nature genetics

Article

https://doi.org/10.1038/s41588-023-01454-3

Suppressing a phosphohydrolase of cytokinin nucleotide enhances grain yield in rice

Received: 11 August 2022

Accepted: 21 June 2023

Published online: 27 July 2023

Check for updates

Bi Wu[1,2], Jianghu Meng[1], Hongbo Liu[1], Donghai Mao[1], Huanran Yin[1], Zhanyi Zhang[1], Xiangchun Zhou[1], Bo Zhang[1], Ahmed Sherif[1], Haiyang Liu[3], Xianghua Li[1], Jinghua Xiao[1], Wenhao Yan[1,2], Lei Wang[1,2], Xingwang Li[1,2], Wei Chen[1,2], Weibo Xie[1,2], Ping Yin[1,2], Qifa Zhang[1,2] & Yongzhong Xing[1,2]

One-step and two-step pathways are proposed to synthesize cytokinin in plants. The one-step pathway is mediated by LONELY GUY (LOG) proteins. However, the enzyme for the two-step pathway remains to be identified. Here, we show that quantitative trait locus *GY3* may boost grain yield by more than 20% through manipulating a two-step pathway. Locus *GY3* encodes a LOG protein that acts as a 5′-ribonucleotide phosphohydrolase by excessively consuming the cytokinin precursors, which contrasts with the activity of canonical LOG members as phosphoribohydrolases in a one-step pathway. The residue S41 of GY3 is crucial for the dephosphorylation of iPRMP to produce iPR. A *solo-LTR* insertion within the promoter of *GY3* suppressed its expression and resulted in a higher content of active cytokinins in young panicles. Introgression of $GY3^{02428}$ increased grain yield per plot by 7.4% to 16.3% in all investigated *indica* backgrounds, which demonstrates the great value of $GY3^{02428}$ in *indica* rice production.

Owing to their high relative abundance and affinity to their receptors, the active cytokinin types N^6-(Δ^2-isopentenyl) adenine (iP) and *trans*-zeatin (tZ) play critical roles in regulating cell proliferation and differentiation, especially during inflorescence meristem development[1–3]. In particular, tZ-type cytokinins play major roles in aerial organ development[4–6]. Two models have been proposed for isoprenoid cytokinin release from cytokinin nucleotides: the one-step pathway mediated by LOG activation[7] and the two-step cytokinin synthesis pathway[8,9]. In the one-step pathway, cytokinin nucleobases are directly generated from their corresponding cytokinin nucleotides, which has been verified for LOG family proteins acting as cytokinin nucleoside 5′-monophosphate phosphoribohydrolases in rice, *Arabidopsis thaliana* and plant-interacting organisms[7,10–15]. The two-step pathway proposes that cytokinin nucleotides such as iP riboside 5′-monophosphate (iPRMP) are successively converted to corresponding cytokinin nucleosides such as iP riboside (iPR) by nucleotidase and then to cytokinin nucleobases (such as iP) by nucleosidase. However, knowledge about the two-step pathway is very limited. Phylogenetic analysis of 123 LOG proteins in both plants and plant-interacting organisms revealed two clusters (type I and type II) and synapomorphic residues at the active site indicated that all LOG proteins are involved in the same catalytic mechanism for cytokinin release[16]. However, the two diverse clusters and further subclusters may produce somewhat different types of cytokinins[17]. Stable isotope-labeled tracers applied to cytokinin analysis showed that the conversion rate from cytokinin nucleosides to cytokinin nucleobases in the two-step pathway is much slower than LOG-catalyzed conversion from cytokinin nucleotides to cytokinin nucleobases[18]. LOG family enzymes have not been previously identified to be involved in the two-step pathway in plants.

[1]National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University, Wuhan, China. [2]Hubei Hongshan Laboratory, Wuhan, China. [3]Hubei collaborative Innovation Center for Grain Industry, Yangtze University, Jingzhou, China. e-mail: yzxing@mail.hzau.edu.cn

https://doi.org/10.1038/s41588-023-01454-3

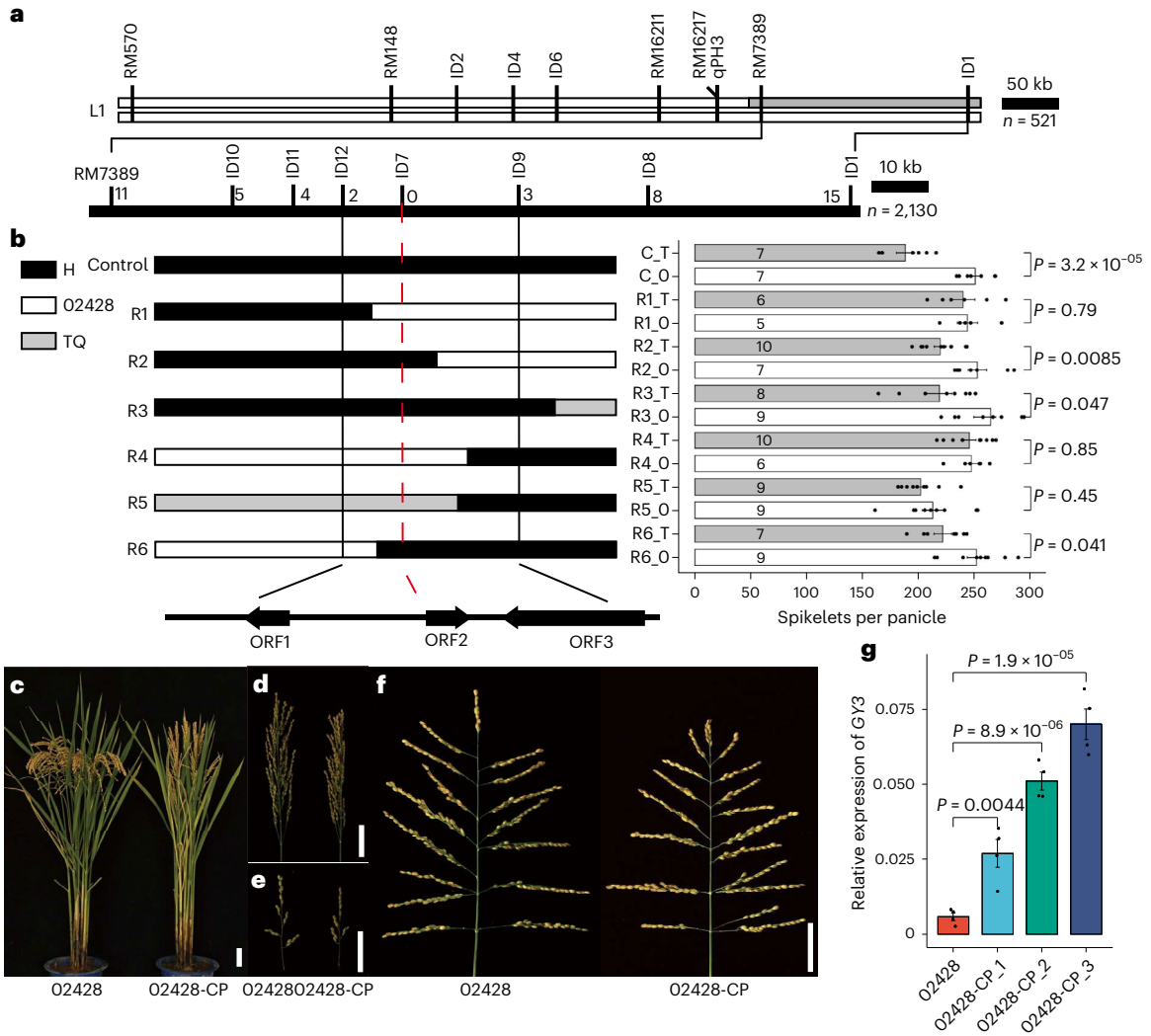


Fig. 1 | Map-based cloning of *GY3*. **a**, Fine mapping of *GY3* via two randomly segregating populations (n = 521 and 2,130, respectively). The numbers near the bars represent the numbers of recombinants between *GY3* and the markers. **b**, Local genomic components of key recombinants and the performance of their progeny in terms of spikelets per panicle (SPP). The black, white and gray bars represent heterozygous, 02428 and TQ homozygous genomic regions, respectively. The control was heterozygous at *GY3* and R1 to R6 represent six recombinants. The SPP of homozygotes TQ (R#_T) and 02428 (R#_0) are shown as the means ± s.e.; three ORFs were identified in the mapping region, of which ORF2 cosegregated with InDel marker ID7; the numbers in the bars indicate the number of individuals used for phenotyping. **c**–**f**, Comparison of the whole plants (**c**), panicles (**d**), primary branches (**e**) and panicle architecture (**f**) at the maturation stage between 02428 and the complementary lines in the background of 02428 (02428-CP). Scale bars, 5 cm. **g**, Relative expression level of *GY3* in the leaves of 02428 and 02428-CP seedlings. The data are the means ± s.e. (four biological and three technical replicates). *P* values were calculated by two-sided paired Student's *t*-test (**b** and **g**).

In our previous study, two tightly linked quantitative trait loci (QTLs), *qPH3* and *GY3*, which control plant height and grain yield per plant, respectively, were fine mapped onto rice chromosome 3 (ref. 19). It was shown that *OsGA2OOX1* underlies *qPH3* (refs. 19,20) but the gene underlying *GY3* has not been identified. By increasing the numbers of secondary branches and spikelets per panicle, *GY3* could increase grain yield per plant by more than 20% (ref. 19). Here, we show that *GY3* encodes a LOG-like protein with a distinctive function in cytokinin synthesis. A *Gypsy solo-LTR* insertion within the promoter of *GY3* suppressed its expression and promoted panicle branching. The introgression of *GY3* into several elite *indica* restorer lines and hybrids demonstrated its usefulness in increasing grain yield.

