

# GWAS meta-analysis using a graph-based pan-genome enhanced gene mining efficiency for agronomic traits in rice

Received: 23 April 2024

Accepted: 6 March 2025

Published online: 03 April 2025

 Check for updates

Longbo Yang<sup>1,2,6</sup>, Wenchuang He<sup>2,6</sup>, Yiwang Zhu<sup>2,6</sup>, Yang Lv<sup>2,6</sup>, Yilin Li<sup>1,2</sup>, Qianqian Zhang<sup>1,2</sup>, Yifan Liu<sup>1,2</sup>, Zhiyuan Zhang<sup>1,2</sup>, Tianyi Wang<sup>2</sup>, Hua Wei<sup>2</sup>, Xinglan Cao<sup>2</sup>, Yan Cui<sup>2</sup>, Bin Zhang<sup>2</sup>, Wu Chen<sup>2</sup>, Huiying He<sup>2</sup>, Xianmeng Wang<sup>2</sup>, Dandan Chen<sup>2</sup>, Congcong Liu<sup>2</sup>, Chuanlin Shi<sup>2</sup>, Xiangpei Liu<sup>2</sup>, Qiang Xu<sup>2</sup>, Qiaoling Yuan<sup>2</sup>, Xiaoman Yu<sup>2</sup>, Hongge Qian<sup>2</sup>, Xiaoxia Li<sup>2</sup>, Bintao Zhang<sup>2</sup>, Hong Zhang<sup>2</sup>, Yue Leng<sup>2</sup>, Zhipeng Zhang<sup>2</sup>, Xiaofan Dai<sup>2</sup>, Mingliang Guo<sup>2</sup>, Juqing Jia<sup>1</sup>, Qian Qian<sup>2,3,4,5</sup>✉ & Lianguang Shang<sup>2,4,5</sup>✉

Genome-wide association studies (GWASs) encounter limitations from population structure and sample size, restricting their efficacy. Though meta-analysis mitigates these issues, its application in rice research remains limited. Here, we report a large-scale meta-analysis of six independent GWAS experiments in rice to mine genes for key agronomic traits. By integrating a rice pan-genome graph to identify structural variants, we obtained 6,604,898 SNP and 42,879 PAV variants for the six panels (7765 accessions). Meta-analysis significantly improved quantitative trait loci (QTLs) detection and hidden heritability by up to 43 and 37.88%, respectively. Among 156 QTLs identified for six agronomic traits, 116 were exclusively detected through meta-analysis, highlighting its superior resolution. Two novel QTLs governing grain width and length were functionally validated through CRISPR/Cas9, confirming their candidate genes. Our findings underscore the utility and potential advantages of this pan-genome-based meta-GWAS approach, providing a scalable model for efficiently gene mining from diverse rice germplasms.

Rice (*Oryza sativa* L.) stands as a crucial food crop globally, with the imperative to enhance its yield to cater to the increasing demands of the world's population<sup>1</sup>. As pan-genomic research progresses, there is a burgeoning abundance of variant information in terms of quantity and diversity. Delving into these genetic variants to unearth new beneficial genes forms a pivotal basis for rice genetic enhancement<sup>2–4</sup>. Among various population genetics techniques, GWAS remains the predominant and representative approach for

unraveling complex traits. It can be used to explore and assess trait-associated functional genes efficiently across extensive germplasm resources<sup>5–7</sup>. Nonetheless, GWAS has several limitations: it is sensitive to population structure and sample size, and its capability to analyze low-frequency variants is constrained, hampering its efficacy in pinpointing new QTLs<sup>8</sup>.

Meta-analysis can provide a remedy by integrating outcomes from multiple independent GWAS experiments, thereby enlarging the

<sup>1</sup>College of Agriculture, Shanxi Agricultural University, Shanxi 030801, China. <sup>2</sup>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. <sup>3</sup>State Key Laboratory of Rice Biology, China National Rice Research Institute, Hangzhou 310006, China. <sup>4</sup>Yazhouwan Laboratory, No. 8 Huanjin Road, Yazhou District, Sanya City, Hainan Province 572024, China. <sup>5</sup>Academician Workstation, National Nanfan Research Institute (Sanya), Chinese Academy of Agricultural Sciences, Sanya 572024, China. <sup>6</sup>These authors contributed equally: Longbo Yang, Wenchuang He, Yiwang Zhu, Yang Lv. ✉e-mail: [qianqian188@hotmail.com](mailto:qianqian188@hotmail.com); [shanglianguang@cass.cn](mailto:shanglianguang@cass.cn)

sample size and notably enhancing the precision and scope of gene mining research. Thus far, meta-GWAS has found extensive application across diverse animal and plant species, including humans<sup>9</sup>, tomatoes<sup>10</sup>, and cattle<sup>11</sup>, underscoring the benefits of meta-GWAS for in-depth gene exploration. In recent decades, extensive GWAS research data have accumulated for rice via various germplasm populations, resulting in the creation of numerous outstanding databases<sup>12–14</sup>. Nevertheless, to the best of our knowledge, meta-GWAS studies that leverage large-scale population data for comprehensive gene mining in rice are still uncommon.

In this study, we carried out a large-scale meta-GWAS analysis to identify new functional genes associated with important rice agronomic traits, utilizing a total of 7765 cultivated rice (*Oryza sativa*) accessions. Independent GWAS studies were conducted for six distinct agronomic traits across these six panels. Following this, a meta-analysis was performed to consolidate the findings from all panels for each trait, resulting in the identification of 156 QTLs, 116 of which were entirely novel. We subsequently undertook a detailed analysis of two novel QTLs associated with grain size to pinpoint and validate two new genes responsible for rice grain width and grain length. Our research clearly highlights the feasibility and potential benefits of meta-GWAS in the study of rice population genetics, offering a robust tool and practical model for harnessing large-scale and multi-population data to mine favorable genes related to crucial agronomic traits in rice.

