Differentially directional selection on grain size genes identified by RapMap during rice domestication

Yibo Li (✉ liyibo@mail.hzau.edu.cn)

Huazhong Agricultural University <https://orcid.org/0000-0001-6105-1969>

Juncheng Zhang

Huazhong Agricultural University

Dejian Zhang

Huazhong Agricultural University

Yawei Fan

Huazhong Agricultural University

Cuicui Li

Huazhong Agricultural University

Pengkun Xu

Huazhong Agricultural University

Wei Li

Huazhong Agricultural University

Qi Sun

Huazhong Agricultural University

Xiaodong Huang

Huazhong Agricultural University

Chunyu Zhang

Huazhong Agricultural University

Linyue Wu

Huazhong Agricultural University

Huaizhou Yang

Huazhong Agricultural University

Shiyu Wang

Huazhong Agricultural University

Xiao Su

Huazhong Agricultural University

Xingxing Li

Huazhong Agricultural University

Yingying Song

Huazhong Agricultural University

Meng-en Wu

Huazhong Agricultural University

Article

Keywords: rice, RapMap, grain size genes

Posted Date: November 23rd, 2020

DOI: <https://doi.org/10.21203/rs.3.rs-89471/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Abstract

The ability to identify quantitative trait locus (QTL) genes controlling natural variation of important traits would facilitate the construction of a bridge between gene function analyses and domestication investigations. Grain size in rice is an important yield, quality and domestication trait. The molecular architecture and domestication signatures of grain size at population level remain unknown. Here we first introduce RapMap, a method for rapid and high-throughput QTL mapping employing a series of F_2 gradient populations (F_2 GPs) constructed by trait-adjacent accessions from diverse germplasms. A co-segregation standard is set for simultaneously integrating the three-in-one framework in RapMap: detecting a real QTL, confirming its effect and obtaining its near-isogenic line-like line (NIL-LL), which greatly enhances the efficiency and robustness of QTL identification. Using 15 gradient populations, eight genes including two novel genes (GL1 and *GW5.1*) with different effects for grain size in rice were simultaneously discovered by RapMap in three years, which make more than two thirds of genetic contributions to grain size and shape in a mini-core collection worldwide. In-depth analyses of the eight genes employing a large and geographically diverse population, including 446 wild, 2462 landrace and 784 cultivar accessions, revealed that directional artificial selection of slender- and long- grains and their alleles played an important role during *indica* rice domestication. The genetic effects of the eight grain-size genes positively correlate with the differential selection intensities and the nucleotide variation intensities. Major grain-size genes fixed before the completion of rice domestication have been heavily selected for human requirements and cultivation practices. Generalizable application of the method and mimicking of the domestication nature will accelerate breeding modern cultivars.

Introduction

Tremendous natural variations of food plants and farm animals provide the original genetic basis for their domestication, diversification and subsequent improvement^{1–3}. Agronomically, pathologically and economically important traits with numerous natural variations within a species are controlled by multiple QTLs^{4–6}. Fully understanding of naturally occurring variations of a trait demands the complete identification of the causal genes, which is a long-standing challenge to go from a QTL to its gene^{7–9}. In addition, identifying the functional natural variations of QTL genes (QTGs) contributing to the overall phenotypes in various subtle ways that may not be found through mutant strategies, is an important goal for a good illustration of how basic and applied researches on biology can be synergistic^{4, 6, 9}.

During the last two decades, the isolation of QTGs have been increasing in several plants^{4, 10}. The map-based cloning approach employing enough recombinants has been proved as the most successful and reliable strategy for isolating a QTL gene in crops owing to the tight linkage feature of its functional variation with many other variations in a QTL region^{9–11}. However, the identification of a QTL depending on the complex construction of an advanced generation population, and the construction of its NILs derived from repeated backcrossing and genome selection to dissect one QTL into a single Mendelian factor, which are decisive rate-limiting steps of cloning a QTL, are **time-consuming** and **laborious**^{4, 6, 10}.

Moreover, a very limited number of QTGs reported so far using limited bi-parental combinations could only explain partial phenotypic variations in a population^{9, 12-14}. Genome-wide association studies (GWAS), a strategy complementary to classical bi-parental linkage mapping with the potential to identify multiple loci, mostly requires further analyses with many additional experiments to discover the causal genes, because of massive population structure, the lower power for rare alleles and the lower rate of local linkage disequilibrium decay¹⁵⁻¹⁷. Therefore, conventional strategies of QTL mapping and cloning are not only very complex, inefficient, time- and cost-consuming, but also represent only partial understanding of limited specific materials, ignoring numerous variations in a population.

Grain size or shape with diverse variations in rice is an important target trait of stable yield, grain appearance and milling quality, domestication and breeding, and about ten QTL genes controlling the trait have been identified in past fifteen years¹⁸⁻²⁰. Grain size is determined by its three geometrical dimensions: grain length, width and thickness. G protein signaling, biosynthesis and signaling of BR, IAA and CK, peptide signaling, MAPK signaling, the ubiquitin-proteasome degradation pathway, epigenetic pathways and transcriptional regulation are revealed to be involved in regulation of grain size¹⁹. However, understanding the regulatory mechanisms among natural variations of grain size is still a grey system and becoming an important research field in agricultural science in many subtle ways that may not be found based on analysis of artificial mutants. Much more QTL genes controlling grain size in rice need to be identified for their regulatory mechanism studies based on QTG interactions^{18, 19}.

The enlargement of specific plant organs is the most widely emphasized feature among all the changes arising from domestication²⁰⁻²⁵. Grain size is an important and typical domestication trait in that crops are distinguished from their wild progenitors by increased seed size. With great superiority in deeper seed burial in soil and the easily harvested feature under human agriculture cultivation, bigger grains as reproductive organs could germinate more effectively and provide more nutrition for progenies to produce larger vigorous seedlings with greater fitness²⁴. As edible organs of many cereal crops, larger grains selected by humans could increase grain yield and more food. Archaeological studies in wheat and barley demonstrate that seed size enlargement may arise during the early history of cultivation but before the completion of plant domestication, and that an increase in grain size may be one of the first traits to experience selective pressures associated with human cultivation²⁴. Fully deciphering the molecular bases of variations of a domestication trait is essential for providing a bridge between molecular analyses of gene function and evolutionary investigations of adaptation and natural selection^{3-5, 25}. However, global molecular basis and domestication investigations of grain-size variations at rice population level are still unclear²¹⁻²⁵.

To address these various problems and challenges, there is an urgent need to design novel methods that are fast, powerful and comprehensive enough to simultaneously discover sufficient QTLs controlling a trait in nature in lesser easy-to-perform steps. Here we introduce such a method called RapMap featured by a three-in-one framework integrated and ensured by a co-segregation standard. RapMap is also featured as making a series of gradient populations derived from many trait-adjacent parents in a diverse

germplasm, to easily and greatly meet the quality-control standard, and overcome many limitations of current approaches. We successfully applied RapMap to simultaneously discover eight genes controlling natural variations of grain length and grain width in rice in three years. The domestication signatures of grain-size variations at a population level has been revealed through in-depth analyses of the eight genes employing a large and geographically diverse natural population.

Results

Principles of RapMap

Here we first introduce a method called Rapid Mapping (RapMap) that is fast, powerful and comprehensive enough to simultaneously discover sufficient QTLs controlling a trait in easy-to-perform steps. The principle and method of RapMap (Fig. 1) are explained using rice as the example. To clone a QTL, the first priority is to obtain a mapping population with Mendel's single gene segregation. To most probably achieve this major goal, we first collect a rice mini-core collection with large genetic diversity across the world, phenotype them for a target trait, divide accessions into several gradient groups with similar trait values for each group and minor phenotype differences between adjacent groups, and select gradient parents (from P_1 to P_n) from each gradient group (Left panel of Fig. 1a). Then, using gradient parents, a series of gradient crosses, such as P_1 with P_2 , P_2 with P_3 , P_3 with P_4 , and so on, are made to most probably minimize the number of segregating loci responsible for phenotypic variations of a genetic population and to discover more QTL genes for the trait using all crosses. Then the subsequent F_2 gradient populations (F_2 GPs) are constructed and independently phenotyped for their target traits (Right panel of Fig. 1a). Because F_2 progenies of each F_2 GP are derived from a cross between trait-adjacent bi-parents, most genes responsible for the trait in F_2 are homologous or function-similar alleles in F_1 , and the number of segregating loci responsible for their phenotypic variations is minimal, in most cases one, even though the bi-parents with closely-related phenotypes could be far diverse in other genome regions. Unequivocal phenotype segregation caused by the one locus can be found in F_2 progenies and the corresponding QTL can be easily mapped and cloned. Therefore, RapMap has the simplicity and rapidity of constructing genetic populations over current approaches (such as RIL, DH, NAM and MAGIC²⁶), coupled with the flexible and comprehensive selection of target traits and parents (Fig. 1a). To detect the QTL region of the causal gene for each F_2 GP, then bulked segregant analysis (BSA) is performed using two phenotype-extreme DNA pools by bulking > 30 individuals for each pool from segregating progenies of each gradient population, with two phenotype extremes in the top and bottom 15%²⁷. The SNPs or InDels responsible for target-phenotype variations are homozygous in at least one extreme pool (Fig. 1b). Therefore, by subjecting two DNA pools of each population to SNP chip or sequencing, a set of clustered SNPs or InDels between two pools are very likely to be tightly linked to the causal QTL gene controlling the phenotypic variations (Fig. 1b).