Results

GY3 encodes LOGL5 and regulates grain yield

To eliminate the genetic interference of *qPH3*, a QTL for plant height which also influences grain yield[19], a near-isogenic line (NIL) for *GY3* in the variety 02428 background was obtained from a BC_5F_2 population. In this line, *qPH3* was fixed for the 02428 allele and *GY3* was fixed for the donor parent allele of Teqing (TQ) (Fig. 1a). The per-plant grain yield of plants with the $GY3^{02428}$ allele (02428) was 6.3 g (21%) higher than that of the plants with the $GY3^{TQ}$ allele (02428-$GY3^{TQ}$), which was mainly caused by ten additional secondary branches resulting in 39 additional grains per panicle (Extended Data Fig. 1a–h). The panicle length, biomass weight per plant and 1,000-grain weight of 02428 were significantly higher than that of 02428-$GY3^{TQ}$ but no differences were observed in heading date, plant height, number of tillers per plant, harvest index, seed setting ratio and number of primary branches (Extended Data Fig. 1i–q). Similarly, introgressing $GY3^{02428}$ in the TQ genetic background (TQ-$GY3^{02428}$) increased grain yield per plant by 7.5 g (29%) compared with TQ (Supplementary Table 1 and Extended Data Fig. 2a–d).

To identify the gene underlying *GY3*, we performed positional cloning. A total of 2,130 plants derived from the self-pollinated progeny of L1 were genotyped with eight markers, which narrowed *GY3* to

a 21 kilobase (kb) interval flanked by ID12 and ID9 and cosegregating with ID7 (Fig. 1b). There were two intact open reading frames (ORFs) and one truncated ORF encoding a *Ty1-copia*-like retrotransposon. *ORF1* (*LOC_Os03g64060*) encodes an expressed unknown protein and *ORF2* (*LOC_Os03g64070*) encodes the LONELY GUY-LIKE 5 (LOGL5) protein (http://rice.uga.edu/), which is predicted to directly catalyze the release of cytokinin nucleobases[7]. Real-time quantitative polymerase chain reaction (qPCR) analysis showed that *LOGL5* was expressed fivefold higher in the leaves of 02428-*GY3*^{TQ} than in the leaves of 02428 (Extended Data Fig. 1r). To confirm the candidacy of *LOGL5*, a nearly 4 kb genomic fragment of the TQ allele, including a 2 kb promoter region, was introduced into 02428 (Extended Data Fig. 3b). *GY3* expression in the transgene-positive plants (02428-CP) increased 3.6–11.1-fold and the phenotype of 02428-CP was similar to that of 02428-*GY3*^{TQ}, with the former presenting a reduction in grain yield from 5.7 to 18.7 g and an average decrease of 32% compared with 02428 (11.7 g) (Fig. 1c–g and Extended Data Fig. 3c–n). Similar results were observed when an overexpression cassette containing the *LOGL5* coding sequence driven by the maize *Ubiquitin* promoter was introduced into 02428 (Extended Data Fig. 4). In addition, *LOGL5* knockout plants in the background of a *japonica* variety Zhonghua 11 significantly increased spikelets per plant by increasing about ten secondary branches (Extended Data Fig. 5 and Supplementary Table 2). These results demonstrated that *LOGL5* is the causal gene underlying *GY3*, which elevates the expression of *LOGL5* thus reducing grain yield.

A *solo-LTR* insertion downregulates *GY3* expression

Comparative sequencing of *GY3* between 02428 and TQ revealed 27 single-nucleotide polymorphisms (SNPs) and eight insertions/deletions (InDels) within the 2 kb promoter region (the TQ allele sequence was used as control), two SNPs and two InDels within the 165 base pair (bp) 5′-untranscribed region, one SNP and one InDel within the 400 bp 3′-untranscribed region and three synonymous SNPs within the coding sequence (Extended Data Fig. 3a). A 3,882 bp insertion was detected at 579 bp upstream of the *GY3* transcriptional start site in 02428 relative to the TQ sequence; this insertion sequence belongs to the typical *Gypsy* subclass of long terminal repeat (LTR) retrotransposons (Fig. 2a). To confirm the regulatory role of the *solo-LTR* insertion in the expression of *GY3*, a knockout plant was generated with a 3,983 bp deletion that included the entire *solo-LTR* (Fig. 2a). The expression level of *GY3* in the homozygous *solo-LTR*-deletion plants (02428$^{LTR\text{-}KO}$) was 5.4-fold higher than that in 02428. Moreover, the per-plant yield of 02428$^{LTR\text{-}KO}$ was equivalent to that of 02428-*GY3*^{TQ}, which was 6.9 g less than that of 02428, a 25% decrease (Fig. 2e–m, Supplementary Fig. 1 and Supplementary Data 1). These results strongly support that the *solo-LTR* insertion within the promoter region of *GY3* is responsible for the high grain yield in 02428 plants.

To further answer how the insertion downregulates *GY3* expression, a bisulfite conversion sequencing PCR (BSP) in the promoter region of *GY3* showed that a higher average methylation level of CHG contexts was detected (59.6%) in 02428 allele than that in 02428-*GY3*^{TQ} (36.3%) but little differences were observed in CG and CHH contexts between genotypes (Fig. 2c). Whole-genome bisulfite sequencing (WGBS) of 02428 and 02428-*GY3*^{TQ} also showed that the CHG contexts are much higher in 02428 than those in 02428-*GY3*^{TQ} (Fig. 2b). A self-reinforcing loop model was confirmed that CHG methylation is associated with a heterochromatin mark, H3K9 dimethylation (H3K9me2) modification in previous study[21]. The enhanced chromatin immunoprecipitation sequencing (eChIP–seq) analysis of H3K9me2 showed a peak with fivefold enrichment in the *solo-LTR* insertion region in 02428 compared with 02428-*GY3*^{TQ} without enrichment (Fig. 2b,d). The eChIP–seq analysis of H3K27ac, which activated gene expression, showed a higher enrichment in the *GY3* gene body region of 02428-*GY3*^{TQ} compared to that of 02428. Accordingly, the repressive histone modification, H3K27me3, was significantly less enriched in 02428-*GY3*^{TQ} than that of 02428 (Fig. 2b,d). Thus, the 3,882 bp insertion enhances DNA methylation level and leads to heterochromatinization and consequently represses the *GY3* expression in 02428 plants, which explained a sixfold higher expression level of *GY3* in the young panicles of 02428-*GY3*^{TQ} than in those of 02428 (Fig. 3a).

GY3 acts as cytokinin 5′-ribonucleotide phosphohydrolases

Alignment of LOG family member sequences from rice, maize and wheat showed that GY3, An-2 (ref. 10) and LOG (ref. 7) shared the same catalytic residues, AMP-binding and prenyl-group binding amino acid residues (Supplementary Fig. 2). The protein sequences of GY3 and An-2 were 59.8% identical. To test whether GY3 regulates cytokinin biosynthesis in vivo, we measured the contents of cytokinins in the young panicles of 02428, 02428-*GY3*^{TQ}, 02428-CP, 02428-OE and 02428$^{LTR\text{-}KO}$ (hereafter, all these lines in which *GY3* expression was higher than that in 02428 are collectively referred to as 'higher-expression lines') (Fig. 3a and Supplementary Fig. 3). As expected, the content of the cytokinin nucleotide iPRMP, the substrate of LOG proteins, was lower in the young panicles of all higher-expression lines compared with 02428 (Fig. 3b); but it was out of expectation that iP was not accumulated in higher-expression lines, except in 02428-*GY3*^{TQ} and 02428-CP (Fig. 3c). However, the contents of cytokinin nucleobases, tZ and dihydrozeatin (DHZ), which are thought to be the products of GY3, were even lower in the higher-expression lines except DHZ in 02428-*GY3*^{TQ} (Fig. 3d and Extended Data Fig. 6a). Moreover, the content of the cytokinin nucleoside iPR was higher in all higher-expression lines than in 02428, whereas the contents of tZ riboside (tZR) and most DHZ riboside (DHZR) were lower except DHZR in 02428-*GY3*^{TQ} (Fig. 3e,f and Extended Data Fig. 6b). In addition, the iPR and tZR contents were lower whereas the iP and tZ contents were higher in leaves of *GY3*-knockout lines than that in the wild type (Extended Data Fig. 5i). These results seemingly contradict the canonical one-step cytokinin-releasing function of LOG proteins, whereby cytokinin nucleobases such as iP and tZ should accumulate in higher-expression lines, whereas the changes in the contents of cytokinin nucleosides such as iPR and tZR should be consistent[7,11,15,18]. Therefore, we proposed that, unlike its homologs that directly generate cytokinin nucleobases, GY3 may instead promote the generation of cytokinin nucleosides. To test this, using its homolog An-2 as a control[10], we evaluated the enzymatic characteristics of GY3. Both An-2 and GY3 react with iPRMP rather than with iP and iPR (Fig. 3g and Extended Data Fig. 6c). As in the previous studies[10,11], the major product of the An-2 catalytic reaction is iP. However, the major product of the reaction of GY3 is iPR coupled with little iP. These results suggested that GY3 acts as a 5′-ribonucleotide phosphohydrolase, which shared the same enzymatic properties as the EC 3.1.3.5 purified from the cytosol of wheat[9]. Analysis of the enzymatic properties of GY3 and An-2 under optimum pH conditions (pH 6.5) showed that both proteins shared a similar affinity to iPRMP. However, the maximum reaction rate at a given amount of enzyme, V_{max}, of GY3 is one-half that of An-2 and the K_{cat} and K_{cat}/K_m, where K_{cat} is the number of substrates converted by each enzyme molecule per minute, of GY3 are one-quarter those of An-2 (Supplementary Table 3). To characterize their relative activities, GY3 and An-2 proteins were mixed together such that a gradient of different ratios was produced. A decreased ratio of iP to iPR along with a reduction in An-2 content was observed (Extended Data Fig. 6d). When the amount of GY3 was approximately sevenfold that of An-2, equivalent amounts of iP and iPR were produced in a single reaction, which corresponds to the fourfold reaction efficiency of An-2 against GY3.