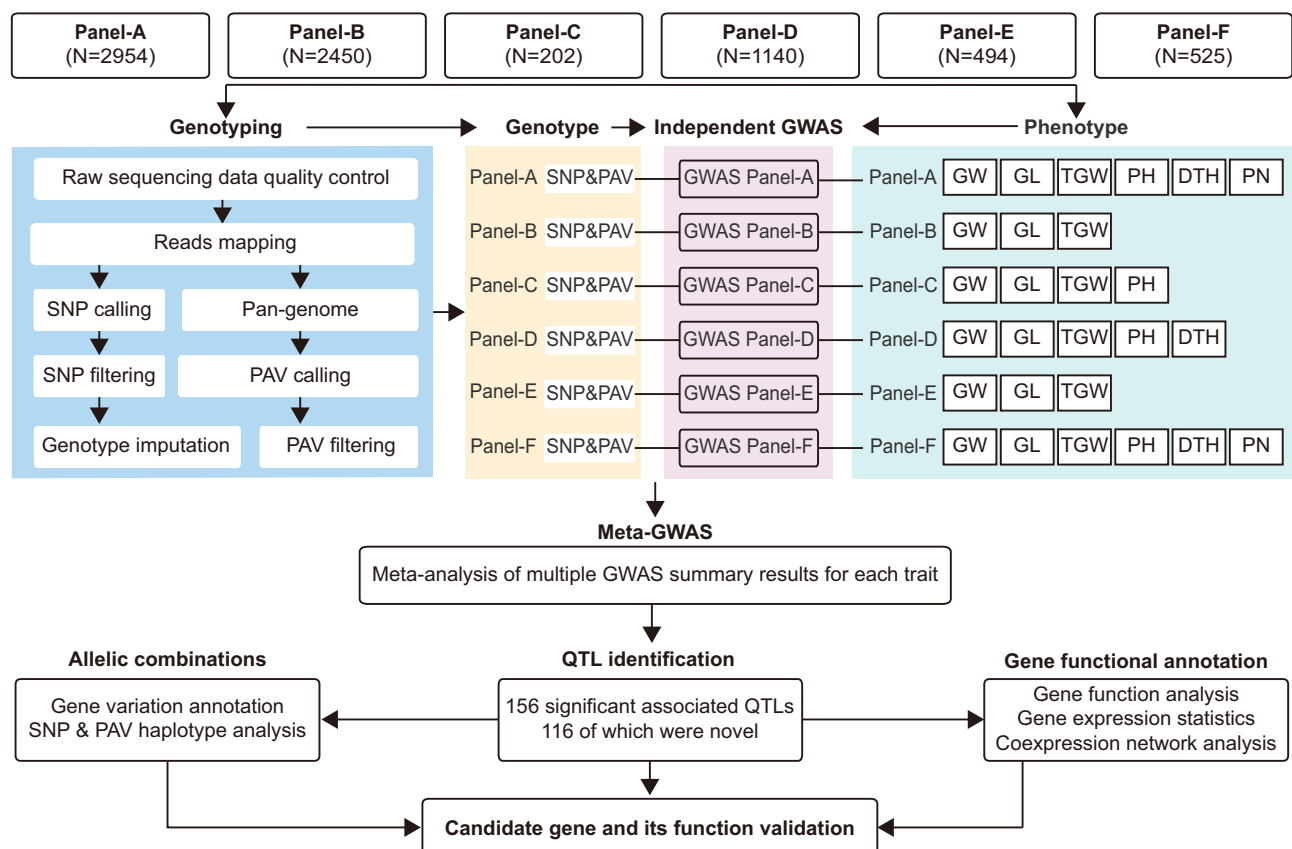
## Results

### Meta-analysis

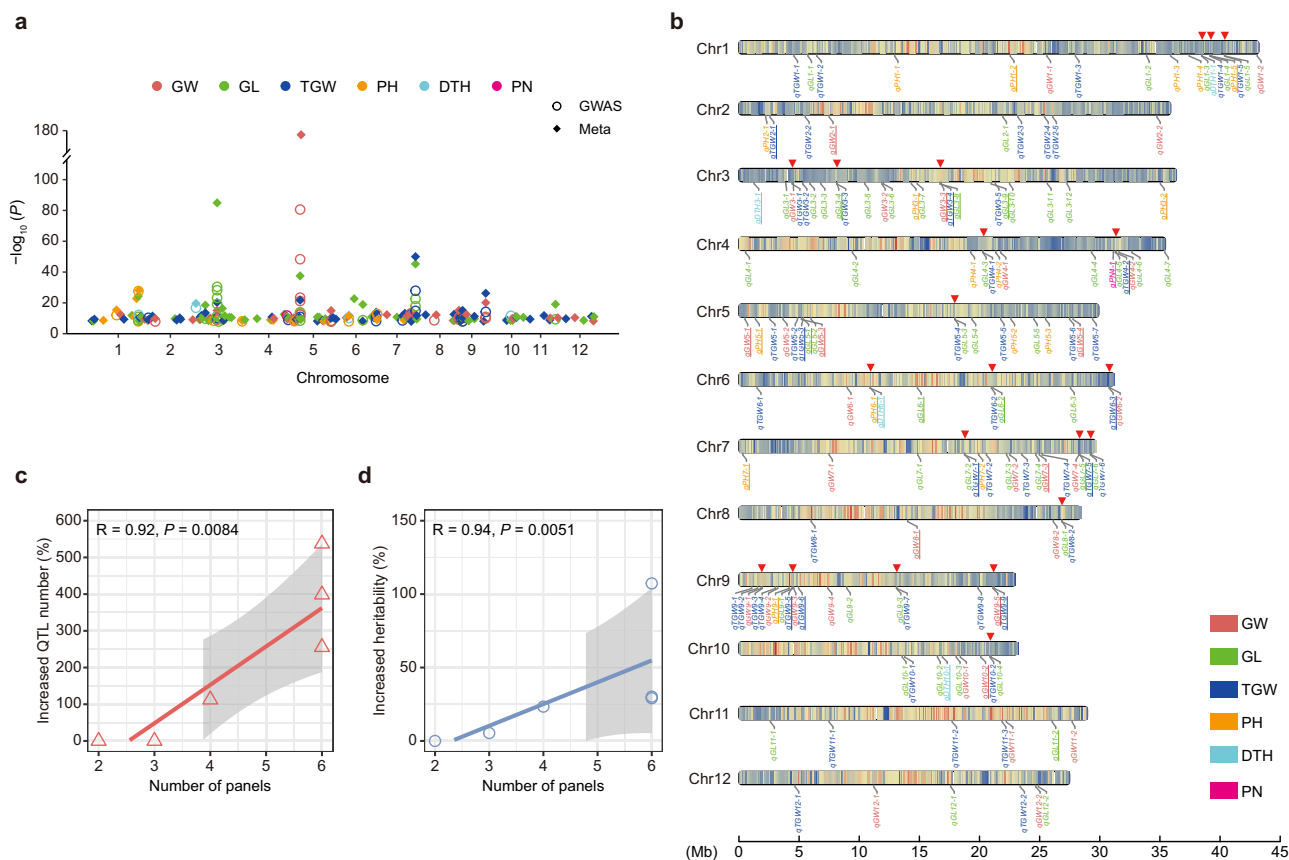
In this study, we report the largest meta-GWAS analysis conducted in rice research to date (Fig. 1). The raw genomic sequencing data and phenotypic data were sourced from six independent rice population

panels, denoted as Panel A–F. These panels encompassed a total of 7765 cultivated rice accessions derived from 126 countries or regions worldwide, including 2954<sup>15</sup>, 2450<sup>16</sup>, 202<sup>4</sup>, 1140<sup>17</sup>, 494<sup>18</sup> and 525<sup>19</sup> rice accessions in Panel A–F, respectively (Supplementary Data. 1). Quality control was conducted on the raw sequencing data for all panels using consistent standards, followed by reads mapping and variant calling processes. The raw SNPs from each panel were filtered (MAF 0.05, missing rate 0.1), resulting in 6,604,898 high-quality SNPs, with each panel containing between 2,890,946 and 4,786,218 SNPs. To obtain a high-quality structural variant dataset, the sequencing reads were mapped to a pan-genome, containing 356 rice accessions as described in our previous study<sup>4,13</sup>, yielding 42,871 high-quality presence/absence variants (PAVs), varying from 9041–34,175 across panels (Supplementary Fig. 1). Six agronomic traits, including grain width, grain length, thousand-grain weight, plant height, heading date, and panicle number, were collected from the six panels (Supplementary Fig. 2). Independent GWAS was carried out between the combined SNP + PAV dataset and phenotypic dataset of each trait for all panels (Supplementary Figs. 3–8). Bonferroni's correction was applied to GWAS outcomes to set the thresholds ( $0.05/n$ , where  $n$  is the total variants count) for false positive control and trustworthy conclusions. Across all panels, 9004 significant variants were identified, with individual panel variants ranging from 0–2600, resulting in a total of 40 QTLs for the six agronomic traits (Supplementary Data 2).