To find a general quality-control standard of confirming whether a gradient population has a Mendel's single-locus segregation pattern, in RapMap, we updated the concept of Mendel's single gene segregation

into 12 single-locus genetic models totally depending on three different genetic effects (Fig. 1c): an allele could be a low-value or high-value complete dominant gene for the phenotypic distribution with a discrete (with no environmental or other effects) or partially overlapping (due to environmental or other effects) boundary and with a ratio of 3:1 or 1:3 (Upper panel of Fig. 1c), or could be a semi-dominant or additive gene for the phenotypic distribution with two discrete or partially overlapping boundaries and with a ratio of 1:2:1 (Middle panel of Fig. 1c), or could be an over-dominant gene for the phenotypic distribution with two discrete or partially overlapping boundaries and with a ratio of 1:1:2 or 2:1:1 (Lower panel of Fig. 1c). Therefore, the common and general nature of the 12 single-locus genetic models is the co-segregation standard that the phenotypes of two homozygous genotypes (AA and aa) can be distinguished (Fig. 1c), regardless of heterozygous genotypes. The traditional judgment in textbooks and literatures that a single locus or multiple loci control is determined by a bimodal (or 3:1) or normal distribution of progeny phenotypes, respectively, should be rationally updated to the co-segregation standard. The co-segregation standard is also the quality-control standard for the simultaneous verification of a candidate QTL and its near isogenic line-like lines (NIL-LLs) in RapMap: if two homozygous genotypes of both flanking molecular markers of a candidate QTL region in progenies from any randomly segregating population can observably differentiate target phenotypes, the region is confirmed as a real QTL for this trait, and the heterozygous lines for the QTL in each gradient population are the corresponding NIL-LLs suitable for fine mapping and cloning of the QTL coupled with various candidate-gene methods (Fig. 1c and d). Therefore, mapping a QTL, verifying its effect and obtaining its NIL-LLs, three decisive rate-limiting steps of cloning a QTL in traditional methods, are a three-in-one step featured in RapMap (Fig. 1a–d), which is integrated and ensured by the co-segregation standard and greatly improves the efficiency and accuracy of QTL identification over traditional methods (Fig. 1e). The genetic population not fitting the co-segregation standard of single-locus genetic models could be replaced by constructing a new gradient population using bi-parents of a similar gradient level, or replaced by a phenotype-well-segregating F_3 or F_4 family derived from a randomly selected low-, medium- or high-value F_2 or F_3 line, respectively, to most probably make minor-effect loci homozygous and one major-effect locus heterozygous through selfing and selecting phenotype-well-segregating F_2 or F_3 progenies (Fig. 1f). To sum up, the co-segregation standard as the necessary and sufficient condition of single-locus genetic models is the cornerstone of RapMap, while the construction of multiple F_2 gradient populations to most probably meet the quality-control standard is the core technique of the method. Thus, the method could be immediately generalizable to any crops that are easy to crossbreed and reproduce.

Identification of four QTL genes for grain-length variations in rice by RapMap, including a novel *GL1* gene

As a proof-of-principle experiment, we applied RapMap to map and isolate QTL genes (QTGs) for natural variations of grain length and width in rice. We have collected, maintained and sequenced a rice mini-core collection of more than 541 accessions, including both landraces and improved varieties from 59 countries, representing the high genetic diversity in this species^{28, 29} (Supplementary Table 1). For grain length, we selected 12 gradient parents, varying from the shortest one of 5.58 mm to the longest one of 13.76 mm, and developed 8 gradient populations using these trait-adjacent parents with minor grain-

length differences within 12 months (Fig. 2a). After measuring grain length of about 200 F_2 progenies of each population and obtaining their DNA, we constructed a high-value bulk (about the top 15%) and a low-value bulk (about the bottom 15%) for each population (Fig. 2b). Two bulked DNA pools for each of the 6 populations (Cross 1–3 and 5–8) were subjected to the commercial rice whole-genome SNP arrays (RICE6K)³⁰. SNPs between two extreme pools detected by the SNP arrays were liable to be tightly linked to the causal QTL gene for grain-length variations (Fig. 2c). Alternatively, following the principle of the SNP-index estimation in QTL-seq analysis³¹, whole-genome sequencing of two DNA bulks from Cross 4 progenies identified a cluster of SNPs with SNP index close to 1 for the locus on the short arm of chromosome 3 (Fig. 2c), but nearby 0.5 for the other loci (Supplementary Fig. 1), and the SNP cluster was most likely near the causal QTG. To confirm a candidate QTL, verify the QTL effect and select its corresponding NIL-LLs suitable for map-based cloning, we first designed 2–4 InDel markers in about 2–4 Mb region covering each candidate QTL region (Fig. 2c), employing the database of rice genomic variations³², and then genotyped all the individuals of each gradient population using the InDel markers, and finally checked the co-segregation standard for every F_2 GP and confirmed two boundaries of each QTL based on the recombinant events (Fig. 2d). If two homozygous genotypes of two flanking molecular markers of each SNP region could observably differentiate grain length of each population at the individual level (Fig. 2d), the SNP region was verified as a real QTL for the grain-length variations, and the heterozygous lines of the QTL region could be the corresponding NIL-LLs. In fact, the Cross 5 population did not fit the co-segregation standard of single-locus genetic models in F_2 (Fig. 1f) and was replaced by a grain length-well-segregating F_4 family derived from a selected medium-value F_3 line (Fig. 2d). Finally, we verified each QTL per gradient population and confirmed the minimum QTL regions on chromosome 3, 3, 3, 3, 7, 1, 7 and 2 for Cross 1 to Cross 8, respectively (Fig. 2e).

For fine mapping of each grain-length QTL identified by each gradient population in RapMap procedures, enough recombinants in progenies of corresponding heterozygous NIL-LLs were screened using the KASP high-through genotyping technique³³ based on two flanking markers and the rapid DNA extraction technique in the greenhouse. Using both high-density marker genotype and grain-length phenotype information of these recombinants, we resolved grain-length QTLs to regions of 20 kb, 18 kb, 14 kb, 54 kb, 24 kb, 19 kb, 56 kb and 3.7 kb for Cross 1 to Cross 8, respectively, in which there is only one to three predicted open reading frames (ORFs) (Fig. 2e). To find out if the known gene in each fine-mapping region is the causal gene, comparative sequencing analyses of their alleles between two parents of corresponding gradient populations were conducted and known functional variations of all three grain-length genes (*GS3*, *GL7* and *GS2*) reported in references^{34–39} were indeed discovered in Cross 1 to Cross 8, except Cross 6 (Fig. 2e). Thus, a transgenic approach was not needed to reassess a known gene underlying the QTL with the same functional variation. Self-pollinating the heterozygous NIL-LLs for each QTL fragment produced two NIL-LLs and their two homozygous genotypes determined by the reported functional marker of each cloned gene could distinguish the long-grain and the short-grain progenies very

well (Fig. 2f), further confirming the three causal genes for grain-length variations of gradient populations.

It should be noted that two novel alleles of the major grain-length gene *GS3*, *GS3-5* from the Cross 1 and 2 population, and *GS3-6* from the Cross 3 population, were identified (Fig. 2 and Supplementary Fig. 2). The *GS3-5* allele has a stronger function (shorter grain) than that of the strongest allele of *GS3*, the *GS3-4* allele from Chuan 7 (Cross 1 of Fig. 2f and Supplementary Fig. 2a)³⁶. The *GS3-6* allele has a stronger function (shorter grain) than that of the *GS3-1* allele (Cross 3 of Fig. 2f and Supplementary Fig. 2a). Compared with the known four alleles of *GS3*, *GS3-5* and *GS3-6* have totally different functional variations with reported ones, with 10 bp and 317 bp deletion in the last exon, respectively, resulting in much shorter frameshift C terminal than that of the *GS3-4* allele (Fig. 2, Supplementary Fig. 2b and c).