To discern the region responsible for the distinct function between the two proteins, both An-2 and GY3 were divided into three segments at two sites located in two completely conserved regions (Supplementary Fig. 2). Five His-tagged recombinant proteins were purified and their enzymatic reaction products were evaluated. The recombinant proteins I-B-C and I-B-III, which contained segment I (with GY3 I, amino acids 1–70), mainly produced iPR as GY3 did, indicating that both have

https://doi.org/10.1038/s41588-023-01454-3

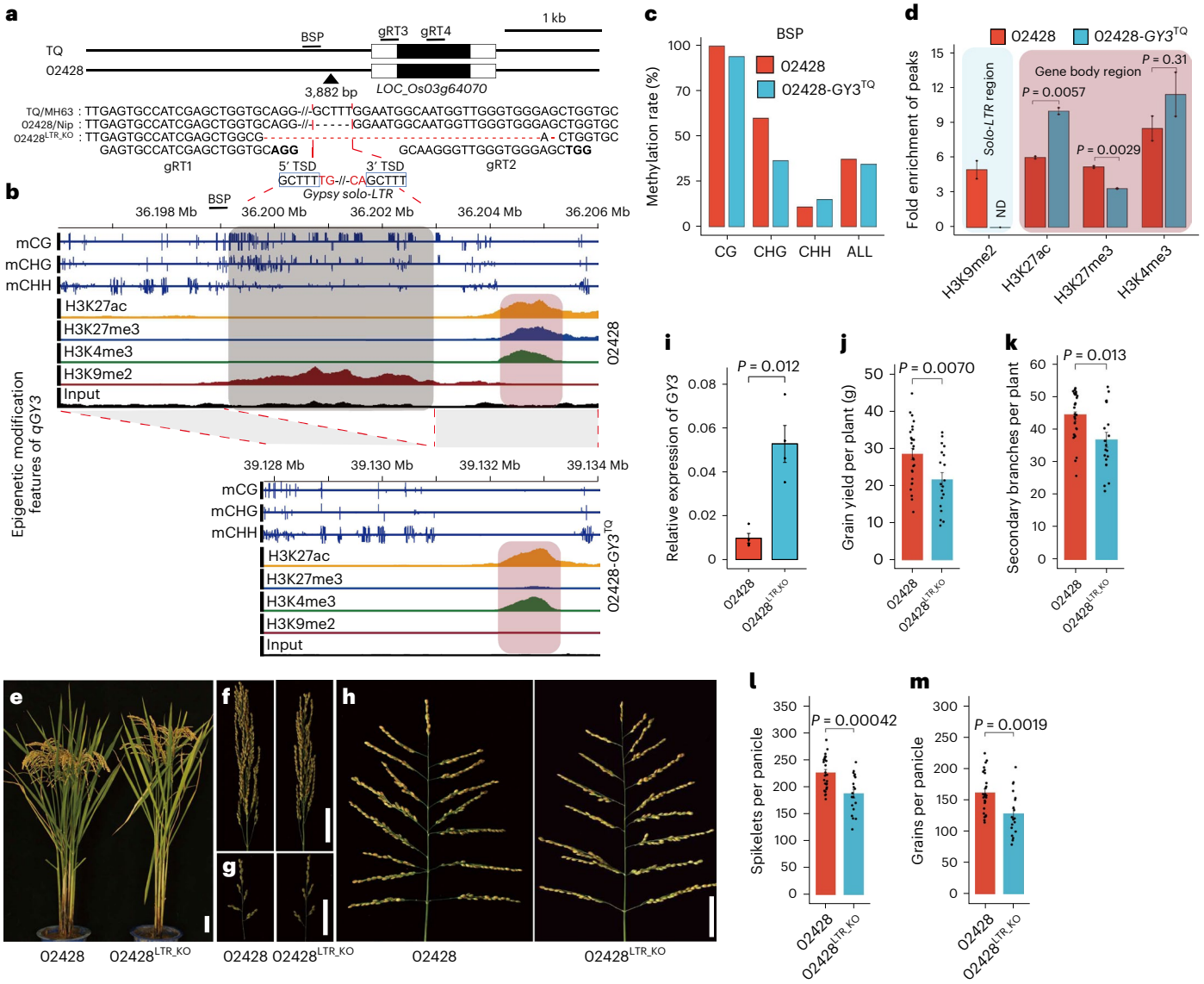


Fig. 2 | Functional identification of the *solo-LTR* insertion within the promoter of *GY3*. **a**, InDel variation of *Gypsy solo-LTR* (*solo-LTR*) within the promoter region of *GY3* between the 02428/Nip and TQ/MH63 alleles. The sequences (GCTTT) in the blue box are the target site duplication (TSD) produced by the *solo-LTR* insertion and a typical LTR sequence termed the TG/CA box (red characters, TG at the 5′ end and CA at the 3′ end of LTR); CRISPR-based knockout line of the *solo-LTR* insertion (02428^{LTR_KO}) in the 02428 background. The red dashed lines indicate deletions; gRT1 and gRT2 are the CRISPR targets for *solo-LTR* knockout in the 02428 background with a protospacer adjacent motif sequence in bold characters. **b**, DNA methylation and chromatin modification features within the promoter and gene body of *GY3* in the young panicles of NILs. **c**, DNA methylation level of BSP region in the young panicles of NILs. **d**, Fold enrichment of H3K9me2 within the *solo-LTR* (light blue) and H3K27ac, H3K4me3 and H3K27me3 within the gene body (red) regions of *GY3* in the young panicles of NILs, respectively ($n = 2$). ND, not determined. **e**–**h**, Comparisons of the whole plants (**e**), panicles (**f**), primary branches (**g**) and panicle architecture (**h**) at the maturation stage between 02428 and 02428^{LTR_KO}. Scale bars, 5 cm. **i**, Relative expression level of *GY3* in the leaves of 02428 and 02428^{LTR_KO} seedlings (four biological and three technical replicates). **j**–**m**, Comparisons of yield per plant (**j**), secondary branches per panicle (**k**), spikelets per panicle (**l**) and grains per panicle (**m**) between 02428 ($n = 27$) and 02428^{LTR_KO} ($n = 19$) lines. The data are the means ± s.e. *P* values were calculated by two-sided paired Student's *t*-test (**d**,**i**–**m**).

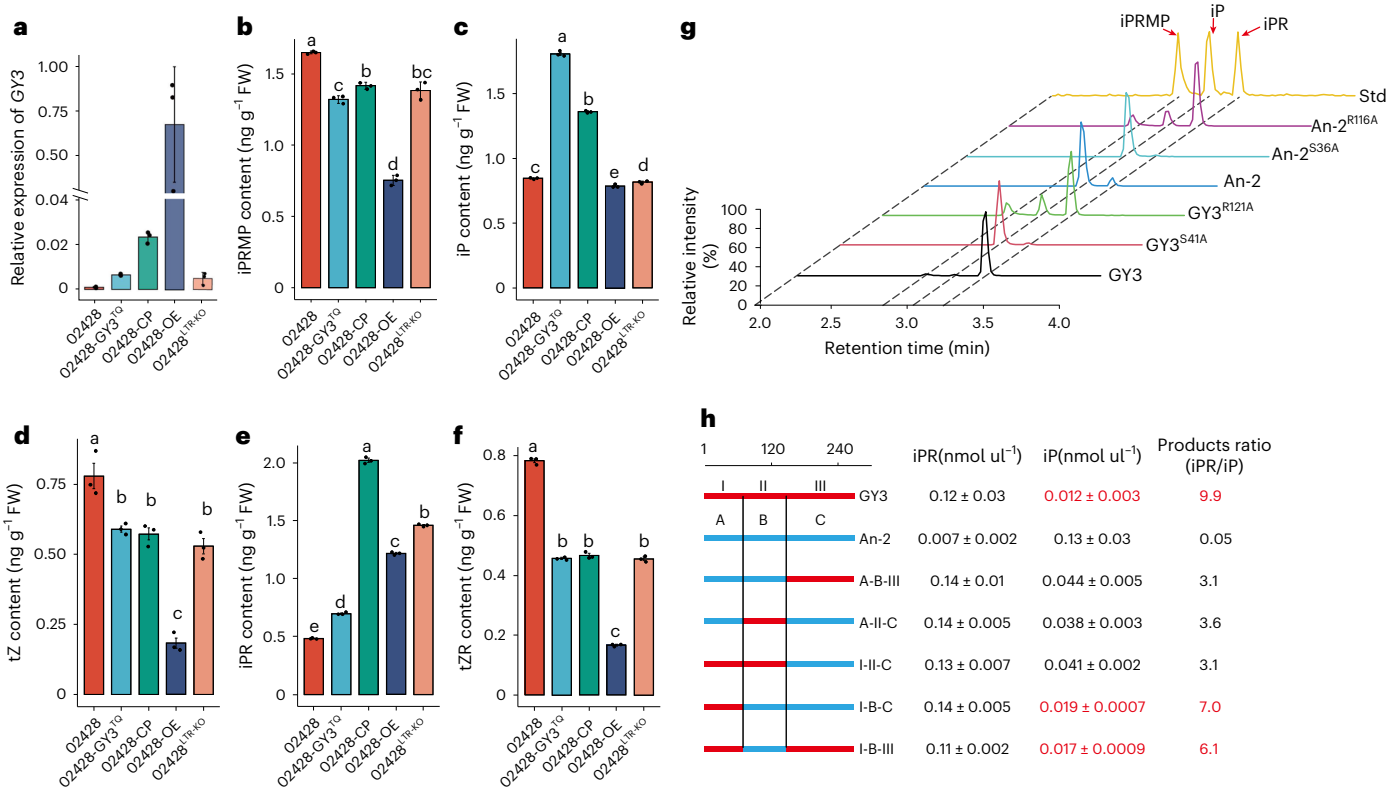
strong phosphohydrolase activity for iPRMP. On the other hand, proteins A-B-III, A-II-C and I-II-C presented reduced phosphoribohydrolase activity in comparison with that of An-2 (Fig. 3h and Extended Data Fig. 6e). Overall, segment I (GY3 I: 1–70) of GY3 is crucial for its function to produce iPR.