For each trait, a meta-analysis was performed to consolidate all the available GWAS findings from different panels using the fixed-effect model. By merging multiple independent GWAS experiments, meta-analysis significantly enhanced the significance of the association  $P$ -values for the trait-associated variants across nearly all the QTLs (Fig. 2a, Supplementary Fig. 9) leading to an average of 6.79 orders of



**Fig. 1 | Flowchart of study design.** PAV presence/absence variations, GW grain width, GL grain length, TGW thousand-grain weight, PH plant height, DTH days to heading, PN panicle number, GWAS genome-wide association study, QTL quantitative trait locus.



**Fig. 2 | Statistics of trait-associated variants detected by meta-analysis of GWAS results in 6 rice panels. a** Lead variants of QTLs from GWAS and meta-analysis for different phenotypes. **b** Illustrated of the 156 QTLs along the rice chromosomes. The colors of the QTL labels represent different phenotypes, as listed in the legend. The density of SNP + PAV is measured in a 100 Kb window for each chromosome. QTLs labeled with underscores were identified by both GWAS and meta-analysis.

Inverted triangles in red indicate that overlapping QTLs of different phenotypes. **c** Percentage increase in QTL number by meta-analysis relative to individual GWAS results. **d** Percentage increase in heritability explained by QTLs of meta-analysis relative to individual GWAS results. In **(c, d)**, the error bands indicate 95% confidence intervals,  $p$ -values were calculated by two-sided spearman correlation test.

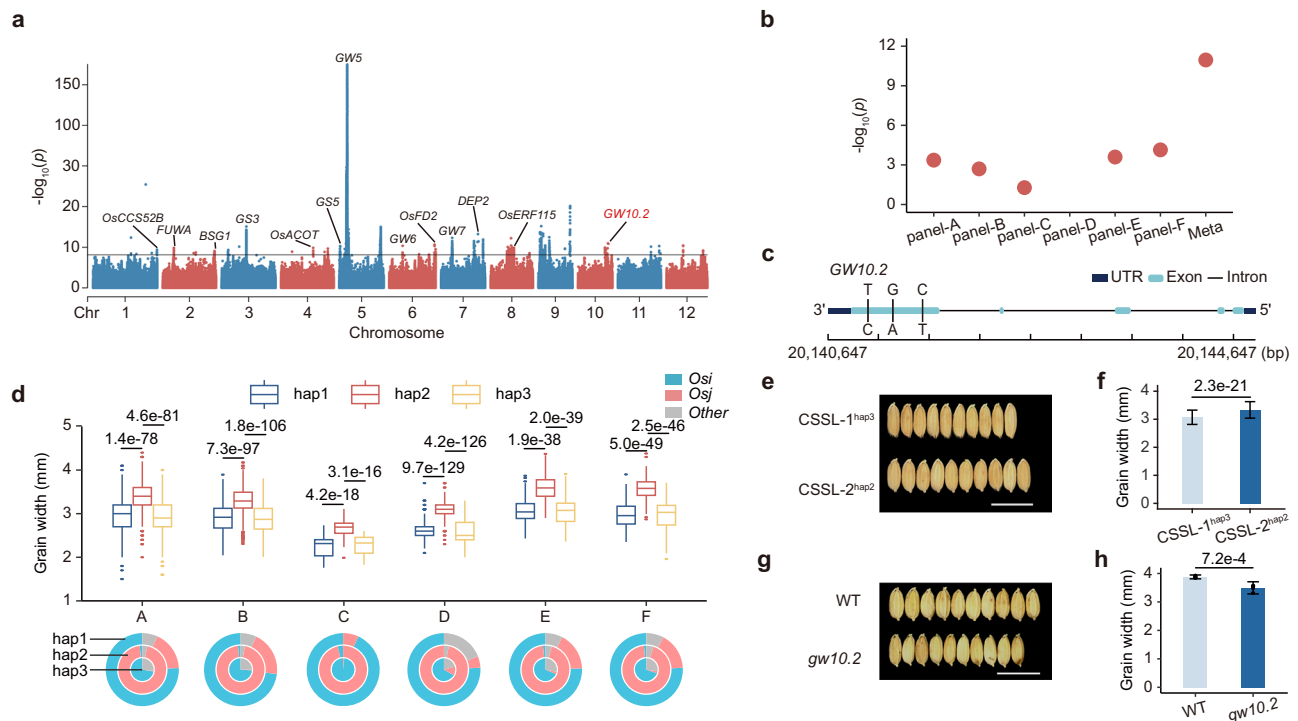
magnitude increase (ranging from -1.14–98.45). We further compared the identified QTLs between the GWAS and meta-analysis (Supplementary Figs. 10–11). Compared with independent GWAS experiments, the meta-GWAS approach detected an equal or greater number of QTLs (ranging from 0–537.5%) and explained more or a similar degree of heritability (ranging from 0% to 107.3%), utilizing a broader range of rice panels (Fig. 2c, d). This enhancement demonstrated a notably positive correlation with both the quantity of QTLs ( $R = 0.92$ ,  $P = 0.0084$ ) and their explained heritability ( $R = 0.94$ ,  $P = 0.0051$ ). Notably, almost all the QTLs (40/40) observed in the individual GWAS results were recovered via meta-analysis, but different GWAS may contribute differently to the meta-analysis; e.g., for panels A–F, a total of 4, 6, 1, 1, 2 and 3 GWAS-based QTLs related to grain width were recovered via the meta-analysis, respectively (Supplementary Figs. 3–8, Supplementary Data 2).

The meta-analysis identified a total of 156 QTLs associated with the six traits, with 116 being unique and not detected in the independent GWAS experiments (Supplementary Data. 3). The QTLs are distributed across all 12 chromosomes (Fig. 2b) and their rate of variation explained for each of the six phenotypes was estimated separately (Supplementary Data. 4). We further screened the functional genes located in the QTLs and identified a total of 141 known genes related to the corresponding traits in rice. Specifically, 29 genes were related to grain width, such as *GW5* ( $qGW5-3$ ,  $P = 8.81e-180$ ), *GW6* ( $qGW6-1$ ,  $P = 4.39e-11$ ), *GW7* ( $qGW7-3$ ,  $P = 5.73e-14$ ), etc.; 37 genes were related to grain length, such as *GS3* ( $qGL3-8$ ,  $P = 1.23e$

–85), *GL6* ( $qGL6-3$ ,  $P = 1.15e-11$ ), *GL7* ( $qGL7-4$ ,  $P = 3.36e-13$ ), and *GL10* ( $qGL10-4$ ,  $P = 1.06e-11$ ), etc.; 46 genes were related to thousand-grain weight, such as *GIF1* ( $qTGW4-1$ ,  $P = 1.49e-11$ ), *GSE5* ( $qTGW5-3$ ,  $P = 1.77e-22$ ), *OsFD2* ( $qTGW6-3$ ,  $P = 1.40e-13$ ), and *FZP* ( $qTGW7-5$ ,  $P = 9.52e-51$ ), etc.; and 26 genes were related to plant height, such as *SD1* ( $qPH1-4$ ,  $P = 1.59e-23$ ), *OsCCT22* ( $qPH6-1$ ,  $P = 2.92e-12$ ), *OsMADS15* ( $qPH7-1$ ,  $P = 9.22e-10$ ), etc.; and three genes were related to days to heading, such as *DTH3* ( $qDTH3-1$ ,  $P = 2.11e-20$ ), *OsCCT22* ( $qDTH6-1$ ,  $P = 1.01e-12$ ), *Ehd1* ( $qDTH10-1$ ,  $P = 2.45e-10$ ). It is worth mentioning that there were 63 QTLs in which no genes known to be associated with phenotypes were found, and these QTLs are valuable for further analysis. Among the six phenotypes, the grain shape can affect both the yield and rice quality traits of rice, thus QTLs for grain width and length were prioritized for further in-depth analysis. The meticulous exploration of these genes is essential for understanding intricate traits. These findings underscore the precision efficacy, and potential benefits of meta-GWAS analysis in mining genes within rice.

