Nucleic Acids Research, 2023, **51**, 10924–10933
https://doi.org/10.1093/nar/gkad840
Advance access publication date: 16 October 2023
Data Resources and Analyses

A rice variation map derived from 10 548 rice accessions reveals the importance of rare variants

Tianyi Wang[1,2,3,†], Wenchuang He[1,†], Xiaoxia Li[1,†], Chao Zhang[1,†], Huiying He[1], Qiaoling Yuan[1], Bin Zhang[1], Hong Zhang[1], Yue Leng[1], Hua Wei[1], Qiang Xu[1], Chuanlin Shi[1], Xiangpei Liu[1], Mingliang Guo[1], Xianmeng Wang[1], Wu Chen[1], Zhipeng Zhang[1], Longbo Yang[1], Yang Lv[1], Hongge Qian[1], Bintao Zhang[1], Xiaoman Yu[1], Congcong Liu[1], Xinglan Cao[1], Yan Cui[1], Qianqian Zhang[1], Xiaofan Dai[1], Longbiao Guo[4], Yuexing Wang[4], Yongfeng Zhou[1], Jue Ruan[1], Qian Qian[1,4,5,*] and Lianguang Shang[1,5,*]

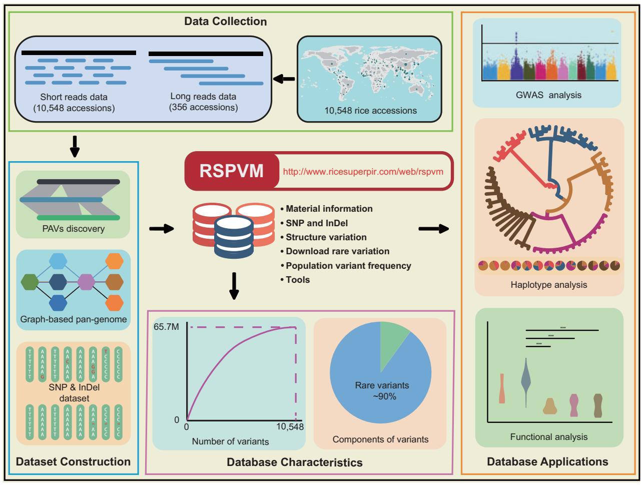
[1]Shenzhen Branch, Guangdong Laboratory of Lingnan Modern Agriculture, Genome Analysis Laboratory of the Ministry of Agriculture and Rural Affairs, Agricultural Genomics Institute at Shenzhen, Chinese Academy of Agricultural Sciences, Shenzhen 518120, China
[2]State Key Laboratory of Crop Stress Adaptation and Improvement, School of Life Sciences, Henan University, Kaifeng 475004, China
[3]Shenzhen Research Institute of Henan university, Shenzhen 518000, China
[4]State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China
[5]Yazhouwan National Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City, Hainan Province 572024, China

[*]To whom correspondence should be addressed. Email: qianqian188@hotmail.com
Correspondence may also be addressed to Lianguang Shang. Email: shanglianguang@caas.cn
[†]These authors contributed equally to this work.

Abstract

Detailed knowledge of the genetic variations in diverse crop populations forms the basis for genetic crop improvement and gene functional studies. In the present study, we analyzed a large rice population with a total of 10 548 accessions to construct a rice super-population variation map (RSPVM), consisting of 54 378 986 single nucleotide polymorphisms, 11 119 947 insertion/deletion mutations and 184 736 presence/absence variations . Assessment of variation detection efficiency for different population sizes revealed a sharp increase of all types of variation as the population size increased and a gradual saturation of that after the population size reached 10 000. Variant frequency analysis indicated that ∼90% of the obtained variants were rare, and would therefore likely be difficult to detect in a relatively small population. Among the rare variants, only 2.7% were predicted to be deleterious. Population structure, genetic diversity and gene functional polymorphism of this large population were evaluated based on different subsets of RSPVM, demonstrating the great potential of RSPVM for use in downstream applications. Our study provides both a rich genetic basis for understanding natural rice variations and a powerful tool for exploiting great potential of rare variants in future rice research, including population genetics and functional genomics.

Graphical abstract



Received: April 20, 2023. **Revised:** September 8, 2023. **Editorial Decision:** September 11, 2023. **Accepted:** September 21, 2023

Introduction

Rice is one of the most important food crops in the world, feeding more than half of the global population (1). Natural genomic variation is an important resource for genetic improvement and modern breeding methods to form new high-yield, high-quality rice varieties. Genomic variation has therefore long been a subject of intensive research. Common genomic variation types include single nucleotide polymorphisms (SNPs), insertion/deletion mutations (InDels) and large structural variations (SVs), all of which contribute extensively to gene functions and phenotypic traits in rice. For example, variations in the coding sequence of *sd1* that alter the amino acid sequence are known to reduce rice plant height to varying degrees (2,3); a 1212-bp deletion can increase rice grain width and weight by regulating *GW5* expression (4,5). Climatic and environmental changes throughout the world give great significance to explorations of natural variations in rice that could be exploited to enhance its ecological adaptability and improve quality and yield.

Use of larger-scale populations can allow a more comprehensive molecular characterization of genomic variations, especially rare variants, than smaller populations. For example, a super-large dataset of 64 000 human exomes demonstrated that human height and weight are largely affected by rare variants (6), which would likely be difficult to detect in a small population. Rare variants in rice have not yet been effectively used. It is therefore important to characterize genetic variations among a large-scale rice population including multiple subpopulations. Several previous rice variation datasets have been generated from thousands of accessions (e.g. 3010, 4726 or 5152 accessions) (7–9). However, those studies primarily focused on simple sequence variants such as SNPs (8); few similar studies have included large-scale SV detection. Although SVs are typically identified from long sequencing reads, current tools support SV identification from large-scale datasets composed of short sequencing reads, which are significantly less expensive to generate. At present, researchers have been generating and accumulating genomic sequencing data for nearly two decades, and these data could now be combined to form a broad rice genomic variation database with a super-large sample size.

In this study, we curated a dataset of both short and long genomic sequencing reads derived from a total of 10 548 cultivated and wild rice accessions; using these data, we constructed the first 10 000-level database of rice variation map (RSPVM) with a sample size of 10 548. The database contained a total of 54 378 986 SNPs, 11 119 947 InDels and 184 736 presence/absence variations (PAVs); 84% of the SNPs and 92% of the InDels both were rare variants, which would be difficult to detect in small-scale populations. Through evaluation of this database, we further demonstrated the great potential of this large variation dataset for studying population structure, genetic diversity, allele distribution and functional diversity in plants.

