome-wide association study of flowering time and grain yield traits in a worldwide collection of rice germplasm. *Nat Genet* 44:32–39
- Huang BE, Verbyla KL, Verbyla AP et al (2015a) MAGIC populations in crops: current status and future prospects. *Theor Appl Genet* 128:999–1017
- Huang X, Yang S, Gong J et al (2015b) Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. *Nat Commun* 6:6258
- Huang X, Yang S, Gong J et al (2016) Genomic architecture of heterosis for yield traits in rice. *Nature* 537:629–633
- Kang HM, Zaitlen NA, Wad CM et al (2008) Efficient control of population structure in model organism association mapping. *Genetics* 178:1709–1723
- Kang HM, Sul JH, Service SK et al (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42:348–354

- Kawakatsu T, Huang SSC, Jupe F et al (2016) Epigenomic diversity in a global collection of *Arabidopsis thaliana* accessions. *Cell* 166:492–505
- Kremling KAG, Chen SY, Su MH et al (2018) Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature* 555:520–523
- Krieger U, Lippman ZB, Zamir D (2010) The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nat Genet* 42:459–463
- Li H, Peng Z, Yang X et al (2013) Genome-wide association study dissects the genetic architecture of oil biosynthesis in maize kernels. *Nat Genet* 45:43–72
- Li W, Zhu Z, Chern M et al (2017) A natural allele of a transcription factor in rice confers broad-spectrum blast resistance. *Cell* 170:114–126
- Lippert C, Listgarten J, Liu Y et al (2011) FaST linear mixed models for genome-wide association studies. *Nat Methods* 8:833–835
- Listgarten J, Lippert C, Robert C et al (2012) Improved linear mixed models for genome-wide association studies. *Nat Methods* 9:525–526
- Liu C, Ou S, Mao B et al (2018) Early selection of bZIP73 facilitated adaptation of japonica rice to cold climates. *Nat Commun* 9:3302
- Liu H, Wang Q, Chen M et al (2019a) Genome-wide identification and analysis of heterotic loci in three maize hybrids. *Plant Biotechnol J*. <https://doi.org/10.1111/pbi.13186>
- Liu J, Li M, Zhang Q et al (2019b) Exploring the molecular basis of heterosis for plant breeding. *J Integr Plant Biol*. <https://doi.org/10.1111/jipb.12804>
- Liu Q, Wang C, Jiao X et al (2019c) Hi-TOM: a platform for high-throughput tracking of mutations induced by CRISPR/Cas systems. *Sci China Life Sci* 62:1–7
- Ma X, Feng F, Zhang Y et al (2019) A novel rice grain size gene OsSNB was identified by genome-wide association study in natural population. *PLoS Genet* 15:e1008191
- Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. *Nature* 461:747–753
- Marouli E, Graff M, Medina-Gomez C et al (2017) Rare and low-frequency coding variants alter human adult height. *Nature* 542:186–190
- Mather KA, Caicedo AL, Polato NR et al (2007) The extent of linkage disequilibrium in rice (*Oryza sativa* L.). *Genetics* 177:2223–2232
- McCouch SR, Wright MH, Tung CW et al (2016) Open access resources for genome-wide association mapping in rice. *Nat Commun* 7:11346
- McMullen MD, Kresovich S, Villeda HS et al (2009) Genetic properties of the maize nested association mapping population. *Science* 325:737–740
- McNally KL, Childs KL, Bohnert R et al (2009) Genomewide SNP variation reveals relationships among landraces and modern varieties of rice. *Proc Natl Acad Sci USA* 106:12273–12278
- Meyer RS, Choi JY, Sanches M et al (2016) Domestication history and geographical adaptation inferred from a SNP map of African rice. *Nat Genet* 48:1083–1088
- Myles S, Peiffer J, Brown PJ et al (2009) Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202
- Navarro JAR, Wilcox M, Burgueno J et al (2017) A study of allelic diversity underlying flowering-time adaptation in maize landraces. *Nat Genet* 49:476–480
- Nordborg M, Weigel D (2008) Next-generation genetics in plants. *Nature* 456:720–723
- Ogawa D, Nonoue Y, Tsunematsu H et al (2018) Discovery of QTL alleles for grain shape in the Japan-MAGIC rice population using haplotype information. *G3 (Bethesda, MD)* 8:3559–3565
- Ouyang Y (2019) Understanding and breaking down the reproductive barrier between Asian and African cultivated rice: a new start for hybrid rice breeding. *Sci China Life Sci* 62(8):1114–1116
- Reynolds D, Ball J, Bauer A et al (2019a) CropSight: a scalable and open-source information management system for distributed plant phenotyping and IoT-based crop management. *Gigascience*. <https://doi.org/10.1093/gigascience/giz009>
- Reynolds D, Baret F, Welcker C et al (2019b) What is cost-efficient phenotyping? Optimizing costs for different scenarios. *Plant Sci* 282:14–22
- Si L, Chen J, Huang X et al (2016) OsSPL13 controls grain size in cultivated rice. *Nat Genet* 48:447–456
- Shen L, Wang C, Fu Y et al (2018) QTL editing confers opposing yield performance in different rice varieties. *J Integr Plant Biol* 61:122–125
- Sun S, Wang T, Wang L et al (2018) Natural selection of a GSK3 determines rice mesocotyl domestication by coordinating strigolactone and brassinosteroid signaling. *Nat Commun* 9:2523
- Takeda S, Matsuoka M (2008) Genetic approaches to crop improvement: responding to environmental and population changes. *Nat Rev Genet* 9:444–457
- Tian Z, Qian Q, Liu Q et al (2009) Allelic diversities in rice starch biosynthesis lead to a diverse array of rice eating and cooking qualities. *Proc Natl Acad Sci U S A* 106(51):21760–21765
- Wang Q, Tian F, Pan Y et al (2014) A super powerful method for genome wide association study. *PLoS ONE* 9:e107684
- Wang Q, Xie W, Xing H et al (2015) Genetic architecture of natural variation in rice chlorophyll content revealed by a genome-wide association study. *Mol Plant* 8:946–957
- Wang H, Xu X, Vieira FG et al (2016) The power of inbreeding: NGS-based GWAS of rice reveals convergent evolution during rice domestication. *Mol Plant* 9:975–985
- Wang DR, Agosto-Perez FJ, Chebotarov D et al (2018a) An imputation platform to enhance integration of rice genetic resources. *Nat Commun* 9:3519
- Wang W, Mauleon R, Hu Z et al (2018b) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* 557:43–49
- Wojcik GL, Graff M, Nishimura KK et al (2019) Genetic analyses of diverse populations improves discovery for complex traits. *Nature* 570:514–518
- Xiao N, Gao Y, Qian H et al (2018) Identification of genes related to cold tolerance and a functional allele that confers cold tolerance. *Plant Physiol* 177:1108–1123
- Xie W, Wang G, Yuan M et al (2015) Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc Natl Acad Sci USA* 112:5411–5419
- Xing Y, Zhang Q (2010) Genetic and molecular bases of rice yield. *Annu Rev Plant Biol* 61:421–442
- Yan H, Xu W, Xie J et al (2019) Variation of a major facilitator superfamily gene contributes to differential cadmium accumulation between rice subspecies. *Nat Commun* 10:2526
- Yang W, Guo Z, Huang C et al (2014) Combining high-throughput phenotyping and genome-wide association studies to reveal natural genetic variation in rice. *Nat Commun* 5:5087
- Yano K, Yamamoto E, Aya K et al (2016) Genome-wide association study using whole-genome sequencing rapidly identifies new genes influencing agronomic traits in rice. *Nat Genet* 48:927–934
- Yu JM, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38:203–208
- Zeng D, Tian Z, Rao Y et al (2017) Rational design of high-yield and superior-quality rice. *Nat Plants* 3:17031
- Zhang Q (2007) Strategies for developing green super rice. *Proc Natl Acad Sci USA* 104:16402–16409