Extensive mutagenesis of potential catalytic residues for phosphohydrolase activity revealed that mutants GY3^{S41A} and An-2^{S36A} showed a complete loss of dephosphorylation activity, indicating that the serine residues (S41 in GY3 and S36 in An-2) are crucial for generating iPR (Fig. 3g). The arginine residue, which corresponds to R121 and R116 of GY3 and An-2, respectively, is the candidacy of catalytic residues in LOGs[16] (Supplementary Fig. 2). An-2^{R116A} mutant sharply reduced the phosphoribohydrolase activity whereas the GY3^{R121A} mutant shows a low phosphoribohydrolase activity as does GY3 (Fig. 3g). This indicates that R116 in An-2 is vital for iP production whereas R121 in GY3 is of no use for iP production.

GY3 inhibits cell division in the branching meristem

Expression analysis showed that *GY3* was preferentially expressed in the leaves and young panicles, especially in panicles longer than 5 mm. In all the investigated tissues except leaf sheath, *GY3* was expressed at higher levels in 02428-*GY3*^{TQ} plants than in 02428 plants (Extended Data Fig. 7a). RNA in situ hybridization clearly showed the presence of *GY3* transcripts in the tips of the branching meristems in both 02428 and

https://doi.org/10.1038/s41588-023-01454-3



	iPR(nmol ul^{-1})	iP(nmol ul^{-1})	Products ratio (iPR/iP)
GY3	0.12 ± 0.03	0.012 ± 0.003	9.9
An-2	0.007 ± 0.002	0.13 ± 0.03	0.05
A-B-III	0.14 ± 0.01	0.044 ± 0.005	3.1
A-II-C	0.14 ± 0.005	0.038 ± 0.003	3.6
I-II-C	0.13 ± 0.007	0.041 ± 0.002	3.1
I-B-C	0.14 ± 0.005	0.019 ± 0.0007	7.0
I-B-III	0.11 ± 0.002	0.017 ± 0.0009	6.1

Fig. 3 | Biochemical features of GY3 in association with synthesis of cytokinins. a–**f**, Relative expression level of *GY3* (**a**), contents of endogenous cytokinin nucleotide (iPRMP) (**b**), contents of the cytokinin nucleobases iP (**c**) and tZ (**d**) and contents of the cytokinin nucleosides iPR (**e**) and tZR (**f**) in the young panicles of 02428, 02428-*GY3*^{TQ}, 02428-CP, 02428-OE and 02428$^{LTR\text{-}KO}$. Data are means ± s.e. (three biological and three technical replicates). The different letters above the bars indicate significant differences at $P < 0.01$ via Duncan's test (**b**–**f**). FW, fresh weight. **g**, Detection of the enzyme-catalyzed reaction products of GY3, GY3^{S41A}, GY3^{R121A}, An-2, An-2^{S36A} and An-2^{R116A} to iPRMP. Standards of iPRMP, iP and iPR. **h**, Diagram of the recombinant enzyme and its major enzymatic products. The red and blue lines represent the protein fragments of GY3 (I, II and III) and An-2 (A, B and C), respectively. The letters at the end of the line represent the compositions of the newly recombinant enzymes and the products and fold differences (iP/iPR) represent the major enzyme product and product ratios of iP and iPR of each recombinant enzyme.

02428-*GY3*^{TQ}, with a higher level found in 02428-*GY3*^{TQ} than in 02428 (Extended Data Fig. 7b–h).

The *KNOX* family genes (*OSH1*, *OSH6* and *OSH15*), which are positively regulated by cytokinins, were repressed in the young panicles of the higher-expression lines compared to 02428 (Supplementary Fig. 4a and Supplementary Data 1). *OSH1*, a meristematic activity marker, showed a higher expression level in 02428 than in 02428-*GY3*^{TQ} (Supplementary Fig. 4d,j and Supplementary Data 1). The mitosis regulator *CYCD1;3*, which encodes a D-type cyclin and is involved in the G1-to-S-phase transition by phosphorylation of retinoblastoma-related protein[22], was downregulated in the *GY3* higher-expression lines (Supplementary Fig. 4a and Supplementary Data 1). Moreover, the cell division activity marker *Histone H4* (ref. 23) was expressed more in meristematic cells of branching panicles in 02428 compared with 02428-*GY3*^{TQ} (Supplementary Fig. 4e,f,k,l and Supplementary Data 1). Taken together, all these results implied that cell division in the inflorescence meristems of 02428 is more active than that in 02428-*GY3*^{TQ}.

Three (*OsRR3*, *OsRR5* and *OsRR6*) of nine cytokinin-responsive type-A genes (*OsRRs*) were downregulated in the panicles of the higher-expression lines compared with 02428 (Supplementary Fig. 4b and Supplementary Data 1). *OsRR6* showed higher expression in the branching primordia of 02428 plants than in 02428-*GY3*^{TQ} plants (Supplementary Fig. 4g,m and Supplementary Data 1). This phenomenon is in line with the increased content of active cytokinins in the panicles of 02428. The expression of both *OsIPT7* and *OsIPT8*, which encode enzymes whose catalytic activities represent the first step of isoprenoid cytokinin biosynthesis, was significantly induced and led to a high content of iPRMP in the 02428 compared with the higher-expression lines (Supplementary Fig. 4c and Supplementary Data 1).

The *solo-LTR* insertion originated from *Oryza rufipogon*

The vital role of *solo-LTR* insertion in regulating *GY3* expression prompted us to investigate its distribution and evolutionary trajectory in rice. By genotyping the minicore germplasm collection consisting of 533 accessions[24], we found that 144 of 146 *japonica*, 42 of 221 *indica* and 19 of 39 *aus* accessions contained the *solo-LTR* insertion (Supplementary Data 2). In general, the *indica* accessions with the *solo-LTR* insertion had a lower expression level of *GY3* and higher grain yield per plant than did those without the insertion (Supplementary Fig. 5 and Supplementary Data 1). Intact *Gypsy LTR* retroelements (LTR-RTs) are usually arranged in two or three ORFs with *gag*, *pol* and *env* and flanked by two LTRs in the 5′ upstream and 3′ downstream regions[25]. However, unequal homologous recombination can partially remove LTR-RTs, resulting in a *solo-LTR* at the original insertion site lacking ORFs and one LTR border region, which is universal among eukaryotes[26], including yeast[27], primates[28] and plants[29–31]. Alignments of 100 kb sequences surrounding *GY3* from eight *Oryza* AA reference genome sequences showed that only Nipponbare and W1943 (*O. rufipogon*) contained the *solo-LTR* insertion and no intact *Gypsy LTR* was identified in Nipponbare or W1943 (Supplementary Fig. 6 and Supplementary Data 3 and 4). To better trace the trajectory of *solo-LTR* insertions, we expanded this analysis to a comprehensive germplasm population comprising 3,010 diverse accessions of Asian cultivated rice[32], 25 African cultivated