A novel grain-length gene *GL1* was discovered in the RapMap programs using the Cross 6 population (Fig. 2). The only one ORF (LOC_Os01g63930) in the 19-kb fine-mapping region (Fig. 2e) was determined as the reliable candidate gene of *GL1*. We sequenced the 2.3-kb promoter region, the coding region and the 3'-UTR of *GL1* in WZ1, 9311 and Zhonghua 11 (ZH11), and found many variations among them (Supplementary Fig. 3a). Compared to ZH11, there is a 2-bp CT insertion in WZ1 and a 751-bp deletion at 3'-UTR in 9311. Compared with NIL(*GL1*), the grains of NIL(*gl1*) were 9.4% longer and 11.3% heavier (Fig. 2g). To further confirm the function of the candidate gene for the *GL1* QTL, we prepared the vectors for gene editing by CRISPR and complementation transformation. Because of the fully dominant feature of the long-grain WZ1 allele as revealed by analyzing the F₂ population (Fig. 2d) and the difficulty in regenerating plants from the *indica* calli of both WZ1 and 9311, we performed complementation transformation by introducing the WZ1 allele into the *japonica* variety ZH11 with short grains. We first knocked out the *GL1* allele in ZH11 background by CRISPR with double target positions, and twenty positive lines of total 26 transgenic plants were generated. However, no significant differences in grain length between negative and positive plants with single or double target positions were found (Fig. 2h, i and Supplementary Fig. 3b, c). Therefore, the *GL1* allele in ZH11 was a non- or weak-functional one, and we transformed the complementation vector contained the coding region of *GL1* in WZ1 driven by its 2.3-kb native promoter into ZH11. Thirty of 36 complementation plants were positive lines, and most of them showed increased grain length and grain weight (Fig. 2h, i, j and Supplementary Fig. 3d). Therefore, these results suggest that only one ORF in the QTL region is the right candidate for *GL1*.

Cloning of four QTL genes for grain-width variations in rice by RapMap, including a new *GW5.1* gene

We also applied RapMap to clone QTGs of rice grain width (Fig. 3). First, we selected 10 gradient parents, varying from the narrowest one of 1.99 mm to the widest one of 3.43 mm, and constructed 7 gradient populations using these gradient parents with adjacent or near width values (Fig. 3a). It should be noted here that if the grain-width difference between the two parents was not very big, the cross would be compatible and suitable for RapMap (Fig. 3a). Similar to the application case of grain length, after RICE6K array or genome sequencing analyses of two bulked DNA pools of each grain-width population (Fig. 3b, c and Supplementary Fig. 4), 7 real QTLs from 7 populations were determined by the co-

segregation standard, which were located on chromosome 8, 8, 8, 7, 5, 5 and 5 for Cross 1 to Cross 7, respectively (Fig. 3d). To clone each grain-width QTL gene, we screened sufficient recombinants for each real QTL. Consequently, using the marker genotypes and the grain-width phenotypes of these recombinants, we resolved the grain-width QTLs to regions of 168 kb, 36 kb, 168 kb, 24 kb, 28 kb, 18 kb and 18 kb for Cross 1 to Cross 7, respectively, which contain only few ORFs (Fig. 3e) for each region after some candidate-gene analyses for two bigger fine-mapping regions. To check whether the known grain-width genes *GW8*, *GW7* and *GW5* within these regions were the causal genes, known functional variations of them^{40–44} were finally discovered by comparative sequencing analyses of their alleles (Fig. 3e). The corresponding NIL-LLs determined by the functional markers of each grain-width gene further confirmed the three causal genes for grain-width variations of these gradient populations (Fig. 3f).

A new grain-width gene *GW5.1* was identified in the RapMap processes using the Cross 5 population (Fig. 3). In the 28-kb fine-mapping region, there are only two ORFs annotated by the RAP-DB website. We identified their full-length cDNAs corresponding to ORF1 and ORF2 (NCBI accessions AK121364 and AK065965, respectively). By searching the public expression database of Zhenshan 97 and Nipponbare (<http://crep.ncpgr.cn/>; <http://ricexpro.dna.affrc.go.jp/>), very low expression of ORF2 was detected in panicles. Moreover, the second exon of the ORF1 allele from Zhenshan 97 or the AA genotype was disrupted by an 11.8-kb LTR-retrotransposon insertion as exhibited by comparative sequencing of the two alleles from the two parents (Fig. 3, Supplementary Fig. 5a), which results in loss-of-function of the ORF1 allele (Fig. 3g). Compared with NIL(*GW5.1*), the grains of NIL(*gw5.1*) were 14.9% wider and 12.9% heavier (Supplementary Fig. 5b). Therefore, ORF1 (LOC_Os05g25350) was determined as the reliable candidate gene of *GW5.1*. To validate the candidate gene, gene editing by CRISPR/Cas9 and complementation transformation strategies were carried out. Three independent frame-shift mutants (Fig. 3h, Supplementary Fig. 5c and d) all produced wider grains than that of ZH11, a *japonica* variety easy in regenerating plants from its calli and with the same functional allele as Iksan438 (Fig. 3i). Consistent with mutant results, the positive plant carrying the complementation vector produced the narrower grains than that of AA (Fig. 3i), suggesting ORF1 is a negative regulator of grain width in rice. Taken together, these data indicate that ORF1 is the right candidate for *GW5.1*.

Genetic contributions of the eight genes to grain size and shape in a rice mini-core collection

To assess the power of RapMap, we investigated the effect size and predictive ability of the eight genes cloned using a single round procedure of RapMap. Using the genotypes of the functional variations of the eight genes as the predictor variables and the phenotypes of 541 accessions, we performed multiple linear regression analyses to assess the genetic contributions of the eight genes to grain size and grain shape (length-width ratio) in the rice mini-core collection (Fig. 4a–f and Supplementary Table 1). The four grain-length genes *GL1*, *GS2*, *GS3* and *GL7*, and four grain-width genes *GW5*, *GW5.1*, *GW7* and *GW8*, totally explained 67.4% and 66.8% phenotypic variations in the rice mini-core collection, respectively (Fig. 4a and b). Among the eight grain-size genes, *GS3* and *GW5* accounted for 62.5% grain-length and 44.3%

grain-width variation, respectively, while other genes showed diverse effects on grain length or width (Fig. 4a and b). Because grain shape is an important quality trait in most rice planting countries, we also investigated the contributions of these eight genes for grain shape (Fig. 4c). Totally 77.2% variations of grain shape could be explained by the eight genes, of which *GW5* and *GS3* explained 47.5% and 55.0% phenotypic variations, respectively (Fig. 4c). These results imply we have dissected most of the genetic variation for grain size and shape, although there are still nearly one third phenotypic effects unrevealed due to other genetic variations that could be identified by selecting other gradient accessions, coupled with environmental effects and large-scale experimental errors. We investigated the prediction performance of the grain-size genes based on their functional variations (Fig. 4d–f). The correlation coefficient between the observed phenotypes and the predicted values was 0.82 for grain length (Fig. 4d), 0.82 for grain width (Fig. 4e) and 0.88 for grain shape (Fig. 4f), showing a very high prediction power. In short, through a single round of RapMap with 15 crosses, more than two thirds of the molecular genetic architecture could be revealed for rice grain size and shape, which displayed a huge potential of RapMap for discovering genes and comprehensively understanding the molecular genetic architecture of a trait in rice.

Directional selection of grain-size genes during rice domestication and improvement

Grain size has been recognized as a typical domestication trait. However, how human domesticated grain size of rice was not clear. Selective signatures from domestication include altered allele frequency and a reduction in nucleotide diversity in domestication loci⁴⁵. Here we first investigated the allele proportion of eight grain-size genes identified by RapMap in a large and geographically diverse natural population⁴⁵, including wild, landrace and cultivar rice of 446, 2462 and 784 accessions, respectively (Supplementary Table 3 and 4). For each of the eight genes, we discovered the stepwise accumulation with different extent of all the long-grain alleles (Fig. 4g) and all the slender-grain alleles (Fig. 4h) from wild to landrace and from landrace to cultivar during rice domestication and improvement (Fig. 4i), indicating the directional and differential selection and enrichment of preexisting genetic variants rather than the acquisition of mutations after domestication for all genes except *GS2*. The extent of the directional selection of these long-grain and slender-grain alleles in *indica* subspecies is much higher than that in *japonica* subspecies, especially for major-effect genes *GS3*, *GW5* and *GL7/GW7* (Supplementary Fig. 6a–e). The average grain length and width in cultivar rice are much longer and slenderer than those of landrace rice, respectively (Fig. 4j). The average grain shape in cultivars is also much slenderer and longer than that in landraces (Fig. 4j). The results were consistent with the significant selection signals of the nucleotide diversity reduction detected for each grain-size locus from wild to landrace and from landrace to cultivar (Fig. 4k). These results demonstrated the enrichment of preferred alleles of all grain-size genes coupled with their nucleotide diversity reduction during rice domestication and improvement.