### Identification and validation of the candidate gene associated with grain width

Among the 32 grain-width-associated QTLs, 23 were found exclusively in the meta-analysis and could not identified in individual GWAS (Fig. 3a). Among the meta-exclusive QTLs, 15 reported QTLs related to grain width (Supplementary Data. 5). For the remaining 9 meta-exclusive novel QTLs, *qGW10-2* presented the lowest associated  $P$ -value (lead variant: Chr10\_20159929, A > C,  $P = 1.11e-11$ ) and was



**Fig. 3 | Meta-GWAS for grain width and its functional analysis.** **a** Manhattan plot for the meta-analysis of genome-wide association study of grain width. The genes labeled in the figure are the known grain width genes associated with meta-GWAS, and the red one is the identified candidate gene. The black line represents the genome-wide significance threshold at 7.52e-9. **b**  $P$ -values of Chr10:20159929 in GWAS and meta-analysis. **c** Gene structure, showing 3 missense mutation SNPs in the exon region. **d** Boxplot showing the grain width at different haplotypes in each panel. Statistical differences according to two-tailed Student's  $t$ -test. The pie chart displays the subpopulation distribution of the three haplotypes in each panel (*Osi*,

*Oryza sativa indica*; *Osj*, *Oryza sativa japonica*). **e** The difference of grain width of Chromosome Segment Substitution Lines in different accessions. **g** Difference in grain width between wild type and mutant. Scale bars: 1 cm. Values are presented as means  $\pm$  SD in **f** and **h**, Student's  $t$ -test. CSSL-1<sup>hap3</sup> ( $n = 246$ ) and CSSL-2<sup>hap2</sup> ( $n = 191$ ) in **f**. Wild type ( $n = 5$ ) and mutants ( $n = 5$ ) in **h**. In **d**, the horizontal line represents the median value, the lower and upper quartiles represent the 25th and 75th percentile, and the whiskers show the maximum and minimum values. Source data are provided as a Source Data file.

selected for further examination (Fig. 3b). Four steps were performed to screen candidate genes for *qGW10-2* within a 200 kb vicinity of this variant; (1), Variants in all the genes in this region were assessed, focusing on missense, frameshift, and premature-stop-codon mutations for subsequent haplotype analysis; (2), Genes with distinct haplotypes exhibiting significantly different traits across the six panels were selected for further evaluation using haplotype analysis; (3), Genes lacking proper annotation and with a median expression below 0 in seeds were excluded by integrating gene function data from [ricedata.cn/gene/](http://ricedata.cn/gene/), and transcriptome information from [plantradb.com/ricerna/](http://plantradb.com/ricerna/)<sup>20</sup>. Two genes were identified (*LOC\_Os10g37280* and *LOC\_Os10g37640*). (4), Co-expression networks of these two genes were further evaluated, revealing that *LOC\_Os10g37280* was co-expressed with two known genes, *LOC\_Os02g47970*<sup>21</sup> related to leaf shape and *LOC\_Os11g32900*<sup>22</sup> related to pollen development, whereas *LOC\_Os10g37640* exhibited co-expression with a known gene, *LOC\_Os07g46360*<sup>23</sup> related to grain development. Additionally, we noted that *LOC\_Os10g37640* was involved in auxin signaling pathway regulation (GO:0010928) and cell elongation regulation (GO:0051510), which was also noted for two known genes related to grain size, *GS5*<sup>24</sup> and *GLW7*<sup>6</sup>. Consequently, we identified *LOC\_Os10g37640* as a promising candidate gene for *qGW10-2* and designated it *GW10.2*.

Three SNPs resulting in missense mutations were detected for *GW10.2* in its exon region (Fig. 3c), leading to the formation of three haplotypes (Supplementary Table. 1). Among them, hap2 was the predominant haplotype, displaying a significantly larger grain width than the other two haplotypes across all six panels. Additionally, we investigated the frequency distribution of these three haplotypes in

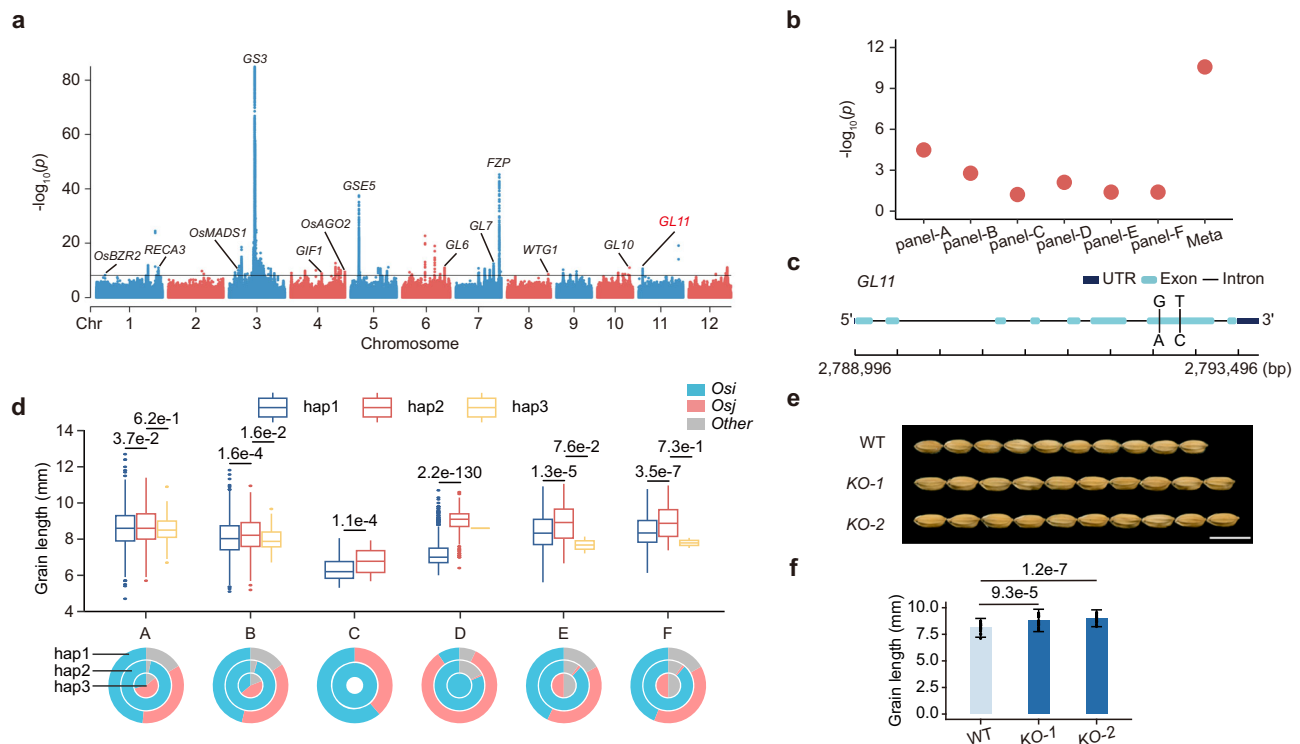
the *indica* and *japonica* subgroups. The result revealed that the hap2 had a notably higher frequency averaging 95.14% in *japonica* rice, only 1.75% in *indica* rice and 3.12% in other rice (Fig. 3d). This suggests that hap2 plays a role in the divergence of grain size between the two subspecies in rice.