Materials and methods

Material collection and identification of variation dataset

We collected relevant resequencing data from public database, including NCBI, GSA and ENA (Supplementary Table S1). Quality control of short sequencing reads were conducted by using Trimmomatic (10) (v.0.39 parameter: MINLEN: 75 LEADING: 20 TRAILING: 20 SLIDINGWINDOW: 5:20; MINLEN = 40, while the read length is <75 bp). The reads were mapped to Nipponbare genome (MSU v.7.0) (11) with BWA software (12) (v.0.7.17-r1188) and then were used for SNP calling in Sentieon software (13) (v.sentieon-genomics-202112.02). Genetic variant annotation and functional effect prediction were conducted by using SnpEff (14) (v.4.3t). The long reads of Pacbio and Nanopore from 356 rice accessions were collected (Supplementary Table S2), mapped to the Nipponbare genome (MSU v.7.0) (11) with minimap2 (16) and NGMLR (17) and were further used for SV calling using Sniffles (17) (v.1.0.11, parameters: -l 50 -genotype) and cuteSV (18) (v.1.0.13, parameters: –max_cluster_bias_INS 100 –diff_ratio_merging_INS 0.3 –max_cluster_bias_DEL 200 –diff_ratio_merging_DEL 0.5 -l 50 -L 1000000 –genotype -S for PacBio reads and –max_cluster_bias_INS 100 –diff_ratio_merging_INS 0.3 –max_cluster_bias_DEL 100 –diff_ratio_merging_DEL 0.3 -l 50 -L 1000000 –genotype -S for Nanopore reads). Raw SV results from the two softwares were combined for each accession and further merged to call SVs for the entire population in SURVIVOR software (19) (v.1.0.7, parameters: 1000 1 1 -1 -1 50). The SVs with lengths from 50 bp to 1 Mb were filtered for constructing the graph-based pan-genome by using the vg software (20) (v.1.36.0). PAVs calling were conducted with short sequencing reads and the pan-genome by using vg giraffe (21) and SURVIVOR (v.1.0.7, parameters: 1000 1 1 -1 -1 50). PAVs with low-quality or unexpected length (>1 Mb or <50 bp) were removed. From 1000–10 000 samples, each increase of 1000 samples was a gradient with 50 replicates per gradient set to detect saturation of SNP, InDel and PAV.

Detection of rare and deleterious variants

Allele frequencies for both SNPs and InDels were calculated with VCFtools (15) (v.0.1.16). These variants with MAF <0.01 were defined as rare variants. PCR experiments were conducted to verify seven and eight selected PAVs from short- and long-read datasets, respectively (Supplementary Figure S1, Supplementary Table S3). Primers were showed in supplementary information (Supplementary Table S3). The non-redundant (nr) protein sequence database was downloaded from NCBI (https://ftp.ncbi.nlm.nih.gov/blast/db/FASTA/) and was used for annotating the functional effects of the genomic variations with SIFT4G (22). These variants with SIFT_SCORE <0.05 were annotated as deleterious variations.

Analysis of population structure

To construct a core variant dataset, the VCFtools (15) (v.0.1.16) was used to remove the variant site with a high missing rate (–max-missing 0.9, –maf 0.05) and the remained dataset was further filtered to remove the rice accessions with missing genotypes >20%. Genotype imputation of missing sites and phasing were performed using Beagle (23,24) (v.5.4). The results then were filtered based on linkage disequilibrium (–indep 50 5 2) in plink (25) (v.v1.90b6.26). Phylogenetic trees were constructed using FastTree (26) (v.2.1.11) with default parameters, and were visualized by iTOL (27) (v.6.3.1). Principal component analysis (PCA) was conducted with plink (25) (v.v1.90b6.26). Population structure of the rice accessions was estimated by using ADMIXTURE (28) (v.1.3.0).

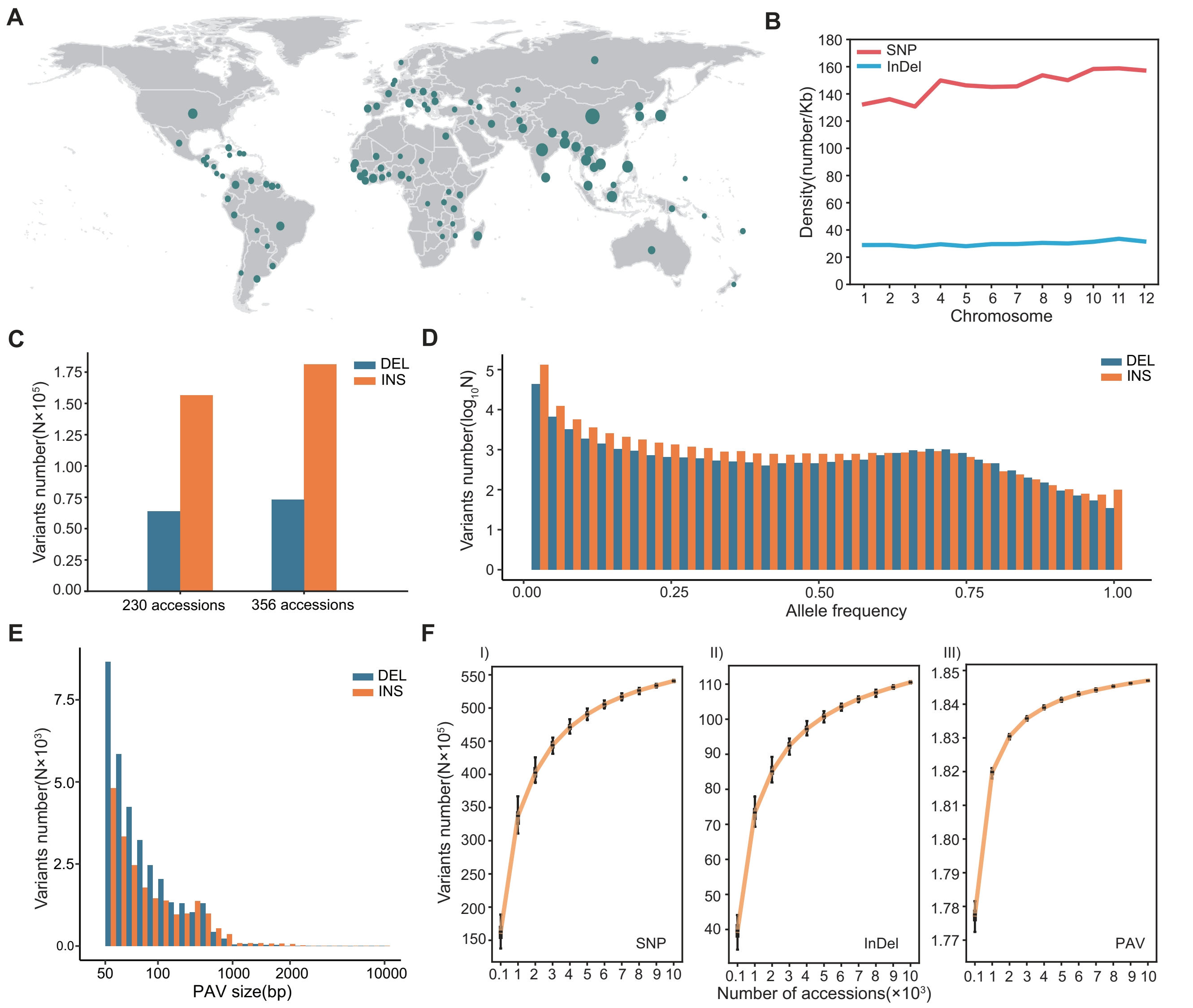


Figure 1. Identification and evaluation of the genomic variants. (**A**) Geographical distribution characteristics of the 10 548 accessions in this study ($n_{min} = 1$, $n_{max} = 4620$). (**B**) Distribution of SNP and InDel on different chromosomes. (**C**) Comparison of deletion (DEL) and insertion (INS) detected in 356 rice accessions in this study and a previously reported population with 230 rice accessions. This comparison presented the additional detected variations caused by a larger population. (**D**) Allele frequency distribution for PAV sites in 356 rice accessions. (**E**) Size distribution characteristics of PAV in 10 548 accessions. (**F**) Saturation curves of different variations in 10 548 accessions.

Haplotype analysis of known functional genes

All mutant loci in the transcript region were used to analyze the haplotypes of functional genes. And the nonsynonymous mutant loci were used to estimate functional haplotypes (alleles) and construct the phylogenetic tree in FastTree (26) (v.2.1.11). The significance of differences between phenotypic traits of different haplotype groups was calculated by using a *t*-test.