To understand the phenotype effect of different allele combinations (haplotypes) of eight grain-size genes and guide the molecular design breeding, we identified nine haplotypes for grain length

(Supplementary Fig. 7a) and 16 haplotypes for grain width (Supplementary Fig. 7b) based on the responsible functional variations in the two diverse panels of landrace and cultivar rice. Three haplotypes with the most proportion of cultivars for grain length, Hap-L7 (78%, 9.50 mm), Hap-L8 (83%, 8.98 mm) and Hap-L9 (100%, 9.05 mm), were higher than the average value of 8.56 in landraces (Supplementary Fig. 7a). Three haplotypes with the most proportion of cultivars for grain width, Hap-W14 (35%, 2.80 mm), Hap-W15 (50%, 3.08 mm) and Hap-W16 (50%, 2.78 mm), were lower than the average value of 3.16 in landraces (Supplementary Fig. 7b). The grain shape of all the six haplotypes with the most proportion of cultivars is much higher than the average level in landraces (Supplementary Fig. 7a and b). It should be noted that the haplotypes with most proportion of cultivars predominantly owned *indica* subspecies with much slenderer and longer grains than those of landrace rice, but not the case for *japonica* subspecies (Supplementary Fig. 7c–e). The slender- and long-grain haplotypes and *indica* subspecies are enriched in low latitude regions with warm temperature, while the wide- and short-grain haplotypes and *japonica* subspecies are enriched in high latitude regions with low temperature (Supplementary Fig. 8a–c). The landraces broadly distributing worldwide include most *japonica* rice, while cultivars distributing most in low to moderate latitude regions include most *indica* rice (Supplementary Fig. 8d). These observations suggest that rice grain size shifted toward slenderer and longer grain shape in *indica* subspecies, which might be due to the preferential and unparallel combinatorial selection and differentially geographical distribution of these grain-size alleles between *indica* and *japonica* rice (Supplementary Fig. 6–8). Furthermore, these results provided a basis and flexibility for human to select or design diverse rice grain shapes.

Selection sweeps of grain-size gene loci and correlation among their selection intensities, DNA variations and grain-shape effects

To find the selection evidence of the grain-size genes, we scanned each sweep region which is always generated by the linkage disequilibrium of the selection target and its surrounding loci, and expected to affect the genetic diversity (Fig. 5a–h and Supplementary Fig. 9). By comparing the ratio of the silent-site nucleotide diversity of each gene in a 500-kb genome region of the *GS3* locus between long/short allele accessions in cultivar population (Fig. 5a–h), we discovered a distinct valley of reduced nucleotide diversity for *GS3* which contributes to grain length mostly, spanning ~263 kb, among accessions carrying the long *gs3* allele (Fig. 5a). For *GW5*, the most contributor for grain width, there is a ~307-kb valley with reduced nucleotide diversity among accessions carrying the wide *gw5* allele (Fig. 5e), which has been recognized as a domestication-related gene located at a 4.8Mb selection sweep by comparing the nucleotide diversity of the wild and cultivated rice previously⁴⁵. Similarly, there is a typical valley for *GW7/GL7*, the second important gene for grain width and grain length, spanning a ~229-kb region (Fig. 5b and f). However, there is no evident valley with reduced nucleotide diversity and linkage disequilibrium discovered for other genes with small grain-size effects (Fig. 5c–d, g–h and Fig. 4a–c). To confirm whether this reduced nucleotide diversity across the target region of *GS3*, *GW5* and *GW7/GL7* is an artifact of demographic history, we compared the long-to-short and wide-to-slender π ratios of the target gene loci. The average π ratios of *GS3*, *GW5* and *GW7/GL7* within their corresponding 500-kb regions are

0.14, 0.27 and 0.42, compared to 0.74, 0.85 and 1.73 at the whole chromosome level, respectively; these more than three times' differences are statistically significant ($P < 2.2 \times 10^{-16}$). Thus, artificial and directional selection but not demographic factors played an important role for the reduced nucleotide diversity in these grain-size gene loci during rice domestication. In addition to cultivar, we also investigated the selection sweeps of these gene loci in wild (Supplementary Fig. 9a–h) and landrace rice (Supplementary Fig. 9i–p). We found that, for *GS3*, *GW5*, *GW7*, *GS2* and *GL1*, the genetic diversity decreased gradually with the domestication process from wild to landrace or cultivar, while *GW5.1* and *GW8* did not display clear decreasing (Supplementary Fig. 9 and Fig. 5a–h). The selection sweeps of major genes *GS3* and *GW5* have been shaped during the landrace period, indicating the early domestication has fixed their selection signatures (Supplementary Fig. 9a, i, e, m and Fig. 5a, e). These results together showed the differential extent of selection on the eight genes is likely to be related with the phenotype effect size in the large diverse population.

To further understand the relationship among the size of phenotype effect, the extent of artificial selection and nucleotide variations during grain-size domestication and improvement, we analyzed the pairwise correlation for the eight grain-size genes (Fig. 5i and j). The results showed the strong positive correlation among nucleotide variations, selection intensities, grain-shape effects and phenotype association significance, and this correlation is much higher through selection than that not through selection (Fig. 5i and j). The genes (such as *GW5*, *GS3* and *GW7/GL7*) under stronger artificial selection due to their location in stronger selection sweep regions (Fig. 5a–h) had more genetic contribution to the whole grain-shape variations (Fig. 5i), usually with a large amount of natural variations between wild and cultivar rice. Although *GS2* has a huge effect for grain length in bi-parental linkage populations (Fig. 2)^{38, 39}, it has little contribution to the whole phenotype variations in the large natural population due to its fewest nucleotide variations and weak selection (Fig. 5i and Fig. 4g). However, the favored natural variation of *GS2*, already present in cultivars, also had been identified by RapMap. Thus, selection plays an essential role in increasing the correlation between DNA variations and grain-size phenotypes.

Discussion

We establish a new strategy and a new quality-control standard of QTL mapping and cloning in RapMap, which can quickly identify QTLs and isolate QTGs responsible for a given phenotype on a large scale (Fig. 1). RapMap has several notable advantages compared with traditional schemes of linkage and GWAS analysis, and recently reported methods for mapping and cloning functional genes (QTL-seq, QTG-seq, MutMap and others)^{31, 46–48}. (1) F_2 gradient populations constructed from trait-adjacent parents from diverse germplasms are technically central and innovative in RapMap (Fig. 1a). Due to large trait differences between two crossed lines usually used for construction of traditional genetic populations, many genes responsible for the phenotype between trait-distant bi-parents most likely differ and the number of segregating loci responsible for phenotype variations of a genetic population is maximal^{4–7}. So, segregation of a phenotype in such genetic populations often follows a complex and continuous pattern of phenotype variations determined by the combined effects of multiple genes. In addition, half of