https://doi.org/10.1038/s41588-023-01454-3

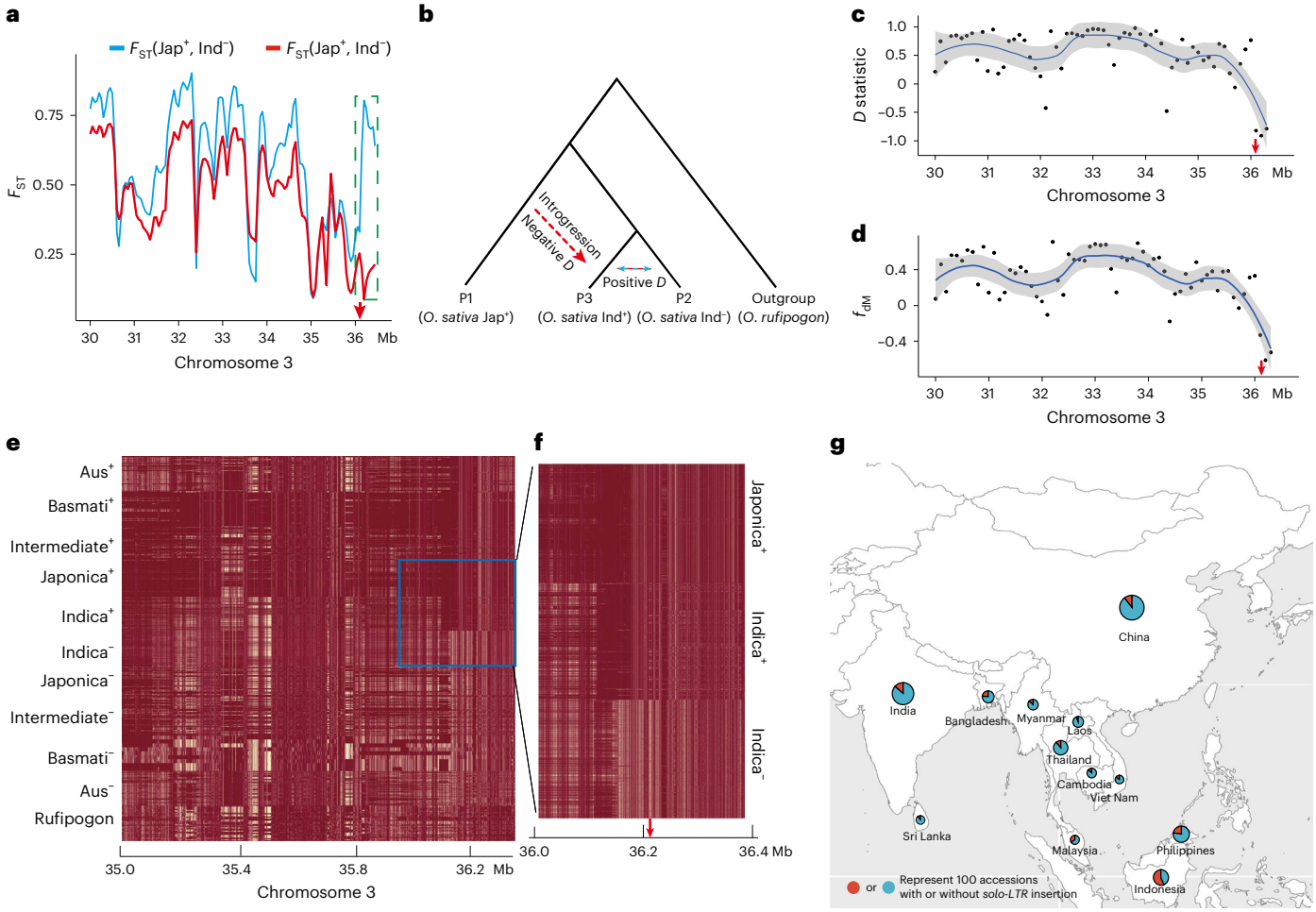


Fig. 4 | Introgression of the *solo-LTR* insertion allele of *GY3* from *japonica* to *indica* cultivars. a, Values of F_{ST} (*Jap*, *Ind*$^-$) and F_{ST} (*Jap*, *Ind*$^+$) for the long-arm terminus of chromosome 3. **b**, According to a phylogenetic analysis restricted to four populations (*O. sativa japonica*$^+$, (*O. sativa japonica*$^+$, *O. sativa indica*$^+$), *O. rufipogon*), the interflow between *O. sativa japonica*$^+$ and *O. sativa indica*$^+$ resulted in positive *D* and f_{dM} values, whereas the introgression between *O. sativa japonica*$^+$ and *O. sativa indica*$^+$ led to negative *D* and f_{dM} values. **c**,**d**, Plot of the *D* statistic (**c**) and f_{dM} statistic (**d**) for the long-arm terminus of chromosome 3, the smooth line conducted with locally weighted regression (loess) with confidence interval at 0.95 level. **e**,**f**, Haplotype analyses of introgression within the long-arm terminus of chromosome 3. Included is a comparison of the haplotypes among *indica*, *japonica*, *aus*, *basmati* and an admix (intermediate) with or without *solo-LTR* insertion (**e**), with a focus on haplotype differentiation within the *indica* population (**f**); the red color represents the reference allele. The red arrows (**a**,**c**,**d**,**f**) indicate the *GY3* location on the chromosome. **g**, Percentage and distribution of *solo-LTR* insertions in *indica* cultivars in Asian countries. Credits for panel **g**: background tiles from Stamen Design under a Creative Commons CC BY 3.0 license; map plotted using ggmap[40] and data from ref. 41 under an Open Data Commons Open Database License (ODbL).

rice accessions (*O. glaberrima*) and 245 wild rice accessions from OryzaGenome (http://viewer.shigen.info)[33] (Supplementary Fig. 6). One-half (29 of 56) of *O. rufipogon* and 98.1% (838 of 854) of *japonica* accessions contained the *solo-LTR* insertion. No insertion was detected in 12 accessions of *O. nivara*, which is considered the ancestor of *O. sativa indica*[32,34], while 20.6% (369 of 1,789) of *indica* accessions contained the insertion (Supplementary Fig. 6 and Supplementary Data 3–5). Similarly, no *solo-LTR* insertion was identified in *O. glaberrima* (0 of 25) or *O. meridionalis* (0 of 20). The *solo-LTR* insertion is rare or infrequent in the other three *Oryza* species with an AA genome, namely, *O. barthii* (0.8%, 1 of 117), *O. glumipatula* (4.8%, 1 of 22) and *O. longistaminata* (16.7%, 3 of 18). Notably, most *basmati* accessions (73 of 76) and one-half (116 accessions) of 201 *aus* accessions contained the *solo-LTR* insertion (Supplementary Data 5). These findings are consistent with those of a previous study showing that *basmati* rice is sister to *japonica* rice; both share the same ancestor, *O. rufipogon* and the high-frequency admixture between *aus* and *basmati* accessions[35] may contribute to the high frequency of this insertion in the *aus* subgroup. Taken together, these results revealed that the retrotransposon element (*solo-LTR*) within the promoter of *GY3* was first inserted in *O. rufipogon* and, subsequently, has become present in *O. sativa japonica* through domestication.

Introgression of the *solo-LTR* insertion to *indica* rice

To understand how the 20.6% *indica* rice (*Ind*$^+$, *indica* rice with the *solo-LTR* insertion) obtained the *GY3* allele with this insertion and to determine whether the *GY3*-containing region of *Ind*$^+$ was obtained from *japonica*, we first analyzed the F_{ST} values, which represent the inbreeding coefficient among different populations, corresponding to the ~6 megabase (Mb) region (from 30 Mb to the end of chromosome 3) flanking *GY3* between *Ind*$^-$ and *Jap*$^+$ populations (F_{ST} (*Ind*$^-$, *Jap*$^+$)) and between *Ind*$^+$ and *Jap*$^+$ populations (F_{ST} (*Ind*$^+$, *Jap*$^+$)) (Fig. 4a). The extremely low F_{ST} (*Ind*$^+$, *Jap*$^+$) of 0.087 and the very high F_{ST} (*Ind*$^-$, *Jap*$^+$) of 0.89 clearly showed that *Ind*$^+$ and *Jap*$^+$ displayed high sequence similarity in the *GY3* region. A 'gene flow' analysis further revealed positive *D* and f_{dM} values corresponding to an adjacent region (30–36 Mb) of *GY3*, indicating that *Ind*$^-$ and *Ind*$^+$ shared the most identical alleles. However, the values of *D* and f_{dM} corresponding to the region surrounding *GY3* (nearly