Conduction of the Tools module in RSPVM

Four phenotypic datasets were collected from previous studies and denoted as 3kRice (7), BMCPB (29), SCLS-CN-Mix (30) and SCLS-NE-GJ (30), respectively. Those phenotypic data were combined with the genetic variant data in this study to conduct genome-wide association study (GWAS) analysis. The vcftools (15) (v.0.1.16) were used to filter the variants (–max-missing 0.9, –maf 0.05,–min-alleles 2 –max-alleles 2). The first five principal components and matrix of IBS kinship were calculated by using plink (25) (v.v1.90b6.26) (–pca 10) and EMMAX (31) (v.beta-07Mar2010) (emmax-kin -v -h -d 10), respectively, and further used as covariates for GWAS analysis. GWAS was performed using a mixed linear model in EMMAX (31) software (v.beta-07Mar2010). The threshold for GWAS was calculated using the Bonferroni test (0.05/SNPs). The SNPhub package (32) was used to construct the SNP and InDel, Variation map, Haplotype network, Sequence maker, Phylogenetic tree and Visualization of variant frequency sections in RSPVM. The geneHapR pacage (33) was used to construct the ANOVA (analysis of variance) of haplotypes section.

Results

Construction of a large genomic variation dataset from a 10 000-level population

Resequencing data (7,30,34–50) were collected for a total of 10 548 accessions of Asian cultivated rice (*Oryza sativa*) and

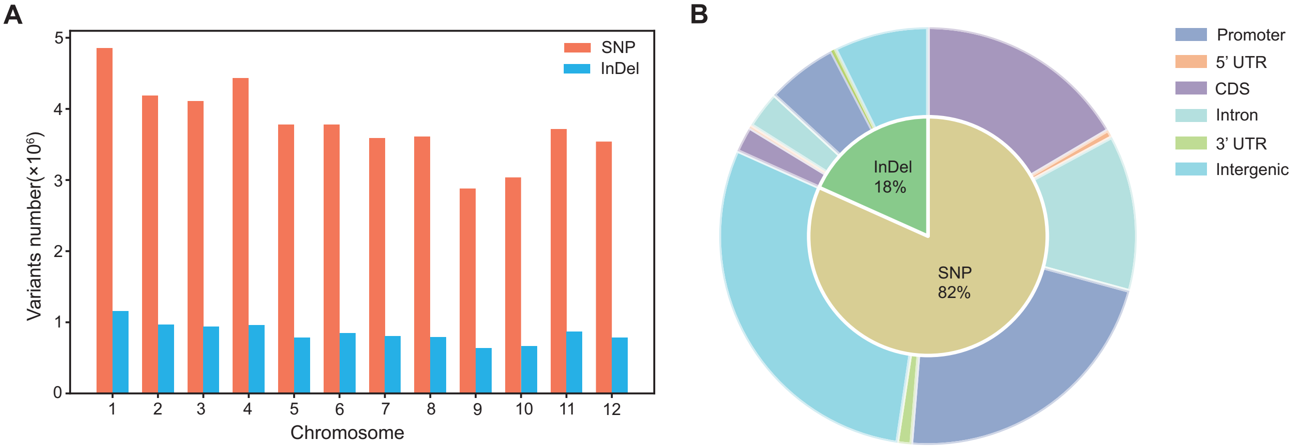


Figure 2. Distribution characteristics of the rare variants. (**A**) Distribution of the rare SNP and InDel on different chromosomes. (**B**) Distribution characteristics of rare variants over different gene structures.

Asian wild rice (*Oryza rufipogon*) from 98 countries in four continents (Figure 1A). These data were used to generate a super-large rice genomic variation dataset. Using Nipponbare (MSU v.7.0) (11) as the reference genome, a total of 54 378 986 SNPs and 11 119 947 InDels were identified among all accessions, with average densities of 146 SNPs/Kb and 30 InDels/Kb. Chromosome 11 showed the highest variation density, with 159 SNPs/Kb and 34 InDels/Kb; chromosome 3 had the lowest variation density, with 131 SNPs/Kb and 28 InDels/Kb (Figure 1B). This indicated a potentially abnormal distribution of genetic diversity between chromosomes. To accurately identify PAVs from the resequencing dataset, a graph-based pan-genome was generated using long sequencing reads from 356 cultivated and wild rice accessions. This dataset constituted a 15.4% increase in the number of genomes compared to the 230 Asian rice accessions included in our previously reported rice pan-genome (49) (Figure 1C). A total of 315 655 SVs, including 254 051 PAVs, were detected in the 356 rice accessions; 94% of the PAVs had a relatively low frequency (<0.05) (Figure 1D). Using the pan-genome as a reference, we here performed PAV detection from short sequencing reads of a 10 000-level population for the first time. From the 10 548 accessions, we identified a total of 184 736 PAVs: 116 371 insertions and 68 365 deletions. The lengths of 58% of the PAVs ranged from 51 bp to 1 kb, and only 0.87% of the PAVs exceeded 10 kb (Figure 1E). These results indicated that shorter PAVs were more readily detected with this method. Saturation of the three types of variants was tested by randomly sampling subsets of the entire variation dataset 50 times. The numbers of SNPs, InDels and PAVs all initially showed rapid increases along with the sample size, then gradually stabilized until the sample size surpassed 10 000 (Figure 1F). The number of identified SNP increased by 61, 22 and 7% when 10 000 samples were used compared to 1000, 3000 and 6000 samples, respectively. Compared to SNPs and InDels, the number of PAVs identified reached saturation at a lower sample size (Supplementary Figure S2). These results strongly demonstrated the necessity and advantages of establishing a super-large-scale variation dataset.

Rare and deleterious variants

The allele frequencies of specific variants were investigated to understand the distribution patterns in a super-large population. Variations with a minor allele frequency (MAF) <0.01 were classified as rare variations, which are expected to be difficult to accurately detect in a small population. A total of 45 509 726 SNPs (84% of the total number identified) and 10 197 265 InDels (92%) were classified as rare variants, with per-chromosome averages of 3 792 477 SNPs (ranging from 2 878 047 to 4 855 139) and 849 772 InDels (ranging from 634 897 to 1 153 153) (Figure 2A). Of these, ∼19% of the rare variations were in a coding sequence, indicating that these rare variants may have genetic and phenotypic functional effects (Figure 2B). A total of 5 758 803 (12.65%) rare variants were non-synonymous mutations locating in 55 343 genes, in which 4011 genes are previously reported to be related to important traits such as plant growth and development, yield related traits and rice quality characteristics, biotic and abiotic stress response (51–54). This indicates that these rare variants could contribute largely to the genetic and phenotypic diversity in rice.

Deleterious variations during crop domestication have cumulative effects that are crucial for understanding potential crop improvement methods (55,56). Predictions of deleterious SNP sites were performed using SIFT4G (22), which revealed a total of 1 486 089 deleterious variant sites. These were unevenly distributed across genes; 3513 genes contained more than 100 deleterious SNPs each, whereas 32 881 genes had fewer than 10 deleterious SNPs each. This suggested gene-specific preferential accumulation of deleterious SNPs. We combined the deleterious variant data with the rare variant data and found that only 2.7% of the rare variants identified here were predicted to be deleterious. These results revealed the powerful advantages of using a super-large dataset for mining both rare and deleterious variants.