The function of *GW10.2* was further verified by using a chromosome segment substitution line (CSSL) population (PA64s  $\times$  Nipponbare) and mutant plants. From a CSSL population reported in our previous study<sup>25</sup>, we selected and evaluated a CSSL (CSSL-2<sup>hap2</sup>) containing the hap2 of *GW10.2* in a Nipponbare-derived fragment and a CSSL (CSSL-1<sup>hap3</sup>) containing the hap3 in a PA64s-derived fragment (Supplementary Fig. 12). The CSSL-2<sup>hap2</sup> showed a significantly larger (8.56%) grain width than that of CSSL-1<sup>hap3</sup> (Fig. 3e, f). Furthermore, we utilized the clustered regularly interspaced palindromic repeats/CRISPR-associated protein 9 (CRISPR/Cas9) system to knock out *GW10.2* in Zhonghua11 (ZH11), resulting in homozygous individuals with a *GW10.2* mutation (Supplementary Fig. 13a). The mutant plants exhibited a significant reduction in grain width, with decreases of 4.3% for those cultivated in Shenzhen (2023) and 6.4% for those grown in Hangzhou (2024), compared to the wild type (Fig. 3g, h; Supplementary Fig. 14). These findings confirm that *GW10.2* could play a positive role in regulating grain width in rice.

### Identification and validation of the candidate gene for grain length

Among the 51 grain-length-associated QTLs examined, 43 were novel and were exclusively detected through meta-analysis (Fig. 4a). A novel QTL *qGL11-1* for further analysis using the same strategy as above for





**Fig. 4 | Meta-GWAS for grain length and its functional analysis.** **a** Manhattan plot for meta-analysis of genome-wide association study of grain length. The genes labeled in the figure are the known grain length genes associated with meta-GWAS, and the red one is the identified candidate gene. The black line represents the genome-wide significance threshold at  $7.52 \times 10^{-9}$ . **b**  $P$ -values of Chr11:262726 in GWAS and meta-analysis. **c** Gene structure, showing 2 missense mutation SNPs in the exon region. **d** Boxplot showing the grain length at different haplotypes in each panel. Statistical differences according to two-tailed Student's  $t$ -test. The pie chart

displays the subpopulation distribution of the three haplotypes in each panel (*Osi*, *Oryza sativa indica*; *Osj*, *Oryza sativa japonica*). **e** Difference in grain length between wild type ( $n = 19$ ) and mutants ( $n = 19$ ). Scale bars: 1 cm. Values are presented as means  $\pm$  SD in **f**, Student's  $t$ -test. In **d**, the horizontal line represents the median value, the lower and upper quartiles represent the 25th and 75th percentile, and the whiskers show the maximum and minimum values. Source data are provided as a Source Data file.

grain width (Supplementary Data. 5), with its lead variant being Chr11:262726 ( $A > G$ ,  $p = 2.66 \times 10^{-11}$ ), showing relatively low significance across all panels (Fig. 4b). Employing the same strategy as previously described, we obtained five genes within the 200 kb region surrounding the variants (*LOC\_Os11g05480*, *LOC\_Os11g05552*, *LOC\_Os11g05570*, *LOC\_Os11g05640*, *LOC\_Os11g05930*). Analysis of co-expression networks of these five genes indicated that only *LOC\_Os11g05930* was co-expressed with three known genes (*OsPRR1*, *OsNST1*, and *OsFKF1*)<sup>26–28</sup> involved in regulating grain development. Consequently, *LOC\_Os11g05930*<sup>29</sup> was considered the candidate gene for *qGL11-1* and was designated *GL11*.

Two SNPs resulting in missense mutations were detected for *GL11* in its exon region (Fig. 4c), forming three haplotypes in all accessions, with hap1 and hap2 being the predominant ones (Supplementary Table. 1). Across all panels, hap2 emerged as the dominant haplotype, exhibiting a significantly larger grain length (average 4.42%) compared to the other two haplotypes across all six panels. Moreover, hap2 had a markedly higher frequency of 91.83% in *indica* rice, only 0.79% in *japonica* rice and 7.39% in other rice accessions (Fig. 4d).

To further validate the biological function of *GL11*, we created knock-out mutants for *GL11* by using CRISPR/Cas9 technology (Supplementary Fig. 13b). These mutant plants showed a notable increase (9.83%) in grain length compared with the wild type (Fig. 4e, f), indicating that *GL11* functions as a negative regulator of rice grain length.

## Discussion

Rice yield is typically influenced by factors such as the effective tiller number, grains per panicle, and grain weight<sup>30</sup>. Grain size is closely associated with grain weight and overall rice quality<sup>31</sup>. Over the past

decade, numerous genes that regulate grain size have been identified. However, as grain size and rice quality are typical quantitative trait regulated by multiple genes, the molecular regulatory mechanisms involving most of these genes remain poorly understood<sup>32</sup>. Genome-wide association studies (GWAS), which have been widely employed for decades<sup>33–35</sup>, have limitations in mining genes from populations with limited size and complex population structures, particularly for complex phenotypes such as yield-related traits. Owing to insufficient statistical power and the influence of complex environmental factors, GWAS can explain only a small portion of the missing heritability<sup>36</sup>. As previously reported, meta-analysis facilitated genetic analysis across different populations, enhancing the identification of complex traits<sup>37</sup>. Furthermore, it enhances the detection rate of rare mutations and subsequently improve their contribution to phenotypic heritability<sup>38</sup>. Thus, the meta-analysis serves as an effective tool for these endeavors.

In this study, we systematically applied and comprehensively evaluated the meta-GWAS strategy for mining genes associated with important agronomic traits in rice. We identified 156 QTLs, 116 of which are being novel and present valuable genetic resources for mining new favorable genes in rice. Two of these novel QTLs were chosen for further analysis, leading to the identification of two candidate genes, *GW10.2* and *GL11* related to grain size. Prior research has indicated that auxin can regulate cell size, thereby influencing grain size<sup>39</sup>. *GW10.2* is involved in biological processes such as the regulation of the auxin signaling pathway and cell elongation, suggesting that *GW10.2* might modulate grain width through the auxin pathway. Numerous studies have highlighted that several genes, including *OsGATA6* and *Ghd7*<sup>40,41</sup>, can coordinate the regulation of heading time and grain size. *GL11* may be a novel factor in coordinating the regulation of grain length in rice.

Additionally, previous studies have shown that *GW10.2* can regulate leaf pillow development, leaf inclination and ear development<sup>42</sup> and *GL11* inhibits heading under long day conditions and promotes heading under short day conditions<sup>29</sup>. Consequently, we hypothesize that both *GW10.2* and *GL11* act as pleiotropic genes for different important agronomic traits and can provide useful gene resources for the molecular design of rice breeding.