https://doi.org/10.1038/s41588-023-01454-3

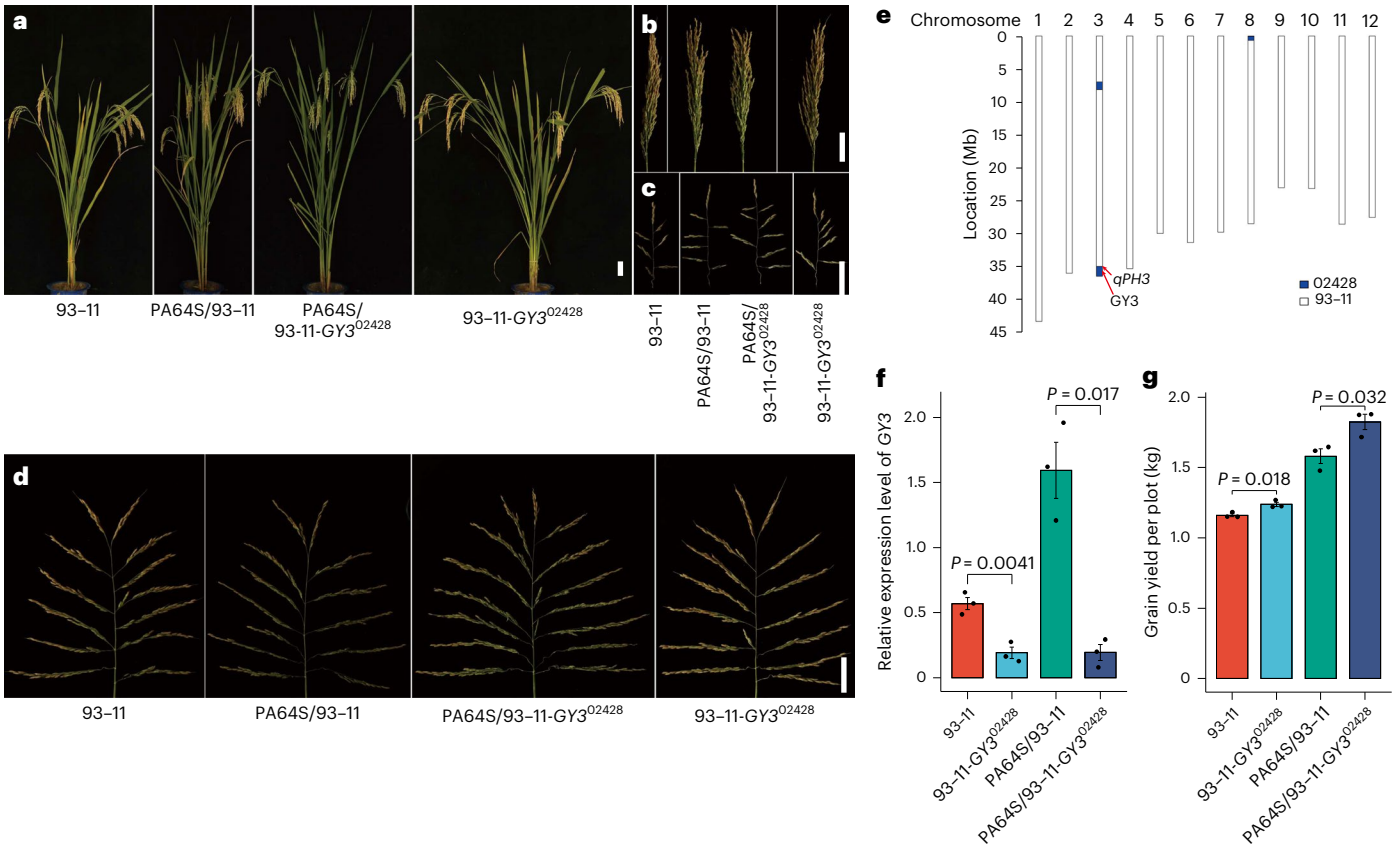


Fig. 5 | Improvements of the yield performances of 93-11 and its related hybrids by *GY3*^{02428}. **a**–**d**, Performances of whole plant (**a**), panicle (**b**), primary branch (**c**) and panicle architecture (**d**) at maturation stage of 93-11, 93-11-*GY3*^{02428} and their hybrids crossed with PA64S. Scale bars, 5 cm. **e**, Genetic components of 93-11-*GY3*^{02428} identified by the 6,000 SNP array. **f**, Relative expression level of *GY3* in leaves at seedling stage. Data are mean ± s.e. (three biological and three technical replicates). *P* values were calculated by two-sided paired Student's *t*-test. **g**, Comparisons of the grain yields per plot (each plot contained 40 plants) of 93-11, 93-11-*GY3*^{02428} and their hybrids crossed with PA64S. Data are mean ± s.e. ($n = 3$). *P* values were calculated by two-sided paired Student's *t*-test.

300 kb in length) were negative, indicating that *Jap*$^+$ and *Ind*$^+$ shared the most identical alleles (Fig. 4b–d). Taken together, these results suggest that 'gene flow' within the *GY3* region occurred from *Jap*$^+$ to *Ind*$^+$.

Haplotype blocks of the target 1.4 Mb region surrounding *GY3* (35.0–36.4 Mb) were constructed for the 3,010 diverse accessions (Fig. 4e,f). Most (68%) of the accessions in *Ind*$^+$ carried *Jap*$^+$ haplotype blocks covering a 280 kb segment (from 36.12 to 36.40 Mb) surrounding *GY3*, indicating that the *GY3* segment was introgressed from *japonica* to *Ind*$^+$. However, a few accessions (32%) of *Ind*$^+$ carried *Jap*$^+$ haplotype blocks that were less than 280 kb in length, indicating that the introgressed segment was shortened, probably due to recombination during *indica* breeding. Conversely, half of the *Jap*$^-$ introgressed *Ind*$^-$ haplotype blocks were located within the whole 1.4 Mb region. Introgression of the *GY3* segment from *Jap*$^+$ was also observed in the *intermediate*$^+$, *aus*$^+$ and *basmati*$^+$ groups (Fig. 4e,f).

GY3^{02428} improved the yield of restorer lines and hybrids

Geographical distribution analysis of *indica* accessions showed that the occurrence of *solo-LTR* insertions varies across Asian countries, ranging from 5.4% in Cambodia to 56.6% in Indonesia (Fig. 4g, Supplementary Table 4 and Supplementary Data 5). Furthermore, the percentages of *solo-LTR* insertions in landraces (traditional rice) are even higher than those in modern varieties (improved varieties), such as those in Indonesia and the Philippines. Similarly, only 8% of 87 modern Chinese *indica* varieties contained this *solo-LTR* insertion (Supplementary Table 4). Moreover, analysis of 95 *indica* backbone parents used for producing hybrid rice showed that only 3 of 21 photothermosensitive genic male sterile (P/TGMS) lines and 1 of 49 restorer lines have the *solo-LTR* insertion. None of the 25 cytoplasmic male sterile (CMS)/maintainer lines have this insertion (Supplementary Data 6). These results indicated that the favorable allele of *GY3* was not a selection target in previous breeding programs.

To verify its value in the genetic improvement of *indica* rice, the *GY3*^{02428} allele was introgressed into an elite restorer line (93-11) for a two-line hybrid system and into two elite restorer lines (MH63 and SH527) for a three-line hybrid system (Fig. 5 and Extended Data Figs. 8 and 9). The introgression lines were termed 93-11-*GY3*^{02428}, MH63-*GY3*^{02428} and SH527-*GY3*^{02428}; compared with their original restorer lines, these three introgression lines exhibited per-plant increases in grain yield of 28.2%, 12.0% and 14.1%, respectively, in 2020 and 25.5%, 24.5% and 25.8%, respectively, in 2021 in Wuhan (Supplementary Tables 5 and 6). *GY3*^{02428} increased the number of secondary branches and spikelets and grains per panicle and ultimately increased the grain yield of all these lines. The grain yield per plot (1.33 m^2) of these improved restorer lines also increased by 9.1%, 9.2% and 13.5% and by 12.4%, 11.3% and 16.3% compared with their counterpart lines in Lingshui, Hianan Island (short-day condition) and Wuhan (long-day condition), respectively (Fig. 5g, Extended Data Fig. 8g and Supplementary Table 7). Accordingly, the grain yields per plot of the new versions of the elite super hybrids LYP9-*GY3*^{02428}, G46B/MH63-*GY3*^{02428} and G46B/SH527-*GY3*^{02428} were 1.82, 1.62 and 1.61 kg, respectively, which were 15.4%, 7.4% and 11.0% higher than those of their original hybrids LYP9, G46B/MH63 and G46B/SH527, respectively (Fig. 5g and Extended Data Figs. 8g and 9g).

https://doi.org/10.1038/s41588-023-01454-3

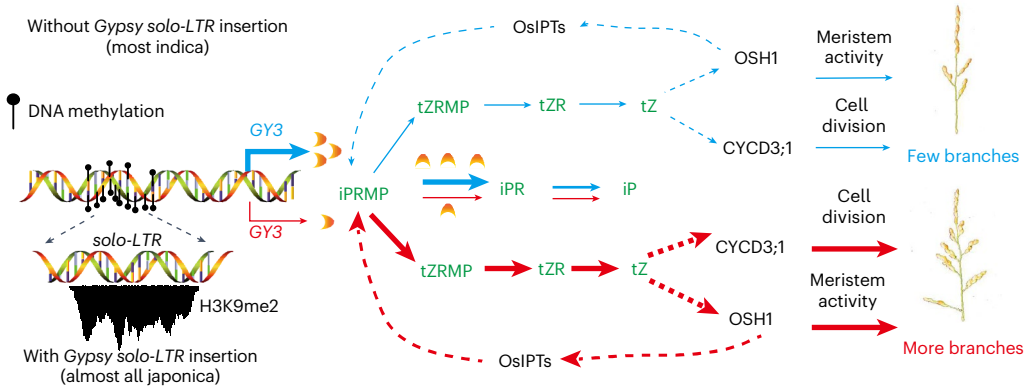


Fig. 6 | Model of how *GY3* determines rice panicle size and grain yield. *GY3* encodes a LOG-like protein that catalyzes the conversion of iPRMP to iPR. The *solo-LTR* insertion induces an H3K9me2 modification and enhances levels of DNA methylation within the promoter of *GY3*, which inhibits the expression of japonica allele of *GY3*. Therefore, more cytokinin nucleotides (iPRMP) in the metabolic pool are converted to tZRMP and further to tZ, an active form of cytokinin that enhances the expression of *OSH1* and *CYCD3;1*, which subsequently maintain meristematic activity and promote cell division, respectively. Finally, the panicle develops increased numbers of secondary branches, which leads to increased grain yields. The chemical molecules in green indicate different forms of cytokinins. The dashed lines represent the regulatory network and the arrow size represents the catalytic or regulatory strength.