all genes in F_2 are heterozygous, which strongly affects a phenotype to be fully observed. Therefore, complex advanced generation genetic populations in traditional linkage analysis, such as RIL, DH, BC, NAM or MAGIC²⁶, are required to separately and fully address the phenotype contribution of each individual gene by making all the loci homologous in each line. However, it takes a lot of time, huge efforts and costs to develop each kind of advanced generation populations before identifying a real QTL (Fig.1e). Moreover, the construction of NILs derived from many times' consecutive backcrosses takes further more years before starting the fine mapping and cloning process (Fig.1e). Thus, the gradient population used in RapMap, only requiring one cross between two trait-adjacent parents followed by one subsequent self-pollination, sidesteps the complex construction of traditional advanced generation genetic populations, and makes a detected QTL easily and most possibly meet one of the 12 single-locus genetic models (Fig.1c). (2) RapMap is a three-in-one framework of rapidly and simultaneously mapping multiple QTLs, verifying their real effects and obtaining their NIL-LLs integrated and ensured by our proposed co-segregation standard (Fig.1a–c), greatly and straightforwardly overcoming the rate-limiting steps of mapping a QTL commonly faced in other mapping methods. (3) We discovered the necessary and sufficient condition of single-locus genetic models and also the quality-control standard for the three-in-one framework featured in RapMap: the co-segregation standard that two homozygous genotypes of a QTL in progenies of any segregating population can distinguish target phenotypes (Fig.1c). The co-segregation standard of benchmarking QTL mapping is the cornerstone of RapMap, while the construction of multiple F_2 gradient populations to meet the quality-control standard is its technique strategy. (4) The co-segregation standard coupled with multiple gradient populations ensures RapMap can be generalizable to crops that are easy to crossbreed and reproduce. (5) As a high-throughput mapping and cloning solution, RapMap holds a tremendous promise for unleashing the full molecular genetic architecture of important traits in crops employing a series of F_2 gradient populations from diverse germplasms, compared with other methods that are based on a single and specific bi-parental population and usually provide limited pieces in the complex jigsaw of a trait^{7–14}. RapMap integrates both merits of the higher reliability of linkage analysis and the higher throughput of GWAS in mapping a QTL, with great focus on the flexible and comprehensive selection of target traits and parents and on the simplicity and rapidity of constructing their genetic populations, allowing us rapid and systematic identification of as many as possible QTGs for a trait utilizing all gradient populations derived from representative resources. (6) Unlike other methods, RapMap is a solid phenotype- and genotype-based but not a software- or statistics-dependent method that may introduce artifacts. Moreover, the power of BSA could be improved by increasing the number and the phenotype precision of extreme individuals, and by time-saving genotyping strategies. We believe the efficiency of RapMap can be greatly improved and maximized by optimizing these advantages or key points.

In this study, eight grain-size genes, including six known genes (*GS3*, *GL7*, *GS2*, *GW5*, *GW8* and *GW7*), two new *GS3* alleles (*GS3-5* and *GS3-6*) and two novel genes (*GL1* and *GW5.1*), not only with major effects but also with minor effects on grain size in a natural germplasm, could be unbiasedly and simultaneously identified by a distinctive RapMap pipeline in 3 years, instead of 8–10 years under normal programs (Fig.2–4). The eight grain-size genes identified by RapMap make more than two thirds of genetic

contributions to grain size and shape in a rice mini-core population worldwide (Fig.4). Fine-mapping in map-based cloning programs using recombinants in a QTL is the most efficient method to identify a gene underlying the QTL^{14–16}. But sometimes, it is very difficult to narrow down a QTL to a very small chromosomal region with few genes, due to the QTL effect, population size and recombination frequency, such as the cases of some grain-width QTLs in this study (Fig.3). To cope with these problems, many candidate gene strategies could be used in a coordinated manner with fine mapping to accelerate the discovery of the functional gene in a large fine-mapping QTL region: finding functional variations (InDels, stop-codons or other large variations resulting in changes of RNA level or protein functions) by comparative sequencing between bi-parents as a reliable feature of uncovering a causal gene, checking homologous genes with similar functions in other species or function-related pathways, discovering differential expression genes between two NILs, checking target tissue-specific or -abundant expression genes or excluding non-expression genes in target tissues, or probing candidate genes using mutants, cDNAs or ESTs as important resources. Complementation transformation of a functional or strong allele driven by its native promoter to a parent or NIL with a null or weak allele is the standard bar to declare isolation of a QTG controlling a natural variation trait, just like the function confirmation of two novel genes (*GL1* and *GW5.1*) in the study (Fig.2i and 3i).

The domestication signatures of grain-size variations at a population level have been revealed through in-depth analyses of the eight genes employing a large and geographically diverse natural population (Fig.4 and 5). Directional selection is the shift in the mean value of a trait from one position in a frequency histogram to a distinctly new value³. We observed the directional accumulation of slender and long allele of the eight grain-size gene during rice improvement process from landrace to cultivar in *indica* but not in *japonica* rice, which resulted in the significant shift of grain width and length towards slenderer and longer directions. Differentially geographical distribution and ecology habit may lead to the differentially directional selection on grain-size genes between *indica* and *japonica* population (Supplementary Fig. 6–8). Usually, slender- and long-grain rice with aroma is considered as good quality rice, while short- and bold-grain rice is considered as the ordinary rice⁴⁹. The reasonable explanations for these preference tendencies may be referring to the increasing suitability to human requirements such as the taste, yield, quality and cultivation practices. For example, the slender *indica* grains always have excellent appearance, cooking and eating quality with much less grain chalkiness than those of wide grains⁴⁹. Farmers preferred to select larger or longer seeds to increase grain yield and obtain more food during early domestication. An increase in grain size is an early adaptive response to human cultivation, and greater grain size is strongly correlated with larger seedlings in many cereal and legume species²⁴. Therefore, human cultivation practices coupled with food demand preferences may result in the main tendency of the directional selection. The allele frequency and haplotype analyses supplied the flexibility to the molecular design breeding of ideal grain-shape cultivars.

Selection sweeps of eight grain-size genes were identified in modern cultivars, landraces and wild rice to provide other evidences for strong selection pressures subjected on these genes and the selection signatures fixed in the early domestication (Fig.5 and Supplementary Fig.9). Artificial selection for human

requirements and cultivation practices in the past was mainly performed on target traits, and thus, the alleles with the major effects on the trait have been heavily selected. The coincidence of large grain-size effects of *GW5*, *GS3* and *GL7* and strong selection pressure acting on them fixed before the completion of rice domestication let us to clarify the relationship between them. As we expected, we discovered the genetic effects of the eight grain-size genes highly and positively correlate with the directional selection intensities and the nucleotide variation intensities, which is a new finding different from previous studies⁵⁰. Differential extent of selection on the eight genes plays an essential role in changing the correlation between sufficient accumulation of DNA variations and grain-size phenotypes in the large diverse population. These conclusions may be applicable to other traits, although more data are needed to furtherly verify its power and robustness. Crop domestication is a lengthy process of selection for DNA variations in a suite of traits that eventually renders a plant cultivable^{51, 52}. Mimicking this process through neo-domestication of wild, landrace or cultivar species using multiple QTL genes coupled with gene editing technologies could be an alternative way to quickly breed modern cultivars⁵².

In this study, we have developed RapMap, a practical and versatile approach to rapidly discover functional genes in high-throughput, which has played an important role in cloning grain-size genes in rice. Although the concept of RapMap was well demonstrated using rice, RapMap could be immediately extended to any other crops that are easy to crossbreed and reproduce. Many crop germplasm collections^{10, 15}, recently available for rice, maize, wheat, barley, potato, soybean, cotton, and pea crops and their wild relatives, will facilitate immediate and comprehensive application of RapMap. The great ability to rapidly clone agriculturally valuable genes by RapMap will further increase the value of these germplasm collections. So, we believe RapMap is proposed as a new paradigm and benchmark for identifying QTL genes for its great potential in functional genomics. Various functional genes from diverse germplasms could be fully applied for accelerating the genetic improvement of valuable traits by precise molecular breeding in plants.

Methods

All the primers used in this study are listed in **Supplementary Table 2**.

Rice mini-core collection

The rice parental lines for generating gradient populations were selected from a diverse mini-core collection including 541 *O. sativa* accessions sequenced in a previous study^{28, 29}. The mini-core collection contains both landraces and the elite varieties including 215 varieties from China's core collection of *Oryza sativa* L. and 326 varieties from other 58 countries worldwide (Supplementary Table 1). All the accessions were grown in the normal season at the Experimental Stations of Huazhong Agricultural University, Wuhan and Lingshui, China. Plants were spaced by 16.5 cm in each row, and rows were 26 cm apart. The rice plants were under normal field management such as irrigation, fertilizer application and pest control.

Evaluation of grain size and other yield traits

Seed size (grain length and width) and other yield traits was measured as described by Li *et al*⁵³. Harvested rice grains were air-dried and stored at room temperature for at least one month before measuring. Fully filled grains were used to measure grain length, width and weight. Ten randomly chosen grains from each plant were lined up length wise along a vernier caliper to measure grain length and arranged width wise to measure grain width.

Construction of gradient populations and field trials

According to the phenotype values of the collections, we selected 12 and 10 accessions with gradient phenotypes of grain length and width, respectively (Fig. 2a and 4a, Supplementary Table 1). The parental materials of the gradient populations used in this study were listed in the Supplementary Table 1. Then we arranged the trait-adjacent accessions as the cross parents to generate F₁ seeds in March 2016 at Hainan. The resulting F₁ seeds for each cross were independently self-pollinated in July 2016 at Wuhan, and each F₂ random gradient population (~200 plants) from each F₁ was grown independently in the field of Hainan in April 2017 for measuring their target phenotypes after harvesting. All the field management of gradient populations was the same as the mini-core collection.