On the basis of our findings, several advantages of meta-analysis in our study are evident: (1) integration of rich phenotypic data and a large sample size from diverse populations; (2) attainment of a more robust and precise statistical effect through the amalgamation of multiple independent experiments, particularly for variants with low frequency; (3) facilitation of a more thorough and systematic analysis of complex agronomic traits; and (4) significant cost savings in sequencing and phenotypic management by leveraging the abundant genotypic and phenotypic datasets amassed in rice research. Furthermore, we also observed that meta-GWAS results from multiple panels proved highly effective in detecting QTLs, underscoring the importance of collecting as many GWAS dataset as possible to fully utilize the advantages of meta-analysis. To date, the field of rice GWAS has been abundant<sup>43–45</sup>, which have established a good foundation for subsequent meta-analyses. Our study can thus serve as a research model for applying meta-GWAS in rice population genetics and gene mining studies. Nevertheless, given the filtering of low-frequency variant sites (e.g., MAF < 0.05) in the GWAS workflow, GWAS and its meta-analysis based on natural populations may have difficulty effectively identifying rare QTLs within populations. Moreover, parameters such as the selection of the threshold for *P*-values of associated loci and the size and composition of the population, which affect the efficiency of GWAS analysis, are crucial factors influencing meta-analysis.

In summary, we conducted a meta-GWAS study encompassing six rice agronomic trait by integrating genotypic and phenotypic datasets from six rice panels. This effort led to the identification of 156 potential QTLs, in which two QTLs related to grain size were selected for in-depth analysis. Through validation experiments, we screened and validated two functional genes associated with grain width and grain length, denoted as *GW10.2* and *GL11*, respectively. Our study demonstrates the successful application of meta-GWAS analysis strategies in gene mining from multiple population panels to glean valuable gene insights for modern rice breeding. Furthermore, this research provides a novel approach to fully utilize the wealth of rice genotype and phenotypic datasets amassed over recent decades, significantly propelling efforts to elucidate pivotal agronomic traits in rice.

## Methods

### Plant materials and growth conditions

The 202 rice accessions of panel-C were grown in Shenzhen, China (N22°32', E114°03'). Seeds were sown on 5 July 2022. At 25 days after sowing, the seedlings were transplanted into a five-row plot with 5 individuals in each row at a spacing of 25 cm × 25 cm. The field management practices, including irrigation, fertilization, and pest control, adhere to standard agricultural protocols. The grain traits, including grain width, grain length and thousand-grain weight were measured in the laboratory after harvest. The phenotypes of other five panels were as described in their original refs. 15–19, and we integrated the location and year of planting in Supplementary Table 2. All phenotypic data can be found in the Source Data.

### Variant calling

We collected raw short sequencing data and phenotypic data from six independent experiments. Trimmomatic<sup>46</sup> was used for quality control of the short sequencing data. (Version: 0.39, parameter: MINLEN: 75 LEADING: 20 TRAILING: 20 SLIDINGWINDOW: 5:20; MINLEN = 40, while the read length is <75 bp). BWA<sup>47</sup> was used to map the quality

control data to the Nipponbare genome (MSU v7.0)<sup>48</sup>. (Version: v.0.7.17-r1188, default parameter). SNP calling was subsequently conducted via Sentieon pipeline<sup>49</sup> (Version: genomics-202112.02, default parameter). Based on the pan-genome we constructed in our previous studies<sup>4,13</sup>, PAV calling was conducted with quality control data and the pan-genome by using VG Giraffe<sup>50</sup> and SURVIVOR (Version: v.1.0.7, parameters: 1000 11 -1 -1 50). PAVs of low-quality or unexpected length (>1 Mb or <50 bp) were removed. Quality control of removed PAVs was performed using VCFtools<sup>51</sup> for subsequent analysis. (Version: 0.1.16, parameter --maf 0.05).

### SNP quality control and genotype imputation

We used VCFtools<sup>51</sup> to filter the SNPs with same standard for each panel. (Version: 0.1.16, parameter: --maf 0.05 --max-missing 0.9). Then, for three panels (Panel-D, Panel-E, Panel-F) with low sequencing depth (average 7× or below)<sup>52</sup>, we used an online method for genotype imputation ([https://gonglab.hzau.edu.cn/Plant\\_imputeDB/](https://gonglab.hzau.edu.cn/Plant_imputeDB/))<sup>53</sup> after quality control to ensure the accuracy of the imputation. In the end, we obtained high quality SNPs for each panel, panel-A, 3,038,707; panel-B, 2,890,946; panel-C, 4,786,218; panel-D, 3,353,170; panel-E, 4,325,815 and panel-F, 4,321,409. A total of 6,604,898 high-quality SNPs were eventually retained for meta-analysis.

### Genome-wide association analysis

A dataset of SNPs and PAVs was used for GWAS analysis. Principal component analysis (PCA) was conducted by PLINK<sup>54</sup> (Version: v1.90b6.21, parameters: --pca 5). Kinship was conducted by GEMMA<sup>55</sup> (version: gemma-0.98.5-linux-static-AMD64, parameters: -gk 2). The first five principal components and standardized matrix of kinship were used as covariates. GWAS was performed using a mixed linear model in genome-wide efficient mixed model association software (GEMMA, parameters: lmm 1 -k -c). The threshold for GWAS was calculated using Bonferroni<sup>56</sup> test (0.05/n, where n is the total number of variants). CMAplot R-package<sup>57</sup> was used to plot Manhattan plots and QQ plots to show the results of all GWSA results.

### Meta-analysis

We performed an independent meta-analysis of all phenotypes by collecting the GWAS summary data available from each panel. Genome inflation is a frequent occurrence in association analysis. There are many reasons for genome inflation, such as population structure, cryptic relatedness and others<sup>58</sup>, and the GenABEL R-package is used to calculate the genome inflation factor<sup>59</sup>. Meta-analysis was performed using the inverse variance-weighted fixed-effect model in METAL<sup>60</sup> with genomic control<sup>60</sup>. Genome inflation control was performed via the following equation:

$$SE_{corrected} = SE \times \sqrt{\lambda} \quad (1)$$

where SE denotes the standard deviation obtained via GWAS, and where  $\lambda$  denotes the genome inflation factor for each GWAS finding.

### Significant sites and QTLs statistics

The genotype frequencies of the significant variants obtained from GWAS and meta-analysis of all phenotypes were calculated with VCFtools<sup>51</sup> (parameter: --freq). The QTL was defined as a chromosomal region where adjacent pairs of significant variants were <200 kb from each other, according to the average LD decay of six panels (195.5 kb, Supplementary Fig. 15) estimated by plink (Version: v1.90b6.21, parameters: --r2 -ld-window-kb 1000 -ld-window 99999 --ld-window-r2 0). Within each QTL, the most significant variant was taken as the lead variant.