Analysis of population structure

To further assess potential applications of the large variation map, we selected 9066 samples as a core collection from the entire population by filtering out samples containing only variants with high missing rate. For the population analysis, variations of the core collection were further filtered by adopting an LD-based SNP pruning procedure to produce a representative dataset consisting of 36 405 variants from the 9066 rice accessions. Based on this representative dataset, common wild rice and Asian cultivated rice accessions were classified as

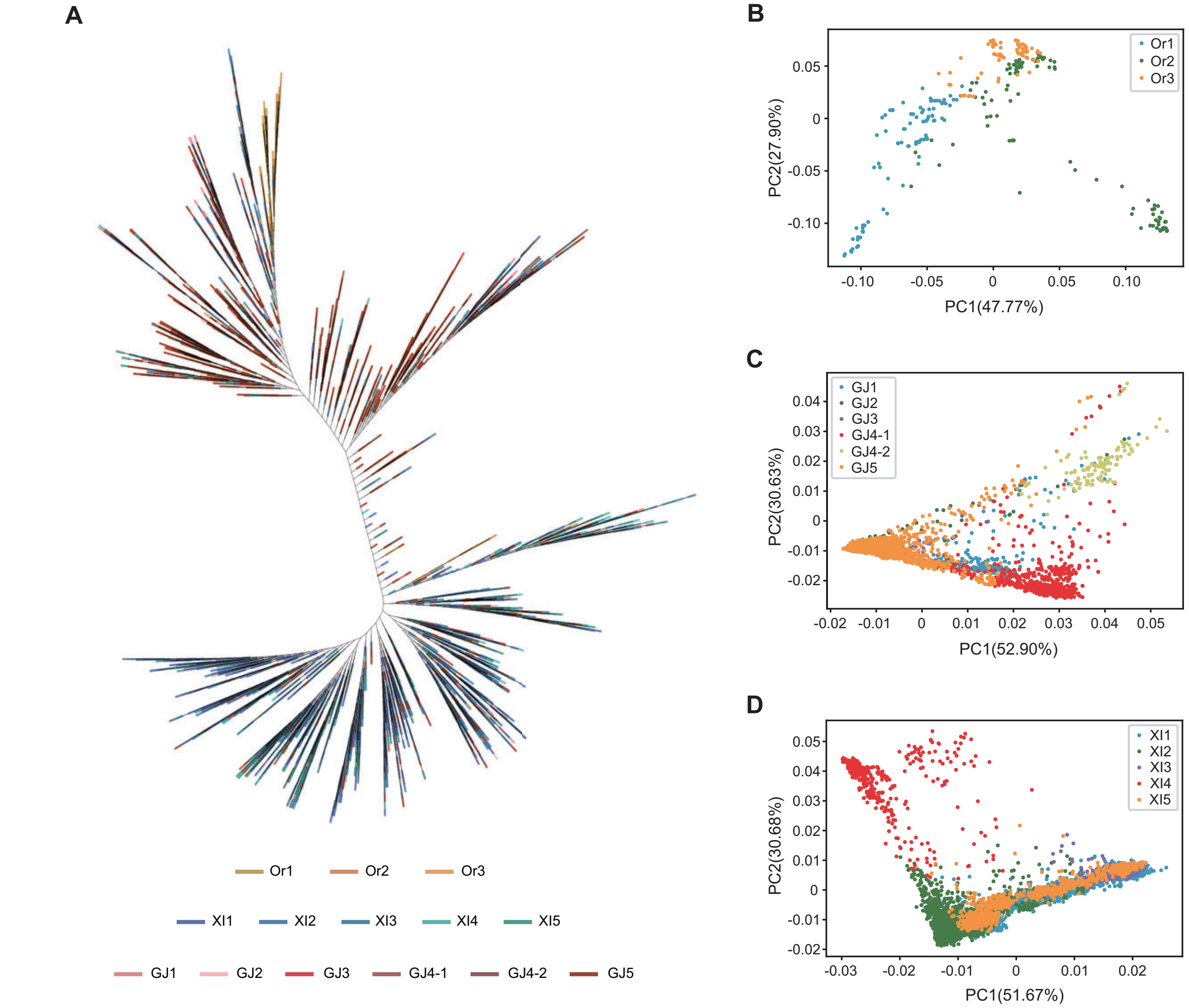


Figure 3. Population analysis of the representative variant dataset containing 9066 rice accessions. (**A**) Phylogenetic tree of the 9066 wild and cultivated rice accessions based on the representative variant dataset. (**B**–**D**) Two-dimensional plotting of the first two principal components based on variations of wild rice (**B**), *japonica* rice (**C**) and *indica* rice (**D**).

two distinct clusters. Common wild rice was divided into three subpopulations, Or1 (containing 246 samples), Or2 (99 samples) and Or3 (67 samples). Asian cultivated rice was divided into 10 subpopulations, namely the five *indica* subpopulations XI1 (1744), XI2 (1796), XI3 (282), XI4 (447) and XI5 (1026) and the six *japonica* subpopulations GJ1 (260), GJ2 (42), GJ3 (99), GJ4-1 (601), GJ4-2 (111) and GJ5 (2246). Most of the accessions in subpopulation XI4 were Aus rice and most of the GJ4-2 members were Basmati rice accessions. The neighbor-joining phylogenetic tree (Figure 3A) and PCA (Figure 3B–D) yielded consistent results; the subpopulations described before were clearly clustered into different clades of the phylogenetic tree and into distinct regions of the PCA map.

Allelic genotypes and associated functional diversity

We further investigated allelic genotypes and associated functional variations in known genes caused by all variations in the core collections. A total of 8223 genes were included in the PAVs of the 9066 wild and cultivated rice accessions. Of these genes, five were selected and analyzed to determine the distinct PAV patterns between subpopulations (Figure 4A). The rice cadmium resistance gene *OsLCD* was present in a 34 708-bp deletion, which caused the widespread absence of this gene in numerous accessions in both the cultivated and wild rice subpopulations. However, it was retained in nearly all accessions in the XI1 and XI3 subpopulations. The rice high-affinity nitrate transporter protein gene *OsNRT2.4* (57), which is an important gene related to nitrogen metabolism, was found to have a rare absence variation only in several accessions in the XI subpopulation. These PAVs of known functional genes provided a valuable basis for further use of diverse rice germplasm.

SNPs were observed on 55 551 genes, and these SNPs generated a large number of haplotypes of functional genes. There was a total of 4522 haplotypes of *GW7*, a major quantitative trait locus controlling grain length and width in rice (58), in the high-quality dataset; the haplotype frequencies ranged from 1 to 700 accessions. To explore possible amino acid changes associated with these variants, nonsynonymous mutation analysis was conducted. This analysis yielded 2155