Genotyping of bulked DNAs by the RICE6K array analysis and high-resolution sequencing analysis

Seeds of F₂ plants with extreme phenotypic values were used for bulking. For each bulk with extreme phenotypes, > 25 plants and 10 seeds per plant should be selected. Seeds of each bulk were grown in a culture room with average day and night temperatures of 30°C and 25 °C, respectively, and 75% relative humidity under a 12-h light and 12-h dark cycle. The leaves of ten-day old seedlings were equally mixed for bulking. DNA extraction of bulked leaves was performed using the cetyltrimethylammonium bromide (CTAB) method. The RICE6K hybridization, scanning and the following further analyses were performed by China National Seed Group Co., Ltd³⁰.

DNA deep sequencing of four samples was performed by Annoroad Gene Technology Co., Ltd (Beijing, China) using Illumina Hiseq 4000 platform in a 2×150 pair-ended manner. After the quality control, short reads in which more than 10% of sequenced nucleotides exhibited a phred quality score of <30 were excluded from the analysis followed. The followed genotyping and primary mapping procedures were described as Takagi *et al*³¹.

Designing InDel and SNP markers for verifying and fine mapping QTLs

After analyzing the QTLs based on RICE6K or NGS, the potential QTL region was available. Based on the RiceVarMap database covering most parental materials in our study (<http://ricevarmap.ncpgr.cn/v1/>)³², we can effectively find insertion-deletion (InDel) polymorphic variations between bi-parents and rapidly design InDel markers in a candidate interval. SNP markers gained by sequencing may be needed for further fine mapping when InDel markers are not available in a small region.

Rapid DNA extraction technique

For the verification of QTLs and identification of recombinants, we used an ultra-fast protocol for rapid DNA extraction as follows⁵⁴. One-centimeter leaf blade was cut off and placed into a 96-well deep-hole plate. Samples were grinded with 200 μ L of NaOH (0.3 M) for each well, centrifuged for 1 min at 1200 rpm and boiled for 45s. Then, samples were mixed with 300 μ L Tris-HCl (0.75M, pH7.4–7.8), boiled for 45s and centrifuged for 1 min at 1200 rpm. 5 μ L of the supernatant and 45 μ L ddH₂O were transferred to the 96-well PCR plate for screening recombinants.

Screening recombinants by KASP

We used the KASP technology to genotype the potential recombinants in each QTL region³³. For each QTL, the flanking markers should be designed using the “SNP Primer” tool (<http://www.snpway.com/snpprimer/>) based on the variations of SNPs retrieved from the RiceVarMap database³⁰. Furthermore, the screening of recombinants was conducted using the KASP primers by genotyping about 5000 individuals of an F_{2:3} population developed from F₂ plants with heterozygous genotypes of a target QTL. The plants with different genotypes at two markers covering the QTL interval were picked as recombinants, whose phenotypes would be tested in the field for determining their corresponding genotypes of the target gene.

RNA isolation and RT-PCR analysis

Total RNA was extracted from the young panicles using the kit (TransZol, TransGen Biotech, Cat. No.ET101-01). The extracted RNA was incubated with RNase-free DNase I (Amplification Grade, Invitrogen, Cat. No.18068-015) to eliminate genomic DNA contamination. The first-strand cDNA was synthesized from 2 μ g total RNA using the M-MLV reverse transcriptase (Invitrogen, Cat. No.28025013). All the experiments were conducted according to the manufacturer’s instructions with at least three biological replicates. Semi-quantitative reverse transcription-polymerase chain reaction (RT-PCR) was performed to determine the expression level of the target gene. The rice *Ubiquitin* gene (LOC_Os03g13170) was used as an internal control.

Construction and genetic transformation of CRISPR/Cas9 and complementation Vectors

CRISPR/Cas9 technology was used to generate the gene-edited plants of target genes⁵⁵. For a single-target vector, the gene-specific guide sequence (sgRNA) was designed and integrated into the *OsU6* promoter by an overlapping PCR. Then, the sgRNA transcriptional unit was inserted into the *pCXUN-CAS9* plasmid at the *KpnI* site through Gibson assembly⁵⁶. For a double-target vector, the second sgRNA was designed and integrated into the *OsU3* promoter by an overlapping PCR. Subsequently, the second sgRNA transcriptional unit was inserted into the obtained single-target vector at the *SacI* site through Gibson assembly. The CRISPR constructs of *GW5.1* with a single target and *GL1* with double targets were transformed into *Agrobacterium tumefaciens* strain EHA105 and then introduced into Zhonghua 11 by *Agrobacterium*-mediated transformation as previously described⁵³.

To validate the function of *GW5.1* on grain width, a 6-kb genomic fragment, containing a 3.5-kb fragment upstream of *GW5.1* and the coding sequence, was amplified from Iksan438 genomic DNA and inserted into the plant binary vector pCAMBIA 1301-Flag between the *EcoRI* and *KpnI* sites using Gibson assembly. After the destination construct was verified by sequencing, it was introduced into loss-of-function “AA” callus by *Agrobacterium*-mediated transformation as previously described⁵³. To validate the function of *GL1* on grain length, a 5.5-kb genomic fragment, containing a 2.3-kb fragment upstream of *GL1* and the coding sequence, was amplified from WZ1 genomic DNA and inserted into the pCAMBIA 1301-Flag vector. Then the sequence-right construct was introduced into ZH11. All the transgenic plants were grown in the transgenic paddy fields of the Experimental Stations of Huazhong Agricultural University, Wuhan and Lingshui, China.

Evaluation of genetic effects of the eight grain-size genes

Multiple linear regressions were performed to estimate the genetic effects of each gene and to predict the phenotype performance by using the function *lm()* in R (Version 3.6.2). The model was as follows:

$$Y = g_0 + g_1\text{Marker1} + g_2\text{Marker2} + \dots + g_i\text{Markeri} + e$$

in which Y , g_0 , g_j ($j=1, 2, \dots, i$), *Marker* and e mean the phenotype, intercept, regression coefficient, genotype of the functional marker from a gene, and random error, respectively. 10-fold cross-validation was conducted to assess the model prediction performance. The relative importance of each variable (allele) for the phenotype explanation was estimated by a method of calculating the relative weight. The 10-fold cross-validation and the estimation of relative importance were executed by using R script. For the comparison of allelic combinations, one-way ANOVA was performed first and then the multiple comparison was conducted by the least significant difference (LSD) method.

Population genetics analyses of the eight grain-size genes

The genotypes of 446 rice wild species *Oryza rufipogon* were retrieved from Huang et al.⁴⁵. The genotypes of 4,726 rice landrace and cultivar were downloaded from RiceVarMap2 (<http://ricevarmap.ncpgr.cn/v2/>)³². The inference for landrace and cultivar was based on the plant height, the representative functional marker of the *sd1* gene and the known information and curated manually³². The nucleotide diversity level of per silent site was estimated as π by VCFtools⁵⁷. We also checked the nucleotide diversity of a 2-Mb flanking genomic region of each gene with a 100-kb sliding window and a 20-kb step. Silent-site nucleotide diversity between accessions with long/short and wide/slender alleles of each grain-size gene and each other gene within its surrounding 500-kb regions was compared for scanning the selective sweep valley. The selection pressure of each gene was the average ratio of nucleotide diversity of cultivars and landraces. The DNA variations of genes were measured by the Kimura 2-parameter which was calculated by MEGA X⁵⁸ after comparing the gene sequences of the cultivars and the wild species *Oryza rufipogon* W1943. Genome-wide association study for length-width

ratio was conducted using the mini core collection (data not shown) by FaST-LMM software⁵⁹ to obtain the association significance of length-width ratio (the minus log transformed *P* value).

Declarations

Acknowledgements

This study was supported by grants from the National Key Research and Development Program of China (2016YFD0100402), National Natural Science Foundation of China (32072042, 31771368, 31821005), Ten-thousand Talents Program and Fundamental Research Funds for the Central Universities (2662020SKPY006). We thank Prof.s Qifa Zhang and Yongzhong Xing for their suggestions in preparing the manuscript. We thank Prof. Bin Han from Shanghai Institutes for Biological Sciences (Chinese Academy of Sciences) for kindly providing the genotypes and phenotypes of the wild, landrace and cultivar rice used in this study.