## Heritability estimation

Based on the lead variant in each QTL, all the variants and phenotypes in the upstream and downstream 200 kb of the lead variant were used to estimate heritability. Plink<sup>54</sup> was used for principal component analysis of the obtained variants, and the first five principal components were used as covariables for heritability estimation. The GRM matrix is then constructed using GCTA software<sup>61</sup>. (Version: gcta-1.94.1, parameter: --make-grm-alg 1 --make-grm) GCTA was used to calculate the heritability of QTLs obtained by GWAS and meta-analysis for each panel, respectively. (parameter: --reml --qcovar --grm). Heritability of QTLs was performed via the following equation:

$$h^2 = V_G / V_P \quad (2)$$

where  $V_G$  is the genotypic variance and  $V_P$  is the phenotypic variance.

## Construction of CRISPR/Cas9 vectors and transformation

The guide RNA (sgRNA) sequences were designed according to the predictions of the CRISPR-P program<sup>62</sup>. The construction of the CRISPR/Cas9 plant expression vector VK005-01 followed the procedures provided by the company's reagent kit (Sbigen Biotech, Beijing, China). The Cas9/sgRNA constructs were transformed into *Agrobacterium* strain EHA105. Two conventional *japonica* rice varieties Xiushui134 (XS134) and ZhongHua11 (ZH11) were used in this study. Rice callus tissue of XS134 and ZH11 were transformed with the *agrobacterium* strains containing the Cas9/sgRNA constructs as previously described<sup>63</sup> to generate transgenic rice plants. Transgene-free plants were grown in a standard greenhouse and in a paddy field at the Shenzhen experimental station (N22°32', E114°03'), Guangdong Province, China and Hangzhou (N30°18', E120°12'), Zhejiang Province, China. Transgenic rice plants were grown in a standard greenhouse.

## Primers

All primers used in this study are provided in Supplementary Table 3.

## Genotype and phenotype analysis of wild and mutant type of rice plants

Genotyping was performed for mutations in the target genes of the transgenic plants. The PCR products were sequenced via the Sanger method. Decoding of the sequencing chromatograms was carried out as described previously<sup>64</sup>. DNA sequences were aligned using Clustal<sup>65</sup>. Agronomic traits were studied in the wild-type parent and T1 generation plants. Rice plants were grown at the Shenzhen experimental station, Guangdong Province, under natural conditions. Plant management followed normal agricultural practices.

## Reporting summary

Further information on the research design is available in the Nature Portfolio Reporting Summary linked to this article.

## Data availability

The sequencing data for panel-A and panel-B are available through NCBI project number [PRJEB6180](#). Genome sequencing data of panel-C accessions can be obtained in the NCBI Sequence Read Archive under BioProjects [PRJNA656318](#) and [PRJNA692836](#). All raw reads generated for rice accessions of panel-D can be obtained from the National Genomics Data Center with BioProject accession [PRJCA000322](#). The raw Illumina sequencing data of panel-E and panel-F could be downloaded from NCBI under accession number [PRJNA171289](#). The pan-genome data has been placed in the RiceSuperPIR database ([ricesuperpir.com/web/rspvm](http://ricesuperpir.com/web/rspvm)). Details of all rice accessions in this study are shown in Supplementary information Data 1. All phenotypes of six panels can be accessed in the Source Data. The source data underlying Fig. 2a–d; Fig. 3b, d–h; Fig. 4b, d, f; Figs. S1–S2; Figs. S9–S11 and Figs.

S14–15 are provided as a Source Data file. Additional datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request. Source data are provided with this paper.

## Code availability

All the data were analyzed with standard programs and packages, as detailed above. The code is available on request.

## References

- Gnanamanickam S. S. Rice and its importance to human life. In *Biological Control of Rice Diseases* 1–11 (Springer, Dordrecht. 2009).
- He, W. et al. Widespread inversions shape the genetic and phenotypic diversity in rice. *Sci. Bull.* **69**, 593–596 (2024).
- Lv, Y. et al. A centromere map based on super pan-genome highlights the structure and function of rice centromeres. *J. Integr. Plant Biol.* **66**, 196–207 (2024).
- Shang, L. et al. A super pan-genomic landscape of rice. *Cell Res.* **32**, 878–896 (2022).
- Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
- Si, L. et al. OsSPL13 controls grain size in cultivated rice. *Nat. Genet.* **48**, 447–456 (2016).
- Sun, X. M. et al. Natural variation of *DROT1* confers drought adaptation in upland rice. *Nat. Commun.* **13**, 4265 (2022).
- Wang, Q., Tang, J., Han, B. & Huang, X. Advances in genome-wide association studies of complex traits in rice. *Theor. Appl. Genet.* **133**, 1415–1425 (2019).
- Yengo, L. et al. A saturated map of common genetic variants associated with human height. *Nature* **610**, 704–712 (2022).
- Zhao, J. et al. Meta-analysis of genome-wide association studies provides insights into genetic control of tomato flavor. *Nat. Commun.* **10**, 1534 (2019).
- Bouwman, A. C. et al. Meta-analysis of genome-wide association studies for cattle stature identifies common genes that regulate body size in mammals. *Nat. Genet.* **50**, 362–367 (2018).
- Sun, C. et al. RPAN: rice pan-genome browser for ~3000 rice genomes. *Nucleic Acids Res.* **45**, 597–605 (2017).
- Wang, T. et al. A rice variation map derived from 10 548 rice accessions reveals the importance of rare variants. *Nucleic Acids Res.* **51**, 10924–10933 (2023).
- Chen, J. X. et al. CropGS-Hub: a comprehensive database of genotype and phenotype resources for genomic prediction in major crops. *Nucleic Acids Res.* **52**, D1519–D1529 (2024).
- Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
- Niu, Y. et al. Identification and allele mining of new candidate genes underlying rice grain weight and grain shape by genome-wide association study. *BMC Genomics* **22**, 602 (2021).
- Li, X. et al. Analysis of genetic architecture and favorable allele usage of agronomic traits in a large collection of Chinese rice accessions. *Sci. China Life Sci.* **63**, 1688–1702 (2020).
- Chen, W. et al. Comparative and parallel genome-wide association studies for metabolic and agronomic traits in cereals. *Nat. Commun.* **7**, 12767 (2016).
- Xie, W. et al. Breeding signatures of rice improvement revealed by a genomic variation map from a large germplasm collection. *Proc. Natl. Acad. Sci. U.S.A.* **112**, E5411–E5419 (2015).
- Yu, Y., Zhang, H., Long, Y., Shu, Y. & Zhai, J. Plant Public RNA-seq database: a comprehensive online database for expression analysis of ~45 000 plant public RNA-Seq libraries. *Plant Biotechnol. J.* **20**, 806–808 (2022).
- Hibara, K.-i et al. The *ADAXIALIZED LEAF1* gene functions in leaf and embryonic pattern formation in rice. *Dev. Biol.* **334**, 345–354 (2009).