- Zhang D, Zhang H, Wang M et al (2009) Genetic structure and differentiation of *Oryza sativa* L. in China revealed by microsatellites. *Theor Appl Genet* 119:1105–1117
- Zhang Z, Ersoz E, Lai CQ et al (2010) Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42:355–360
- Zhang Z, Zhao H, Li W et al (2018) Genome-wide association study of callus induction variation to explore the callus formation mechanism of rice. *J Integr Plant Biol*. <https://doi.org/10.1111/jipb.12759>
- Zhang C, Zhu J, Chen S et al (2019) Wx(lv), the ancestral allele of rice waxy gene. *Mol Plant*. <https://doi.org/10.1016/j.molp.2019.05.011>
- Zhao K, Tung CW, Eizenga GC et al (2011) Genome-wide association mapping reveals a rich genetic architecture of complex traits in *Oryza sativa*. *Nat Commun* 2:467
- Zhao Q, Feng Q, Lu H et al (2018) Pan-genome analysis highlights the extent of genomic variation in cultivated and wild rice. *Nat Genet* 50:278–284
- Zhou X, Huang X (2019) Genome-wide association studies in rice: how to solve the low power problems? *Mol Plant* 12:10–12
- Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44:821–824
- Zhou H, Li P, Xie W et al (2017) Genome-wide association analyses reveal the genetic basis of stigma exertion in rice. *Mol Plant* 10:634–644
- Zhu Q, Zheng X, Luo J et al (2007) Multilocus analysis of nucleotide variation of *Oryza sativa* and its wild relatives: severe bottleneck during domestication of rice. *Mol Biol Evol* 24:875–888
- Zuo J, Li J (2014) Molecular dissection of complex agronomic traits of rice: a team effort by Chinese scientists in recent years. *Natl Sci Rev* 1:253–276

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.