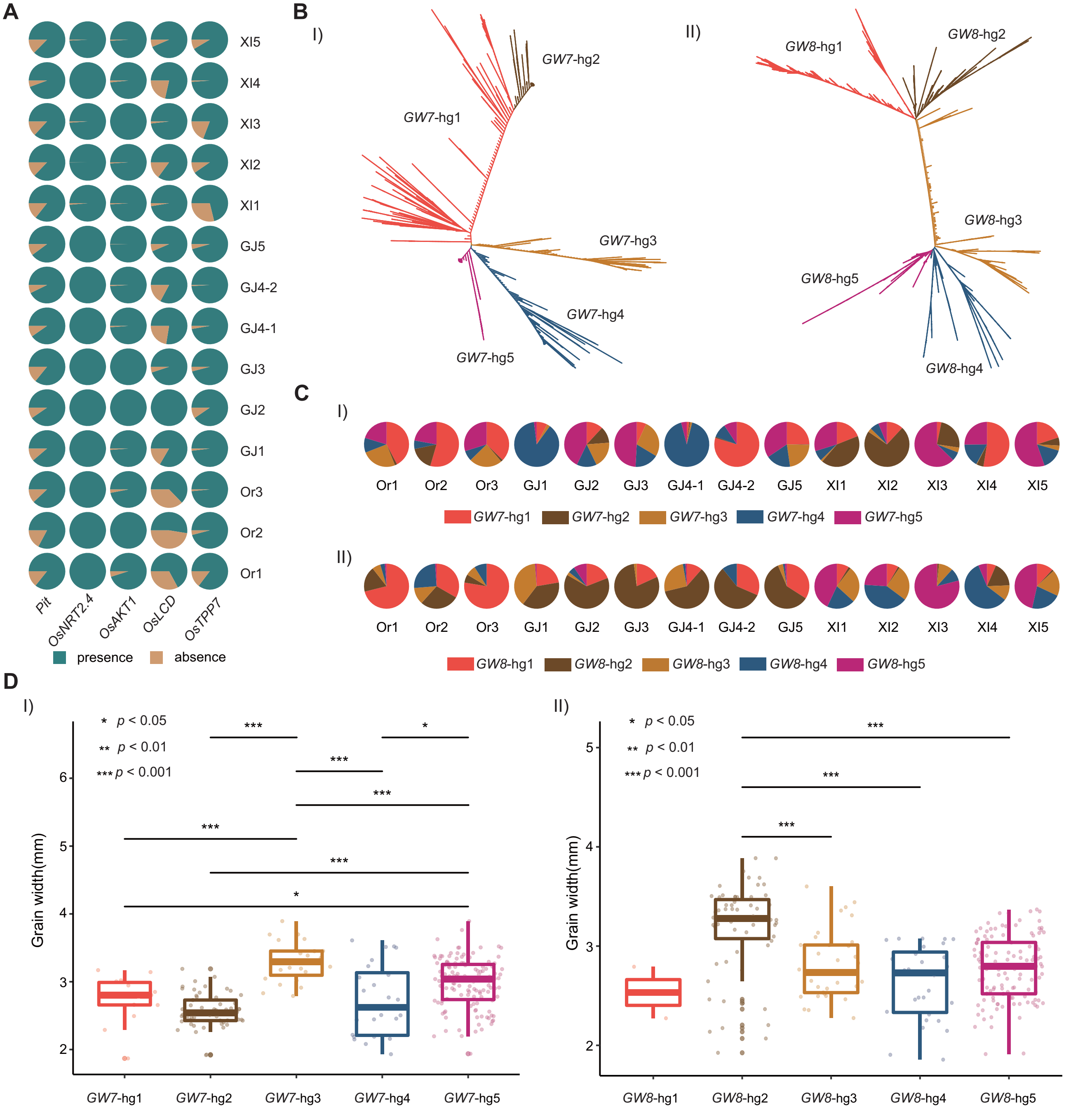


Figure 4. Allele types of some important known genes across different subpopulations. (**A**) Presence and absence of five functional genes in different subpopulations due to large deletions. (**B**) Phylogenetic tree based on functional haplotype sequences of *GW7* (**I**) and *GW8* (II). (**C**) Population frequencies of *GW7* (**I**) and *GW8* (II) haplotype groups in different rice subpopulations. (**D**) The *t*-test analysis of the grain size between different haplotype groups based on 265 rice accessions.

functional haplotypes that could be treated as potential alleles. The functional haplotypes were classified into five haplotype groups (*GW7*-hg1/2/3/4/5) based on neighbor-joining tree (Figure 4B), which showed distinct distributions between subpopulations (Figure 4C). For example, *GW7*-hg3 showed significantly higher grain width than other haplotype groups, and was mainly observed in members of the GJ2, GJ3 and GJ5 subpopulations. In contrast, *GW7*-hg2 was associated with the smallest grain width and was more concentrated in the XI1, XI2 and XI3 subpopulations (Figure 4D). In the entire variation dataset, there were 6923 haplotypes of *GW8*, an important gene that determines grain size, shape and quality (51,59). A total of 3730 functional haplotypes were identified for this gene, which were further classified into five haplotype groups (*GW8*-hg1/2/3/4/5) based on neighbor-joining tree (Figure 4B). These groups were unevenly distributed among 14 rice subpopulations (Figure 4C). Compared with other haplotype groups, *GW8*-hg2 was associated with higher grain width and was primarily distributed among members of XI4 and several GJ subpopulations. *GW8*-hg3, *GW8*-hg4 and *GW8*-hg5 were

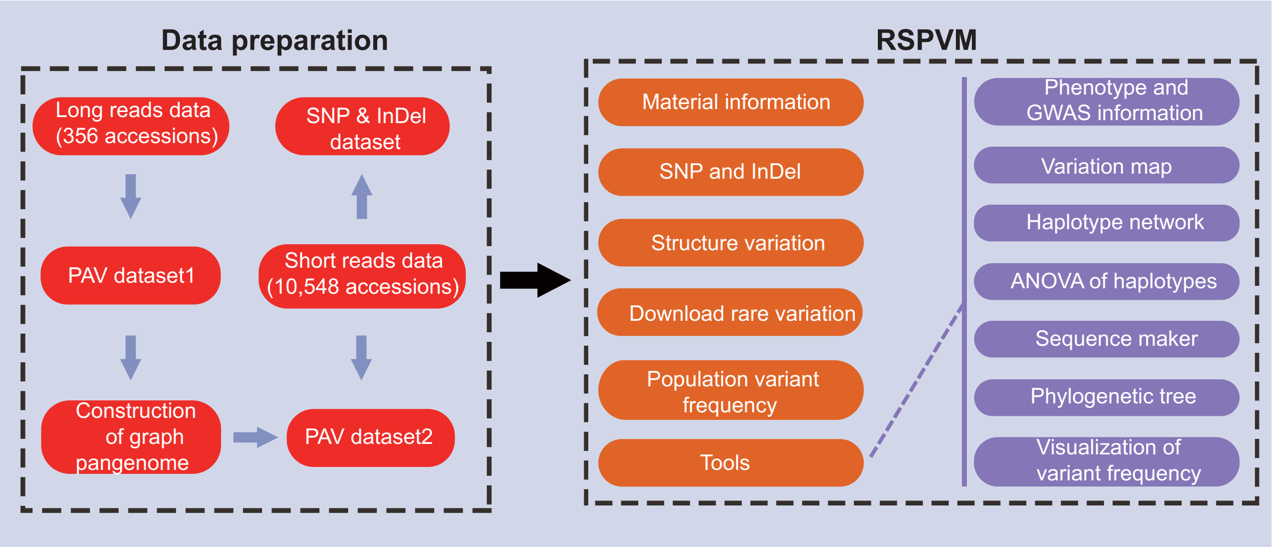


Figure 5. Schematic roadmap of the RSPVM.

associated with relatively low grain width and were mainly observed in XI subpopulations (Figure 4D).