Discussion

The involvement of LOG proteins in the two-step pathway of cytokinin synthesis in plants remains to be explored. The major product of GY3 is iPR, whereas the major product of An-2 is iP. Notably, small amounts of iP and iPR were detected as a result of the enzymatic reaction of GY3 and An-2, respectively (Fig. 3g). The product of LOG via the one-step cytokinin synthesis pathway was previously verified and there was an unconfirmed minor accumulation of LOG and An-2/LABA1 reaction products[7,11], which might be signals of iPR. According to the previously proposed LOG catalytic mechanism, the glutamic acid residue in $PGG_XGT_{XX}E$ motif activated a water molecule to attack the ribose ring and the arginine residue (R116) stabilized transition state then produced iP[36] (Supplementary Fig. 2). Docking analysis showed that R116 of An-2 may interact with the ribose group of iPRMP whereas R121 of GY3 is supposed not to interact with the ribose group of iPRMP (Extended Data Fig. 10a,b,d,e). This is in agreement with the results that R116 of An-2 is essential to phosphoribohydrolase activity whereas R121 of GY3 is of no account for phosphoribohydrolase activity. On the other hand, the residues S41 of GY3 and S36 of An-2 are also able to interact with the phosphate group of iPRMP (Extended Data Fig. 10c,f). This is consistent with the fact that the serine residue (S41 and S36) is responsible for phosphohydrolase activity. The conserved residues of LOG homologs such as arginine (corresponding to R116 of An-2) for phosphoribohydrolase activity and serine (corresponding to S41 of GY3) for phosphohydrolase activity in rice, maize and wheat suggested that all LOGs might possess bifunctional catalytic activities and each LOG has a dominant activity which is determined by the interaction between these two catalytic residues and substrates (Supplementary Fig. 2). For example, GY3 mainly acts as phosphohydrolase, while other members like LOG and An-2 dominantly act as phosphoribohydrolase[12–14] (Supplementary Fig. 7).

The *solo-LTR* insertion enhanced the levels of DNA methylation and induced the enrichment of H3K9me2, which together downregulated *GY3* expression and resulted in high grain yield in 02428 and several other genetic backgrounds (Supplementary Tables 5–7). Therefore, we propose a working model for *GY3* (Fig. 6). The *solo-LTR* insertion within the promoter of *GY3* in the *japonica* variety 02428 suppresses the expression of *GY3* in the panicles by altering histone modification and DNA methylation levels. A lower content of GY3 in 02428 limits the consumption of iPRMP to iPR, while more iPRMP is converted to tZRMP and, finally, more active cytokinins such as tZ are produced, which in turn leads to increased expression of *CYCD3;1* and *OSH1*. A high content of OSH1 subsequently promotes *OsIPTs* expression, which further enhances the synthesis of iPRMP. Ultimately, the increased meristematic activity and cell division activity result in increased numbers of secondary branches and substantially increase grain yield.

In Indonesia and Malaysia, 60.7% and 35.0% of the *indica* accessions with the *solo-LTR* insertion are traditional rice accessions, which indicates that the insertion was introgressed into *indica* rice before modern genetic improvements were made. Both *indica* and *tropical japonica* cultivars are commonly grown in these countries[32], which provides the opportunity for naturally occurring intersubspecies pollination. However, the percentage of the modern varieties with the insertion did not increase (Supplementary Table 4). Although the breeding program for new rice plant types involving *indica–japonica* intersubspecies crossings was initiated by the International Rice Research Institute (IRRI) approximately 50 years ago[37], the proportion (33 of 163) of modern *indica* varieties with the insertion has not increased in the Philippines, where IRRI is located. In China, only 8.0% of improved *indica* varieties and 4.2% of backbone parents of *indica* hybrids were found to carry the high-yielding *GY3* alleles. All these data showed that the higher-yielding japonica *GY3* allele has not been the artificial selection target in previous *indica* breeding programs, suggesting the great potential to introduce japonica *GY3* alleles to improve grain yield in *indica* cultivars. In fact, all the $GY3^{02428}$-improved elite *indica* restorer lines and hybrids presented substantially increased grain yields under both long-day and short-day conditions (increases of 12.8% and 11.3% per plot on average) (Fig. 5, Extended Data Fig. 8 and Supplementary Table 7). Unlike other yield-related genes such as *Ghd7* in rice and *TaCol-B5* in wheat, which increase grain yield by delaying heading date[38,39], *GY3* has no effect on heading date (Supplementary Table 5). This trait may provide an attractive target for improving the grain yield of *indica* cultivars worldwide as it does not alter the adaptation scheme of rice cultivars in terms of the cropping region or cropping season. The optimal transcriptional regulation of *GY3* may maximize grain yields by optimizing tZ-type cytokinins in rice. Thus, the generation of saturated mutations by gene editing within the promoter region may be a suitable alternative for generating new, superior *GY3* alleles.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41588-023-01454-3.

References

1. Osugi, A. & Sakakibara, H. Q&A: how do plants respond to cytokinins and what is their importance? *BMC Biol.* **13**, 102 (2015).
2. Ashikari, M. et al. Cytokinin oxidase regulates rice grain production. *Science* **309**, 741–745 (2005).
3. Sakakibara, H. Cytokinins: activity, biosynthesis, and translocation. *Annu. Rev. Plant Biol.* **57**, 431–449 (2006).
4. Kiba, T., Takei, K., Kojima, M. & Sakakibara, H. Side-chain modification of cytokinins controls shoot growth in *Arabidopsis*. *Dev. Cell* **27**, 452–461 (2013).
5. Sakakibara, H. Cytokinin biosynthesis and transport for systemic nitrogen signaling. *Plant J.* **105**, 421–430 (2021).
6. Osugi, A. et al. Systemic transport of trans-zeatin and its precursor have differing roles in *Arabidopsis* shoots. *Nat. Plants* **3**, 17112 (2017).
7. Kurakawa, T. et al. Direct control of shoot meristem activity by a cytokinin-activating enzyme. *Nature* **445**, 652–655 (2007).
8. Chen, C. M. & Kristopeit, S. M. Metabolism of cytokinin: deribosylation of cytokinin ribonucleoside by adenosine nucleosidase from wheat germ cells. *Plant Physiol.* **68**, 1020–1023 (1981).
9. Chen, C. M. & Kristopeit, S. M. Metabolism of cytokinin: dephosphorylation of cytokinin ribonucleotide by 5′-nucleotidases from wheat germ cytosol. *Plant Physiol.* **67**, 494–498 (1981).
10. Gu, B. et al. *An-2* encodes a cytokinin synthesis enzyme that regulates awn length and grain production in rice. *Mol. Plant* **8**, 1635–1650 (2015).
11. Hua, L. et al. *LABA1*, a domestication gene associated with long, barbed awns in wild rice. *Plant Cell* **27**, 1875–1888 (2015).
12. Hinsch, J. et al. De novo biosynthesis of cytokinins in the biotrophic fungus *Claviceps purpurea*. *Environ. Microbiol.* **17**, 2935–2951 (2015).
13. Radhika, V. et al. Methylated cytokinins from the phytopathogen *Rhodococcus fascians* mimic plant hormone activity. *Plant Physiol.* **169**, 1118–1126 (2015).
14. Samanovic, M. I. et al. Proteasomal control of cytokinin synthesis protects *Mycobacterium tuberculosis* against nitric oxide. *Mol. Cell* **57**, 984–994 (2015).
15. Kuroha, T. et al. Functional analyses of LONELY GUY cytokinin-activating enzymes reveal the importance of the direct activation pathway in *Arabidopsis*. *Plant Cell* **21**, 3152–3169 (2009).
16. Seo, H. & Kim, K. J. Structural basis for a novel type of cytokinin-activating protein. *Sci. Rep.* **7**, 45985 (2017).
17. Naseem, M., Bencurova, E. & Dandekar, T. The cytokinin-activating LOG-family proteins are not lysine decarboxylases. *Trends Biochem. Sci.* **43**, 232–236 (2018).
18. Tokunaga, H. et al. *Arabidopsis* LONELY GUY (LOG) multiple mutants reveal a central role of the LOG-dependent pathway in cytokinin activation. *Plant J.* **69**, 355–365 (2012).
19. Wu, B., Mao, D. H., Liu, T. M., Li, Z. X. & Xing, Y. Z. Two quantitative trait loci for grain yield and plant height on chromosome 3 are tightly linked in coupling phase in rice. *Mol. Breed.* **35**, 156 (2015).
20. Wu, Y. et al. The QTL *GNP1* encodes GA20ox1, which increases grain number and yield by increasing cytokinin activity in rice panicle meristems. *PLoS Genet.* **12**, e1006386 (2016).
21. Du, J. et al. Dual binding of chromomethylase domains to H3K9me2-containing nucleosomes directs DNA methylation in plants. *Cell* **151**, 167–180 (2012).
22. Schaller, G. E., Street, I. H. & Kieber, J. J. Cytokinin and the cell cycle. *Curr. Opin. Plant Biol.* **21**, 7–15 (2014).
23. Marzluff, W. F. & Duronio, R. J. Histone mRNA expression: multiple levels of cell cycle regulation and important developmental consequences. *Curr. Opin. Cell Biol.* **14**, 692–699 (2002).
24. Chen, W. et al. Genome-wide association analyses provide genetic and biochemical insights into natural variation in rice metabolism. *Nat. Genet.* **46**, 714–721 (2014).
25. Kumar, A. & Bennetzen, J. L. Plant retrotransposons. *Annu. Rev. Genet.* **33**, 479–532 (1999).
26. Vitte, C. & Panaud, O. Formation of *solo-LTR*s through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. *Mol. Biol. Evol.* **20**, 528–540 (2003).
27. Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A. & Voytas, D. F. Transposable elements and genome organization: a comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. *Genome Res.* **8**, 464–478 (1998).
28. Goodchild, N. L., Wilkinson, D. A. & Mager, D. L. Recent evolutionary expansion of a subfamily of RTVL-H human endogenous retrovirus-like elements. *Virology* **196**, 778–788 (1993).
29. SanMiguel, P. et al. Nested retrotransposons in the intergenic regions of the maize genome. *Science* **274**, 765–768 (1996).
30. Vitte, C., Panaud, O. & Quesneville, H. LTR retrotransposons in rice (*Oryza sativa*, L.): recent burst amplifications followed by rapid DNA loss. *BMC Genom.* **8**, 218 (2007).
31. Ma, J., Devos, K. M. & Bennetzen, J. L. Analyses of LTR-retrotransposon structures reveal recent and rapid genomic DNA loss in rice. *Genome Res.* **14**, 860–869 (2004).
32. Wang, W. et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature* **557**, 43–49 (2018).
33. Ambavaram, M. M. et al. Coordinated regulation of photosynthesis in rice increases yield and tolerance to environmental stress. *Nat. Commun.* **5**, 5302 (2014).
34. Stein, J. C. et al. Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nat. Genet.* **50**, 285–296 (2018).
35. Choi, J. Y. et al. Nanopore sequencing-based genome assembly and evolutionary genomics of circum-basmati rice. *Genome Biol.* **21**, 21 (2020).
36. Seo, H. & Kim, K. J. Structural insight into molecular mechanism of cytokinin activating protein from *Pseudomonas aeruginosa* PAO1. *Environ. Microbiol.* **20**, 3214–3223 (2018).
37. Denning, G. L. & Mew, T. W. *China and IRRI: Improving China's Rice Productivity in the 21st Century* (IRRI, 1997).
38. Xue, W. et al. Natural variation in Ghd7 is an important regulator of heading date and yield potential in rice. *Nat. Genet.* **40**, 761–767 (2008).
39. Zhang, X. et al. TaCol-B5 modifies spike architecture and enhances grain yield in wheat. *Science* **376**, 180–183 (2022).
40. Kahle, D. & Wickham, H. ggmap: spatial visualization with ggplot2. *R J.* **5**, 144–161 (2013).
41. Haklay, M. & Weber, P. OpenStreetMap: user-generated street maps. *IEEE Pervasive Comput.* **7**, 12–18 (2008).