Author contributions

J.Z. did most data analyses of all figures and Supplementary Information, and wrote some parts of the manuscript; D.Z., C.Z., Q.S. and X.S. carried out the cloning of *GW8*, *GS2* and *GL 1*, and wrote the new *GL 1* gene part of the manuscript; Y.F. carried out the cloning of the novel *GW5.1* gene, wrote the part and edited the manuscript; C.L., Y.S. and H.Y. carried out the cloning of *GS3*; P.X., W.L., X.L. and M.W. carried out the cloning of *GL 7*, *GW7* and *GW5*; X.H., L.W. and S.W. made some F₂GP crosses. X.L. provided the rice mini-core collection data and materials. Y.L. conceived the idea and principle of RapMap, supervised the whole work, and mainly wrote the manuscript. All authors read and approved the final version of the manuscript.

Competing interests: The authors declare no competing interests.

Correspondence and requests for materials should be addressed to Yibo Li.

Reprints and permissions information is available at www.nature.com/reprints.

Accession codes of GenBank: genomic DNA of *GS3-5* in Zhimali and Heen GW, XXX; genomic DNA of *GS3-6* in SYB5; genomic DNA of *GL 1* in WZ1, XXX; genomic DNA of *GL 1* in 9311, XXX; genomic DNA of *GW5.1* in Zhenshan 97, XXX; genomic DNA of *GW5.1* in Ikaan438, XXX.

References

1. Ellegren, H. & Galtier, N. Determinants of genetic diversity. *Rev. Genet.* **17**, 422–433 (2016).
2. Rheenen, W., Peyrot, W., Schork, A., Lee, S. & Wray, N. Genetic correlations of polygenic disease traits: from theory to practice. *Rev. Genet.* **20**, 567–581 (2019).

3. Mitchell-Olds, T., Willis, J. & Goldstein, D. Which evolutionary processes influence natural genetic variation for phenotypic traits? *Rev. Genet.* **8**, 845–856 (2007).
4. Bazakos, C. et al. New strategies and tools in quantitative genetics: how to go from the phenotype to the genotype. *Rev. Plant Biol.* **68**, 435–455 (2017).
5. Goddard, M. & Hayes, B. Mapping genes for complex traits in domestic animals and their use in breeding programs. *Rev. Genet.* **10**, 381–391 (2009).
6. Boyle, E., Li, Y. & Pritchard, J. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
7. Glazier, A., Nadeau, J. & Aitman, T. Finding genes that underlie complex traits. *Science* **298**, 2345–2349 (2002).
8. Mackay, T., Stone, E. & Ayroles, J. The genetics of quantitative traits: challenges and prospects. *Rev. Genet.* **10**, 565–577 (2009).
9. Costanzo, M. et al. Global genetic networks and the genotype-to-phenotype relationship. *Cell* **177**, 85–100 (2019).
10. Wallace, J., Rodgers-Melnick, E., & Buckler, E. On the road to breeding 4.0: unraveling the good, the bad, and the boring of crop quantitative genomics. *Rev. Genet.* **52**, 421–444 (2018).
11. Holland, J. Genetic architecture of complex traits in plants. *Opin. Plant Biol.* **10**, 156–161 (2007).
12. Benfey, P. & Mitchell-Olds, T. From genotype to phenotype: systems biology meets natural variation. *Science* **320**, 495–497 (2008).
13. Rockman, M. Reverse engineering the genotype-phenotype map with natural genetic variation. *Nature* **456**, 738–744 (2008).
14. Yamamoto, T., Yonemaru, J. & Yano, M. Towards the understanding of complex traits in rice: substantially or superficially? *DNA Res.* **16**, 141–154 (2009).
15. Huang, X. & Han, B. Natural variations and genome-wide association studies in crop plants. *Rev. Plant Biol.* **65**, 531–551 (2014).
16. Kover, P. et al. A multiparent advanced generation inter-cross to fine-map quantitative traits in *Arabidopsis* *PLoS Genet.* **5**, e1000551 (2009).
17. Xiao, Y. et al. Genome-wide association studies in maize: praise and stargaze. *Plant* **10**, 359–374 (2017).
18. Li, N., Xu, R. & Li, Y. Molecular networks of seed size control in plants. *Annu. Rev. Plant Biol.* **70**, 435–463 (2019).
19. Fan, Y. & Li, Y. Molecular, cellular and Yin-Yang regulation of grain size and number in rice. *Breeding* **39**, 163–187 (2019).
20. Zuo, J. & Li, J. Molecular genetic dissection of quantitative trait loci regulating rice grain size. *Annu Rev Genet* **48**, 99–118 (2014).
21. Klingenberg, C. Evolution and development of shape: integrating quantitative approaches. *Rev. Genet.* **11**, 623–635 (2010).

22. Hufford, M. et al. Crop biodiversity: an unfinished magnum opus of nature. *Annu. Rev. Plant Biol.* **70**, 727–751 (2019).
23. Olsen, K. & Wendel, J. A bountiful harvest: genomic insights into crop domestication phenotypes. *Annu. Rev. Plant Biol.* **64**, 47–70 (2013).
24. Purugganan, M.D. & Fuller, D.Q. The nature of selection during plant domestication. *Nature* **457**, 843–848 (2009).
25. Purugganan, M.D. Evolutionary insights into the nature of plant domestication. *Curr. Biol.* **29**, R705–R714 (2019).
26. Gage, J. et al. Ten years of the maize nested association mapping population: impact, limitations, and future directions. *Plant Cell* **32**, 2083–2093 (2020).
27. Zou, C., Wang, P. & Xu, Y. Bulk sample analysis in genetics, genomics and crop improvement. *Plant Biotechnol. J.* **14**, 1941–1955 (2016).
28. Xie, W. et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Natl. Acad. Sci. USA* **112**, E5411–E5419 (2015).
29. Chen, W. et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Genet.* **46**, 714–721 (2014).
30. Yu, H. et al. A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol. J.* **12**, 28–37 (2014).
31. Takagi, H. et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. *Plant J.* **74**, 174–183 (2013).
32. Zhao, H. et al. RiceVarMap: a comprehensive database of rice genomic variations. *Nucleic Acids Res.* **43**, D1018–D1022 (2015).
33. Semagn, K. et al. Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement. *Breeding* **33**, 1–14 (2014).
34. Fan, C. et al. *GS3*, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Appl. Genet.* **112**, 1164–1171 (2006).
35. Takano-Kai, N. et al. Evolutionary history of *GS3*, a gene conferring grain length in rice. *Genetics* **182**, 1323–1334 (2009).
36. Mao, H. et al. Linking differential domain functions of the *GS3* protein to natural variation of grain size in rice. *Natl. Acad. Sci. USA* **107**, 19579–19584 (2010).
37. Wang, Y. et al. Copy number variation at the *GL7* locus contributes to grain size diversity in rice. *Genet.* **47**, 944–948 (2015).
38. Hu, J. et al. A rare allele of *GS2* enhances grain size and grain yield in rice. *Plant* **8**, 1455–1465 (2015).
39. Duan, P. et al. Regulation of *OsGRF4* by *OsmiR396* controls grain size and yield in rice. *Plants* **2**, 15203 (2015).

40. Wang, S. et al. The *OsSPL 16-GW7* regulatory module determines grain shape and simultaneously improves rice yield and grain quality. *Genet.* **47**, 949–954 (2015).
41. Wang, S. et al. Control of grain size, shape and quality by *OsSPL 16* in rice. *Genet.* **44**, 950–954 (2012).
42. Shomura, A. et al. Deletion in a gene associated with grain size increased yields during rice domestication. *Genet.* **40**, 1023–1028 (2008).
43. Duan, P. et al. Natural variation in the promoter of *GSE5* contributes to grain size diversity in rice. *Plant* **10**, 685–694 (2017).
44. Liu, J. et al. *GW5* acts in the brassinosteroid signalling pathway to regulate grain width and weight in rice. *Plants* **3**, 17043 (2017).
45. Huang, X. et al. A map of rice genome variation reveals the origin of cultivated rice. *Nature* **490**, 497–501 (2012).
46. Abe, A. et al. Genome sequencing reveals agronomically important loci in rice using MutMap. *Biotechnol.* **30**, 174–178 (2012).
47. Zhang, H. et al. QTG-Seq accelerates QTL fine mapping through QTL partitioning and whole-genome sequencing of bulked segregant samples. *Plant* **12**, 426–437 (2019).
48. Wang, C. et al. Dissecting a heterotic gene through GradedPool-Seq mapping informs a rice-improvement strategy. *Commun.* **10**, 2982 (2019).
49. Fitzgerald, M., McCouch, S. & Hall, R. Not just a grain of rice: the quest for quality. *Trends Plant Sci.* **14**, 133–139 (2009).
50. Lu, L. et al. Natural variation and artificial selection in four genes determine grain shape in rice. *New. Phytol.* **200**, 1269–1280 (2013).
51. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
52. Hickey, L. et al. Breeding crops to feed 10 billion. *Nat. Biotechnol.* **37**, 744–754 (2019).
53. Li, Y. et al. Natural variation in *GS5* plays an important role in regulating grain size and yield in rice. *Nat. Genet.* **43**, 1266–1269 (2011).
54. Lu, J. et al. A direct PCR-based SNP marker-assisted selection system (D-MAS) for different crops. *Mol. Breeding* **40**, 9–18 (2020).
55. He, Y., Zhang, T., Yang, N., Xu, M., Yan, L., Wang, L., Wang, R. & Zhao, Y. Self-cleaving ribozymes enable the production of guide RNAs from unlimited choices of promoters for CRISPR/Cas9 mediated genome editing. *J. Genet. Genomics* **44**, 469–472 (2017).
56. Gibson, D., Young, L., Chuang, R., Venter, J., Hutchison, C. & Smith, H. Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
57. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
58. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: Molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).

59. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).

Figures

Figure 1

A simplified scheme and principle of RapMap. (a) Construction of F_2 gradient populations using trait-adjacent gradient parents from diverse germplasms. Each F_2 gradient population is generated from the self-pollinated F_1 plants and the phenotypes of F_2 progenies are measured for each gradient population. (b) Identification of QTLs. Two DNA pools of high-value bulk and low-value bulk were subjected to microarray chip or next-generation sequencing analysis. (c) The co-segregation standard that two homozygous genotypes of a QTL in progenies of any segregating population can distinguish their target phenotypes, is the general principle of the 12 single-locus genetic models and also the quality-control standard for the simultaneous verification of a real QTL and its NIL-LLs. All the individuals from each

gradient population are genotyped by markers flanking each candidate QTL region and the co-segregation standard is employed to simultaneously confirm each QTL and its NIL-LLs. (d) Fine mapping and cloning of each QTL gene by screening enough recombinants using the heterozygous plants as NIL-LLs of the real QTL, coupled with various candidate-gene methods. (e) Traditional processes of QTL cloning that take about 8 years and great efforts. (f) The genetic population not meeting the co-segregation standard could be replaced by a phenotype-well-segregating F3 or F4 family derived from a randomly selected low-, medium- or high-value F2 or F3 line, respectively.

Figure 2

Cloning of multiple QTL genes for grain-length variations in rice using RapMap. (a) Construction of gradient populations using grain length-adjacent parents selected from the rice mini-core collection worldwide. (b) Distribution of grain length of each gradient population. Green and blue arrows indicate low-value and high-value parents, respectively. Top and bottom 15% of the plants were pooled as low-value and high-value bulks, respectively. (c) Identification of grain-length QTLs by RICE6K or next-generation sequencing analyses. Cross 1–3, 5, 7 and 8 display the QTLs identified by RICE6K, while Cross 4 displays the QTL identified by next-generation sequencing analysis. (d) Confirmation of the candidate grain-length QTLs by the co-segregation standard. A (blue), B (green) and H (grey) represent high-value homozygous, low-value homozygous and heterozygous genotypes of each QTL for each gradient population, respectively. (e) Fine mapping and cloning of each grain-length QTL gene using recombinants from corresponding heterozygous NIL-LLs. The black and white bars indicate the chromosome and the QTL interval, respectively. The markers are marked above the chromosomes. The number of the recombinants supporting each marker was listed under the chromosomes corresponding to the marker. The gene structures of two parents are drawn at the bottom, while the boxes and the lines represent exons and introns of genes, respectively. Triangles and arrows in red indicate functional insertions and SNPs, respectively. (f) Confirmation of the genes by comparisons of grain length between two homozygous NIL-LLs identified by the functional marker of each cloned gene. (g) Grain weight of NIL(gl1) and NIL(GL1) for the novel grain-length gene GL1. (h) Phenotypes of CRISPR lines (Cri1 to Cri3), a negative line (CK) and positive complementary lines (L1 to L4) for the novel GL1 gene. (i, j) Statistics of grain length (i) and weight (j) of CRISPR lines, negative lines and complementary lines (n=15) for GL1. All Data represent mean $\pm$ s.e.m. Student's t-tests were used to generate P values.

Figure 3

Identification of multiple QTL genes for grain-width variations in rice using RapMap. (a) Construction of gradient populations with grain width-adjacent parents selected from the rice mini-core collection. (b) Distribution of grain width of each gradient population. Green and blue arrows indicate low-value and high-value parents, respectively. Top and bottom 15% of the plants were pooled as low-value and high-value bulks, respectively. (c) Identification of grain-width QTLs by RICE6K or next-generation sequencing

analyses. Cross 1–4 and 6 display the QTLs identified by RICE6K, while Cross 7 displays the QTL identified by next-generation sequencing analyses. (d) Confirmation of the candidate grain-width QTLs by the co-segregation standard. A (blue), B (green) and H (grey) represent high-value homozygous genotypes, low-value homozygous genotypes and heterozygous genotypes of each QTL for each gradient population, respectively. (e) Fine mapping and cloning of each grain-width QTL gene using recombinants from corresponding heterozygous NIL-LLs, coupled with various candidate-gene methods. The black and white bars indicate chromosome and the QTL interval, respectively. The markers are marked above the chromosomes. The number of the recombinants supporting each marker was listed under the chromosomes corresponding to the marker. The gene structures of two parents are drawn at the bottom, while the boxes and the lines represent exons and introns of genes, respectively. Triangles in red indicate functional insertions. (f) Confirmation of the genes by comparisons of grain width between two NIL-LLs identified by the functional marker of each cloned gene. (g) Validation of the 11.8-kb LTR-retrotransposon insertion in DNA and mRNA level. (h) Genomic sequence showing the gRNA target site in green and the resulting frameshift mutations in red in the ORF1 gene of the three independent CRISPR plants (CR4-4, CR4-8 and CR9-3). The PAM sequences are in blue. (i) Functional verification of GW5.1 by CRISPR and complementation transformation. Com. is the positive complementary line; ZH11 and AA is the control of the CRISPR and complementation transformation, respectively. Data represent mean $\pm$ s.e.m. ($n>10$). Student's t-tests were used to generate P values.

Figure 4

Genetic contribution and directional selection of grain-size genes during rice domestication and improvement. (a–c) The contribution and relative importance of the identified genes for grain length (a), grain width (b) and grain shape (c) of brown rice in the rice mini-core collection with 541 accessions. The phenotype variance explanation (R^2), determined by multiple linear regression with the genotypes of the functional marker of each gene as the variables, was used for assessing the genetic contribution of these

genes. (d–f) The power of the prediction model for grain length (d), grain width (e) and grain shape (f). 10-fold cross-validation was used to estimate the model prediction performance. (g, h) Proportions and directional selection of alleles of grain-length (g) and grain-width (h) genes within each of the three germplasm groups with wild, landrace and cultivar rice of 446, 2462 and 784 accessions, respectively. (i) The average allele frequency of long-grain and slender-grain alleles from wild to landrace and from landrace to cultivar. (j) The average grain length, grain width and length-width ratio of the landrace and cultivar rice. (k) The nucleotide diversity of grain-size gene loci (2 Mb) in wild, landrace and cultivar rice.

Figure 5

Selection sweeps of the eight grain-size gene loci in cultivar rice and relationship among their selection intensities, DNA variations and grain-shape effects. (a–h) Selection sweep scanning across *O. sativa* genome regions of the eight grain-size gene loci in cultivar rice of 784 accessions. Bars indicate the ratios of silent-site nucleotide diversities of each gene in these loci between long/short allele accessions or wide/slender allele accessions. The smallest bars are equivalent to a ratio of 0. Red arrows are indicative of each grain-size gene. (i–j) The correlation among DNA variations, selection pressure and phenotype effects of the eight grain-size genes. The phenotype variance explanation (explaining percentage for length-width ratio) was the R^2 used in (Fig. 4c). Association significance of length-width ratio was the minus log transformed P value of genome-wide association analysis for length-width ratio. Selection pressure was the π ratio of cultivar versus landrace for each gene using VCFtools. DNA variation is the DNA changes of each gene by comparing the gene sequences of the cultivars and the wild species *Oryza rufipogon* W1943 using MEGA X (j). The arrow width and the number font size are scaled to the size of the correlation coefficients (j).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [Supplementaryinformation.docx](#)
- [SupplementaryTable3.pdf](#)
- [SupplementaryTable4.pdf](#)