Configuration and usage of the variation database

Using the materials discussed before and the identified genomic variations, we constructed an online rice variation database, RSPVM (http://www.ricesuperpir.com/web/rspvm), which included all of the detected variations, the associated population frequencies, pan-genome sequences and metadata for all samples of the analyzed population (Figure 5). There were six sections of this database. The first was basic information for all samples used in this study, including the sample number, material name, classified population, accession number and data source. The second was a query and view service for SNPs and InDels (including rare variants) based on users' specifications such as chromosome position, gene ID, range of MAF, etc. The third was a query and view service for SVs derived from long reads (356 accessions) and short reads (10 548 accessions). The fourth was the variation frequency for different rice populations. Two different search entries were provided to view and compare variant frequencies of different populations. The fifth was a download service for rare variations divided by chromosomes and variation types. The last contained a series of tools for analyzing variants, which could be summarized as follows: (i) phenotype and GWAS information. This dataset was generated by GWAS analysis based on phenotypic traits of 4790 accessions from previous studies (7,29,30) and the genetic variations in RSPVM. Phenotype values and frequencies, significant trait-associated loci for different traits could be obtained from this tool. (ii) Variation map, visualizing the variants according to the customized groups and accessions, chromosome regions or gene IDs, MAF, etc. (iii) Haplotype network, generating a haplotype network from variants in customized chromosome regions or genes. (iv) ANOVA of haplotypes, conducting an ANOVA analysis for a phenotypic trait between different haplotypes and visualizing them in heatmap and boxplot figures. (v) Sequence maker, generating sequences in FASTA format according to the customized accessions, and chromosome regions or gene IDs. (vi) Phylogenetic tree, generating a neighbor-joining (NJ) tree or multidimensional scaling (MDS) plot based on the user-specified variants. (vii) Visualization of variant frequency, visualizing the population frequency and functional annotation for the user-specified variants.

These applications will enable users to quickly obtain genomic variants for a gene of interest and to rapidly analyze frequency differences between subpopulations. Variations can also be analyzed in one target material compared to other accessions. These resources and functions are valuable tools for research involving population genetics, gene mining and rice breeding.

Discussion

Globally, there are abundant rice germplasm resources with very rich genetic and phenotypic diversity (60). As rice molecular genetics methods have been developed, screening of genomic variations at a population level has become essential for many research areas, e.g. population phylogenetics (61), genomics (62), pan-genomics (48–50), genetic diversity analysis, gene mining, allelic polymorphism analysis (63), and investigations into crop origins and domestication histories (45,64). However, the use of relatively small population sizes inestimably causes omission of rare mutations, resulting in the loss of a large amount of genetic information and biasing results.

To demonstrate potential applications of our dataset, we selected 52 genes with known functions (30) and analyzed the distributions and functions of their functional sites in our dataset. The proportions of functionally validated natural variants in each subpopulation were then analyzed (Supplementary Figure S3, Supplementary Table S4). Most of the results showed identical or similar distribution patterns compared to a previous report using a relatively small population (66 accessions) (44). However, in using a much larger population, we discovered many genes from different subpopulations that were previously reported (44) as absent in the corresponding subpopulations. For example, the nitrate-transporter gene *NRT1.1B* (65) was previously shown to be highly favorable alleles frequency only in *indica* rice and low favorable alleles frequency in *japonica* rice while in our study it also showed favorable alleles frequency in a few accessions of the *japonica* population. Similarly, different patterns were also found for *sd1*, *SCM2* and other genes. These results systematically demonstrated the potential advantages of using 10 000-level data and extensive rare alleles for comprehensively understanding the functional variation of target genes.

We here used resequencing data from 10 548 rice accessions to build a comprehensive super-large variation database, RSPVM, containing more variations (e.g. 54 million SNPs) than a previously reported 3000-level database (29 million) (7), 4700-level database (14 million) (8) and 5000-level databse (18 million) (9). Providing abundance of rare variations, RSPVM is a powerful tool with great potential to enable and enhance many downstream studies. For example, a comprehensive understanding of genomic variations based on a 10 000-level population will yield better insights into genetic structure and diversity, more precise molecular fingerprints for germplasm identification, more functional variations and alleles of target genes for population genetics and functional genomics, and more informative loci and greater potential for whole-genome selection breeding compared to similar analyses using smaller populations. These potential applications reveal the broad prospective uses of our database. Some technical bottlenecks remain that prevent full use of the super-large variation dataset. For example, it remains a challenge to accurately estimate the contributions of rare variations in genome-wide association analyses, and many SVs (e.g. inversions) should be identified with long sequencing reads, which limits the possible number of input sequencing datasets.

Data availability

The long reads data and short reads data useful for this study were obtained from public databases (Supplementary Tables S1 and S2) and 126 genomic sequences were added to the blast panel of RiceSuperPIRdb (http://www.ricesuperpir.com/web/blast/blast1). All variation datasets used in this study could be found at RSPVM (available at http://www.ricesuperpir.com/web/rspvm).

Supplementary data

Supplementary Data are available at NAR Online.

Acknowledgements

Author contributions: L.S.: Conceptualization, Funding acquisition, Project administration, Supervision, Writing-review and editing. Q.Q.: Conceptualization, Funding acquisition, Writing-review and editing. T.W.: Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization, Writing-original draft, Writing–review and editing. W.H.: Data curation, Formal Analysis, Methodology, Validation, Visualization, Writing–original draft, Writing-review and editing. X.L.: Data curation, Formal Analysis, Methodology, Validation, Writing–original draft. C.Z.: Data curation, Formal Analysis, Investigation, Methodology, Validation, Visualization. H.H.: Formal Analysis, Visualization. Q.Y.: Visualization, Validation, Formal Analysis. B.Z.: Visualization, Validation. H.Z.: Formal Analysis, Validation. Y.L.: Formal Analysis, Validation. H.W. Funding acquisition, Validation. Q.X.: Visualization. C.S.: Formal Analysis. X.L.: Formal Analysis. M.G.: Formal Analysis. X.W.: Formal Analysis. W.C.: Formal Analysis. Z.Z.: Formal Analysis. L.Y.: Formal Analysis. Y.L.: Formal Analysis. H.Q.: Formal Analysis. B.Z.: Visualization. X.Y.: Visualization. C.L.: Visualization. X.C.: Visualization. Y.C.: Visualization. Q.Z.: Visualization. X.D.: Visualization. L.G.: Conceptualization. Y.W.: Conceptualization. Y.Z.: Conceptualization. J.R.: Conceptualization.

Funding

This work was supported by National Natural Science Foundation of China [32188102 to Q.Q., 32372148 to L.S.]; Guangdong Basic and Applied Basic Research Foundation [2023B1515020053 to L.S.]; Youth innovation of Chinese Academy of Agricultural Sciences [Y20230C36 to L.S.] and Shenzhen Science and Technology Program [20221008093339096 to H.W.]. Funding for open access charge: National Natural Science Foundation of China [32188102].

Conflict of interest statement

None declared.