22. Duan, Y. et al. *RETINOBLASTOMA-RELATED* genes specifically control inner floral organ morphogenesis and pollen development in rice<sup>1</sup>. *Plant Physiol.* **181**, 1600–1614 (2019).
23. Zeng, Z. et al. Overexpression of a glyoxalase gene, *OsGly I*, improves abiotic stress tolerance and grain yield in rice (*Oryza sativa* L.). *Plant Physiol. Biochem.* **109**, 62–71 (2016).
24. Zhao, D.-S. et al. *GS9* acts as a transcriptional activator to regulate rice grain shape and appearance quality. *Nat. Commun.* **9**, 1 (2018).
25. Zhang, B. et al. Development of three sets of high-throughput genotyped rice chromosome segment substitution lines and QTL mapping for eleven traits. *Rice* **12**, 33 (2019).
26. Han, S. H., Yoo, S. C., Lee, B. D., An, G. & Paek, N. C. Rice FLAVIN-BINDING, KELCH REPEAT, F-BOX 1 (*OsFKF1*) promotes flowering independent of photoperiod. *Plant Cell Environ.* **38**, 2527–2540 (2015).
27. Zhang, B. C. et al. Golgi nucleotide sugar transporter modulates cell wall biosynthesis and plant growth in rice. *Proc. Natl Acad. Sci. USA* **108**, 5110–5115 (2011).
28. Wang, F. et al. The rice circadian clock regulates tiller growth and panicle development through strigolactone signaling and sugar sensing. *Plant Cell* **32**, 3124–3138 (2020).
29. Zhang, J. et al. Global analysis of CCT family knockout mutants identifies four genes involved in regulating heading date in rice. *J. Integr. Plant Biol.* **63**, 913–923 (2021).
30. Li, G. M., Tang, J. Y., Zheng, J. K. & Chu, C. C. Exploration of rice yield potential: decoding agronomic and physiological traits. *Crop J.* **9**, 577–589 (2021).
31. Lu, X. D. et al. Grain shape genes: shaping the future of rice breeding. *Rice Sci.* **30**, 379–404 (2023).
32. Ren, D., Ding, C. & Qian, Q. Molecular bases of rice grain size and quality for optimized productivity. *Sci. Bull.* **68**, 314–350 (2023).
33. Yano, K. et al. GWAS with principal component analysis identifies a gene comprehensively controlling rice architecture. *Proc. Natl Acad. Sci. USA* **116**, 21262–21267 (2019).
34. Yoshida, H. et al. Integrated genome-wide differentiation and association analyses identify causal genes underlying breeding-selected grain quality traits in japonica rice. *Mol. Plant* **16**, 1460–1477 (2023).
35. Ruan, B. P. et al. Natural variation in the promoter of *TGW2* determines grain width and weight in rice. *N. Phytol.* **227**, 629–640 (2020).
36. Tam, V. et al. Benefits and limitations of genome-wide association studies. *Nat. Rev. Genet.* **20**, 467–484 (2019).
37. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
38. Marouli, E. et al. Rare and low-frequency coding variants alter human adult height. *Nature* **542**, 186–190 (2017).
39. Qiao, J. Y. et al. A novel miR167a-OsARF6-OsAUX3 module regulates grain length and weight in rice. *Mol. Plant* **14**, 1683–1698 (2021).
40. Zhang, Y. J. et al. The transcription factor *OsGATA6* regulates rice heading date and grain number per panicle. *J. Exp. Bot.* **73**, 6133–6149 (2022).
41. Xue, W. Y. et al. Natural variation in *Ghd7* is an important regulator of heading date and yield potential in rice. *Nat. Genet.* **40**, 761–767 (2008).
42. Chen, S. -H., Zhou, L. -J., Xu, P., Xue, H. -W. SPOC domain-containing protein Leaf inclination3 interacts with LIP1 to regulate rice leaf inclination through auxin signaling. *Plos Genet.* **14**, e1007829 (2018).
43. Su, J. et al. Genome-wide association study and Mendelian randomization analysis provide insights for improving rice yield potential. *Sci. Rep.* **11**, 6894 (2021).
44. McCouch, S. R. et al. Open access resources for genome-wide association mapping in rice. *Nat. Commun.* **7**, 10532 (2016).
45. Yoshida, H. et al. Genome-wide association study identifies a gene responsible for temperature-dependent rice germination. *Nat. Commun.* **13**, 5665 (2022).
46. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
47. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
48. Kawahara, Y. et al. Improvement of the *Oryza sativa* Nipponbare reference genome using next generation sequence and optical map data. *Rice* **6**, 4 (2013).
49. Kendig, K. I. et al. Sentieon DNaseq variant calling workflow demonstrates strong computational performance and accuracy. *Front. Genet.* **10**, 736 (2019).
50. Sirén, J. et al. Pangenomics enables genotyping of known structural variants in 5202 diverse genomes. *Science* **374**, 6574 (2021).
51. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
52. Huang, J. et al. Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat. Commun.* **6**, 8111 (2015).
53. Gao, Y. et al. Plant-ImputeDB: an integrated multiple plant reference panel database for genotype imputation. *Nucleic Acids Res.* **49**, D1480–D1488 (2021).
54. Purcell, S. et al. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
55. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat. Genet.* **44**, 821–824 (2012).
56. Curtin, F. & Schulz, P. Multiple correlations and Bonferroni's correction. *Biol. Psychiatry* **44**, 775–777 (1998).
57. Yin, L. L. et al. rMVP: A memory-efficient, visualization-enhanced, and parallel-accelerated tool for genome-wide association study. *Genomics Proteom. Bioinf.* **19**, 619–628 (2021).
58. Yang, J. et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.* **19**, 807–812 (2011).
59. Aulchenko, Y. S., Ripke, S., Isaacs, A. & Van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* **23**, 1294–1296 (2007).
60. Willer, C. J., Li, Y. & Abecasis, G. R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190–2191 (2010).
61. Yang, J. A., Lee, S. H., Goddard, M. E. & Visscher, P. M. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.* **88**, 76–82 (2011).
62. Lei, Y. et al. CRISPR-P: A web tool for synthetic single-guide RNA design of CRISPR-system in plants. *Mol. Plant* **7**, 1494–1496 (2014).
63. Hiei, Y., Ohta, S., Komari, T. & Kumashiro, T. Efficient transformation of rice (*Oryza sativa* L.) mediated by *Agrobacterium* and sequence analysis of the boundaries of the T-DNA. *Plant J.* **6**, 271–282 (2003).
64. Liu, W. Z. et al. DSDecode: A web-based tool for decoding of sequencing chromatograms for genotyping of targeted mutations. *Mol. Plant* **8**, 1431–1433 (2015).
65. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega. *Mol. Syst. Biol.* **7**, 539 (2011).

## Acknowledgements

This research was supported by the Natural Science Foundation of China (32188102 to Q.Q. and 32372148 to L.S.), the National Key R&D Program of China (2023ZD04073 to L.S.) and Innovation Program of Chinese Academy of Agricultural Sciences, the Youth Innovation of Chinese Academy of Agricultural Sciences (Y2023QC36 to L.S.).



## Author contributions

L.S. and Q.Q. conceived and directed the project. L.Y. and W.H. conducted manuscript writing and editing. Y.Lv., H.W. and Q.Z. screened the literature and collected the data. Y.Li., Y.Liu and Zhiyuan Zhang designed and coordinated field experiments. T.W., Bin Zhang, W.C. and H.H. performed data quality control, comparison and variation detection. X.W., D.C., C.L. and C.S. performed the GWAS analysis and draw the figure. X.Liu, Q.X., Q.Y. and X.Y. conducted the meta-analysis. H.Q., X.Li and Binta Zhang screened significant sites and statistically analyzed H.Z., Y.Leng, Zhipeng Zhang, X.D. and M.G. helped accomplish the heritability estimation. Y.Z., X.C., J.J. and Y.C. completed the construction and transformation of the vector.

## Competing interests

The authors declare no competing interests.

## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-58081-1>.

**Correspondence** and requests for materials should be addressed to Qian Qian or Lianguang Shang.

**Peer review information** *Nature Communications* thanks Reddyamini Bommisetty and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>

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

**Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025