References

1. Sasaki,T. and Burr,B. (2000) International Rice Genome Sequencing Project: the effort to completely sequence the rice genome. *Curr. Opin. Plant Biol.*, **3**, 138–141.
2. Monna,L., Kitazawa,N., Yoshino,R., Suzuki,J., Masuda,H., Maehara,Y., Tanji,M., Sato,M., Nasu,S. and Minobe,Y. (2002) Positional cloning of rice semidwarfing gene, sd-1: rice “Green Revolution Gene” encodes a mutant enzyme Involved in gibberellin synthesis. *DNA Res.*, **9**, 11–17.
3. Sasaki,A., Ashikari,M., Ueguchi-Tanaka,M., Itoh,H., Nishimura,A., Swapan,D., Ishiyama,K., Saito,T., Kobayashi,M., Khush,G.S., *et al.* (2002) A mutant gibberellin-synthesis gene in rice. *Nature*, **416**, 701–702.
4. Weng,J., Gu,S., Wan,X., Gao,H., Guo,T., Su,N., Lei,C., Zhang,X., Cheng,Z., Guo,X., *et al.* (2008) Isolation and initial characterization of GW5, a major QTL associated with rice grain width and weight. *Cell Res.*, **18**, 1199–1209.
5. Liu,J., Chen,J., Zheng,X., Wu,F., Lin,Q., Heng,Y., Tian,P., Cheng,Z., Yu,X., Zhou,K., *et al.* (2017) GW5 acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Nat. Plants*, **3**, 17043.
6. Akbari,P., Gilani,A., Sosina,O., Kosmicki,J.A., Khrimian,L., Fang,Y.Y., Persaud,T., Garcia,V., Sun,D., Li,A., *et al.* (2021) Sequencing of 640,000 exomes identifies GPR75 variants associated with protection from obesity. *Science*, **373**, eabf8683.
7. Wang,W., Mauleon,R., Hu,Z., Chebotarov,D., Tai,S., Wu,Z., Li,M., Zheng,T., Fuentes,R.R., Zhang,F., *et al.* (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43.
8. Zhao,H., Li,J., Yang,L., Qin,G., Xia,C., Xu,X., Su,Y., Liu,Y., Ming,L., Chen,L.L., *et al.* (2021) An inferred functional impact map of genetic variants in rice. *Mol. Plant*, **14**, 1584–1599.
9. Yan,J., Zou,D., Li,C., Zhang,Z., Song,S. and Wang,X. (2020) SR4R: an integrative SNP resource for genomic breeding and population research in rice. *Genom. Proteom. Bioinformatics*, **18**, 173–185.
10. Bolger,A.M., Lohse,M. and Usadel,B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.
11. Kawahara,Y., Bastide,M.d.l., Hamilton,J.P., Kanamori,H., McCombie,W.R., Ouyang,S., Schwartz,D.C., Tanaka,T., Wu,J., Zhou,S., *et al.* (2013) Improvement of the *Oryza sativa*

Nipponbare reference genome using next generation sequence and optical map data. *Rice*, **6**, 4.

12. Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.
13. Kendig,K.I., Baheti,S., Bockol,M.A., Drucker,T.M., Hart,S.N., Heldenbrand,J.R., Hernaez,M., Hudson,M.E., Kalmbach,M.T., Klee,E.W., *et al.* (2019) Sentieon DNASeq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.*, **10**, 736.
14. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2014) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, **6**, 80–92.
15. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., *et al.* (2011) The variant call format and VCFtools. *Bioinformatics*, **27**, 2156–2158.
16. Li,H. (2018) Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*, **34**, 3094–3100.
17. Sedlazeck,F.J., Rescheneder,P., Smolka,M., Fang,H., Nattestad,M., von Haeseler,A. and Schatz,M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**, 461.
18. Jiang,T., Liu,Y., Jiang,Y., Li,J., Gao,Y., Cui,Z., Liu,Y., Liu,B. and Wang,Y. (2020) Long-read-based human genomic structural variation detection with cuteSV. *Genome Biol.*, **21**, 189.
19. Jeffares,D.C., Jolly,C., Hoti,M., Speed,D., Shaw,L., Rallis,C., Balloux,F., Dessimoz,C., Bahler,J. and Sedlazeck,F.J. (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.*, **8**, 14061.
20. Hickey,G., Heller,D., Monlong,J., Sibbesen,J.A., Siren,J., Eizenga,J., Dawson,E.T., Garrison,E., Novak,A.M. and Paten,B. (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, **21**, 35.
21. Siren,J., Monlong,J., Chang,X., Novak,A.M., Eizenga,J.M., Markello,C., Sibbesen,J.A., Hickey,G., Chang,P.-C., Carroll,A., *et al.* (2021) Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science*, **374**, 1461.
22. Vaser,R., Adusumalli,S., Leng,S.N., Sikic,M. and Ng,P.C. (2016) SIFT missense predictions for genomes. *Nat. Protoc.*, **11**, 1–9.
23. Browning,B.L., Tian,X., Zhou,Y. and Browning,S.R. (2021) Fast two-stage phasing of large-scale sequence data. *Am. J. Hum. Genet.*, **108**, 1880–1890.
24. Browning,B.L., Zhou,Y. and Browning,S.R. (2018) A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.*, **103**, 338–348.
25. Purcell,S., Neale,B., Todd-Brown,K., Thomas,L., Ferreira,M.A., Bender,D., Maller,J., Sklar,P., de Bakker,P.I., Daly,M.J., *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.*, **81**, 559–575.
26. Price,M.N., Dehal,P.S. and Arkin,A.P. (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.*, **26**, 1641–1650.
27. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
28. Alexander,D.H., Novembre,J. and Lange,K. (2009) Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.*, **19**, 1655–1664.
29. Zhong,H., Liu,S., Sun,T., Kong,W., Deng,X., Peng,Z. and Li,Y. (2021) Multi-locus genome-wide association studies for five yield-related traits in rice. *BMC Plant Biol.*, **21**, 364.
30. Li,X., Chen,Z., Zhang,G., Lu,H., Qin,P., Qi,M., Yu,Y., Jiao,B., Zhao,X., Gao,Q., *et al.* (2020) Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Science China-Life Sci.*, **63**, 1688–1702.
31. Kang,H.M., Sul,J.H., Service,S.K., Zaitlen,N.A., Kong,S.-y., Freimer,N.B., Sabatti,C. and Eskin,E. (2010) Variance component model to account for sample structure in genome-wide association studies. *Nat. Genet.*, **42**, 348–354.
32. Guo,W., Sun,Q., Yao,Y., Peng,H., Xin,M., Hu,Z., Ni,Z., Li,X., Wang,Z. and Wang,W. (2020) SnpHub: an easy-to-set-up web server framework for exploring large-scale genomic variation data in the post-genomic era with applications in wheat. *GigaScience*, **9**, giaa060.
33. Zhang,R., Jia,G. and Diao,X. (2023) geneHapR: an R package for gene haplotypic statistics and visualization. *BMC Bioinf.*, **24**, 199.
34. Chen,W., Gao,Y., Xie,W., Gong,L., Lu,K., Wang,W., Li,Y., Liu,X., Zhang,H., Dong,H., *et al.* (2014) Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.*, **46**, 714–721.
35. Qiu,J., Zhou,Y., Mao,L., Ye,C., Wang,W., Zhang,J., Yu,Y., Fu,F., Wang,Y., Qian,F., *et al.* (2017) Genomic variation associated with local adaptation of weedy rice during de-domestication. *Nat. Commun.*, **8**, 15323.
36. Xia,H., Luo,Z., Xiong,J., Ma,X., Lou,Q., Wei,H., Qiu,J., Yang,H., Liu,G., Fan,L., *et al.* (2019) Bi-directional selection in upland rice leads to its adaptive differentiation from lowland rice in drought resistance and productivity. *Mol. Plant*, **12**, 170–184.
37. Gutaker,R.M., Groen,S.C., Bellis,E.S., Choi,J.Y., Pires,I.S., Bocinsky,R.K., Slayton,E.R., Wilkins,O., Castillo,C.C., Negrao,S., *et al.* (2020) Genomic history and ecology of the geographic spread of rice. *Nat. Plants*, **6**, 492–502.
38. Lv,Q., Li,W., Sun,Z., Ouyang,N., Jing,X., He,Q., Wu,J., Zheng,J., Zheng,J., Tang,S., *et al.* (2020) Resequencing of 1,143 indica rice accessions reveals important genetic variations and different heterosis patterns. *Nat. Commun.*, **11**, 4778.
39. Mao,D., Xin,Y., Tan,Y., Hu,X., Bai,J., Liu,Z.-y., Yu,Y., Li,L., Peng,C., Fan,T., *et al.* (2019) Natural variation in the HAN1 gene confers chilling tolerance in rice and allowed adaptation to a temperate climate. *Proc. Natl Acad. Sci. USA*, **116**, 3494–3501.
40. Xiao,N., Pan,C., Li,Y., Wu,Y., Cai,Y., Lu,Y., Wang,R., Yu,L., Shi,W., Kang,H., *et al.* (2021) Genomic insight into balancing high yield, good quality, and blast resistance of japonica rice. *Genome Biol.*, **22**, 283.
41. Zheng,X., Pang,H., Wang,J., Yao,X., Song,Y., Li,F., Lou,D., Ge,J., Zhao,Z., Qiao,W., *et al.* (2022) Genomic signatures of domestication and adaptation during geographical expansions of rice cultivation. *Plant Biotechnol. J.*, **20**, 16–18.
42. Yano,K., Yamamoto,E., Aya,K., Takeuchi,H., Lo,P.-c., Hu,L., Yamasaki,M., Yoshida,S., Kitano,H., Hirano,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.
43. Wang,X., Wang,W., Tai,S., Li,M., Gao,Q., Hu,Z., Hu,W., Wu,Z., Zhu,X., Xie,J., *et al.* (2022) Selective and comparative genome architecture of Asian cultivated rice (*Oryza sativa* L.) attributed to domestication and modern breeding. *J. Adv. Res.*, **42**, 1–16.
44. Huang,X., Wei,X., Sang,T., Zhao,Q., Feng,Q., Zhao,Y., Li,C., Zhu,C., Lu,T., Zhang,Z., *et al.* (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.*, **42**, 961–967.
45. Huang,X., Kurata,N., Wei,X., Wang,Z.-X., Wang,A., Zhao,Q., Zhao,Y., Liu,K., Lu,H., Li,W., *et al.* (2012) A map of rice genome variation reveals the origin of cultivated rice. *Nature*, **490**, 497.
46. Verma,R.K., Chetia,S.K., Sharma,V., Baishya,S., Sharma,H. and Modi,M.K. (2022) GWAS to spot candidate genes associated with grain quality traits in diverse rice accessions of North East India. *Mol. Biol. Rep.*, **49**, 5365–5377.
47. Higgins,J., Santos,B., Khanh,T.D., Trung,K.H., Duong,T.D., Doai,N.T.P., Hall,A., Dyer,S., Ham,L.H., Caccamo,M., *et al.* (2022) Genomic regions and candidate genes selected during the breeding of rice in Vietnam. *Evolut. Appl.*, **15**, 1141–1161.
48. Qin,P., Lu,H., Du,H., Wang,H., Chen,W., Chen,Z., He,Q., Ou,S., Zhang,H., Li,X., *et al.* (2021) Pan-genome analysis of 33

genetically diverse rice accessions reveals hidden genomic variations. *Cell*, **184**, 3542.

49. Shang,L., Li,X., He,H., Yuan,Q., Song,Y., Wei,Z., Lin,H., Hu,M., Zhao,F., Zhang,C., *et al.* (2022) A super pan-genomic landscape of rice. *Cell Res.*, **32**, 878–896.
50. Zhang,F., Xue,H., Dong,X., Li,M., Zheng,X., Li,Z., Xu,J., Wang,W. and Wei,C. (2022) Long-read sequencing of 111 rice genomes reveals significantly larger pan-genomes. *Genome Res.*, **32**, 853–863.
51. Wang,S., Wu,K., Yuan,Q., Liu,X., Liu,Z., Lin,X., Zeng,R., Zhu,H., Dong,G., Qian,Q., *et al.* (2012) Control of grain size, shape and quality by OsSPL16 in rice. *Nat. Genet.*, **44**, 950–954.
52. Kuroha,T., Nagai,K., Gamuyao,R., Wang,D.R., Furuta,T., Nakamori,M., Kitaoka,T., Adachi,K., Minami,A., Mori,Y., *et al.* (2018) Ethylene-gibberellin signaling underlies adaptation of rice to periodic flooding. *Science*, **361**, 181–186.
53. Wang,J., Zhou,L., Shi,H., Chern,M., Yu,H., Yi,H., He,M., Yin,J., Zhu,X., Li,Y., *et al.* (2018) A single transcription factor promotes both yield and immunity in rice. *Science*, **361**, 1026–1028.
54. Kawano,Y., Akamatsu,A., Hayashi,K., Housen,Y., Okuda,J., Yao,A., Nakashima,A., Takahashi,H., Yoshida,H., Wong,H.L., *et al.* (2010) Activation of a Rac GTPase by the NLR family disease resistance protein Pit plays a critical role in rice innate immunity. *Cell Host Microbe*, **7**, 362–375.
55. Morrell,P.L., Buckler,E.S. and Ross-Ibarra,J. (2011) Crop genomics: advances and applications. *Nat. Rev. Genet.*, **13**, 85–96.
56. Liu,Q., Zhou,Y., Morrell,P.L. and Gaut,B.S. (2017) Deleterious variants in Asian rice and the potential cost of domestication. *Mol. Biol. Evol.*, **34**, 908–924.
57. Li,K., Zhang,S., Tang,S., Zhang,J., Dong,H., Yang,S., Qu,H., Xuan,W., Gu,M. and Xu,G. (2022) The rice transcription factor Nhd1 regulates root growth and nitrogen uptake by activating nitrogen transporters. *Plant Physiol.*, **189**, 1608–1624.
58. Wang,Y., Xiong,G., Hu,J., Jiang,L., Yu,H., Xu,J., Fang,Y., Zeng,L., Xu,E., Xu,J., *et al.* (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.*, **47**, 944–948.
59. Wang,S., Li,S., Liu,Q., Wu,K., Zhang,J., Wang,S., Wang,Y., Chen,X., Zhang,Y., Gao,C., *et al.* (2015) The OsSPL16-GW7 regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Nat. Genet.*, **47**, 949–954.
60. Chen,R., Deng,Y., Ding,Y., Guo,J., Qiu,J., Wang,B., Wang,C., Xie,Y., Zhang,Z., Chen,J., *et al.* (2022) Rice functional genomics: decades' efforts and roads ahead. *Sci. China Life Sci.*, **65**, 33–92.
61. Qiu,J., Jia,L., Wu,D., Weng,X., Chen,L., Sun,J., Chen,M., Mao,L., Jiang,B., Ye,C., *et al.* (2020) Diverse genetic mechanisms underlie worldwide convergent rice feralization. *Genome Biol.*, **21**, 70.
62. Shang,L., He,W., Wang,T., Yang,Y., Xu,Q., Zhao,X., Yang,L., Zhang,H., Li,X., Lv,Y., *et al.* (2023) A complete assembly of the rice Nipponbare reference genome. *Mol. Plant*, **16**, 1232–1236.
63. Wang,Q., Tang,J., Han,B. and Huang,X. (2020) Advances in genome-wide association studies of complex traits in rice. *Theor. Appl. Genet.*, **133**, 1415–1425.
64. Xie,W., Wang,G., Yuan,M., Yao,W., Lyu,K., Zhao,H., Yang,M., Li,P., Zhang,X., Yuan,J., *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**, E5411–E5419.
65. Hu,B., Jiang,Z., Wang,W., Qiu,Y., Zhang,Z., Liu,Y., Li,A., Gao,X., Liu,L., Qian,Y., *et al.* (2019) Nitrate-NRT1.1B-SPX4 cascade integrates nitrogen and phosphorus signalling networks in plants. *Nat. Plants*, **5**, 401–413.

Received: April 20, 2023. **Revised:** September 8, 2023. **Editorial Decision:** September 11, 2023. **Accepted:** September 21, 2023