Methods

Plant materials and phenotypic evaluations

The NIL for *GY3* (02428-*GY3*^{TQ}) was derived from a cross between the recurrent parent 02428 (a widely compatible variety; *O. sativa* ssp. *japonica*) and the donor parent TQ (*O. sativa* ssp. *indica*), a line harboring homozygous 02428 alleles of *qPH3* and homozygous TQ alleles of *GY3* was developed via marker-assisted selection. A mapping population for *GY3* was developed by self-pollination of a line (L1) that harbored homozygous 02428 alleles of *qPH3* and heterozygous alleles of *GY3*. Following this strategy to use 02428^{GY3} to improve three elite restorer lines, namely 93-11, Minghui63 (MH63) and Shuhui527 (SH527) and TQ, all these lines were used as recurrent parents and crossed with donor parent 02428 followed by four continuous backcrosses to generate BC_4F_1 plants. One BC_4F_1 plant in each background was self-pollinated to generate a BC_4F_2 population. One-hundred plants of each BC_4F_2 population were genotyped with marker ID7 (Supplementary Data 7), a cosegregation marker of *GY3*. A randomly selected BC_4F_2 plant with a homozygous 02428^{GY3} allele from each background (named 93-11-*GY3*^{02428}, MH63-*GY3*^{02428}, SH527-*GY3*^{02428} and TQ-*GY3*^{02428}) was deemed as NIL and genotyped with a RICE6K SNP array[42]. NILs were used for hybrid production and grain yield comparisons. These materials were grown at the experimental stations of Huazhong Agricultural University in Wuhan, Hubei Province, and Lingshui, Hainan Province, China. All the materials were planted such that the distance was 16.5 cm between plants within a row and 25 cm between rows.

The tillers per plant, plant height, panicle length, primary branches per panicle, secondary branches per panicle, spikelets per panicle, grains per panicle, 1,000-grain weight, biomass per plant, harvest index and grain yield per plant were evaluated for each material. Phenotypic measurements were performed on three independent transgenic lines. Significant differences were determined with Student's *t*-test for each set of materials.

Vector construction and genetic transformation

The DNA fragment including the 2 kb promoter region of *GY3* was amplified from TQ with the primers CP-GY3-F and CP-GY3-R and inserted into the complementation vector pCAMBIA1301. The full-length ORF of *GY3* was amplified from TQ by the use of primers GY3U-F and GY3U-R and inserted into the overexpression vector pU1301. To generate clustered regularly interspaced short palindromic repeats (CRISPR)-associated 9 (Cas9) constructs for knockout of the *solo-LTR* in the *GY3* promoter, the *OsU6a* and *OsU6b* promoters were used to drive sgRNA containing the flanking target sequence of *solo-LTR* (OsU6a-sgRNA-LTR1 and OsU6b-sgRNA-LTR2). Then, these two fragments were inserted into the pCXUN-CAS9 vector[43]. These constructs were subsequently introduced into 02428 calli by *Agrobacterium*-mediated transformation. To generate G*Y3*-knockout mutant, a CRISPR vector was constructed with two targets (gRT3-GY3 and gRT4-GY3) and introduced into Zhonghua 11 calli.

RNA extraction and expression analysis

Total RNA was extracted from various rice tissues using TransZol reagent (TransGen Biotech). Genomic DNA was removed using DNaseI (Invitrogen) and first-strand complementary DNA was synthesized using an M-MLV Reverse Transcriptase kit (Invitrogen). Reverse transcription quantitative polymerase chain reaction (RT–qPCR) was performed using gene-specific primers (Supplementary Data 7) in a reaction solution that included FastStart Universal SYBR Green Master Mix (Roche) on a QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems).

Amino acid sequence alignments of LOGs

LOG protein sequence was used to search for LOGs in rice, maize and wheat with BLASTP (https://www.ncbi.nlm.nih.gov/), while ClustalX2 was used to realize amino acid sequence alignment. There are three copies of LOGs in wheat, while only the ones with the most similarity or the longest sequence were selected for sequence alignment. The secondary structure information of aligned sequences was rendered in ESPript 3.0 software (https://espript.ibcp.fr/ESPript/ESPript/).

Recombinant enzymes and enzyme assays

The coding sequences of *GY3* and *An-2* were amplified from cDNA (via primers pET-28a-GY3-F/R, pGEX-6P-1-LOGL6-F/R) (Supplementary Data 7) and inserted into pET-28a and pGEX-6P-1 with His-tagged and GST-tagged sequences, respectively. GY3 and An-2 were expressed in the *Escherichia coli* BL21 (DE3) strain under the induction of 1 mM isopropyl-D-thiogalactopyranoside (IPTG) in LB media for 24 h at 16 °C. GY3 and An-2 protein purification was conducted in Ni-NTA Agarose Resin (Thermo Fisher Scientific) and glutathione Sepharose beads (GE Healthcare) to purify recombinant proteins according to the manufacturers' protocols. The enzymatic activity of the GY3, An-2 and recombinant proteins was measured by incubating 2 μg of purified GY3 and An-2 proteins in a reaction mixture consisting of 50 mM Tris-HCl, 1 mM $MgCl_2$ and 1 mM dithiothreitol (DTT) (pH 6.5) with different substrate contents of iPRMP at 30 °C for 2 h. We also used iPR and iP as the substrates for the identified enzyme activities of GY3 and An-2 under the same reaction conditions. All the reaction products were purified through an Oasis MCX column (Waters Corporation Milford). The products were measured with a LC/MS system (Agilent 6520 Accurate-Mass Q-TOF LC/MS; Agilent Zorbax C18, 4.6 mm 350 mm) under the conditions described by ref. 44. One unit of enzyme activity was defined as the amount of enzyme that produced 1 μmol of products per minute under the reaction conditions. V_{max} is defined as the maximum reaction rate at a given amount of enzyme. K_{cat} is defined as number of substrates converted by each enzyme molecule per minute. The K_m values were calculated using Lineweaver–Burk plots. All cytokinin substrates and standard substances were purchased from OlChemim (Olomouc).

DNA methylation analysis

Genomic DNA from young panicle (<2 mm) tissues (02428 and 02428-TQ^{GY}) were extracted using CTAB method. The CpG island sites and BSP primers (Supplementary Data 7) in the promoter of *GY3* were predicted and designed with MethPrimers (http://www.urogene.org/cgi-bin/methprimer/methprimer.cgi), respectively. DNA was treated by bisulfite conversion method using EpiTect Bisulfite Kit (Qiagen). And PCR products using BSP primers were inserted to the *pEASY*-Blunt cloning vector (TransGen Biotech) and sequenced for more than 30 clones. The reference sequence and corresponding clones' sequence were submitted to an online bisulfite analysis tool (Cytosine Methylation Analysis Tool for Everyone; https://www.cymate.org/cymate.html) for calculating the DNA methylation level of each context. For the whole-genome bisulfite sequencing (WGBS), the bisulfite conversion of DNA, library construction and sequencing were performed at the Beijing Genomics Institute. The raw WGBS data were treated by Fastp (v.0.20.1) and mapped to the MH63RS2 (*O. sativa* ssp. *indica* cv. Minghui 63)[45] and Nip (*O. sativa* ssp. *japonica* cv. Nipponbare)[46] reference genome by Bismark (v.0.23.1) using default parameters[47]. The uniquely mapped reads were retained for DNA methylation level analysis. The DNA methylation level of each cytosine was obtained by dividing the depth of methylated reads by the total coverage of individual cytosines.

Enhanced chromatin immunoprecipitation sequencing library preparation and analysis

Young panicle (<2 mm) tissues (02428 and 02428-TQ^{GY3}) were crosslinked with 1% formaldehyde for 10–15 min and quenched with 0.2 M glycine at room temperature. Approximately 0.2 g of samples were ground in liquid nitrogen into fine powder and used for each eChIP–seq assay according to the previously study[48]. After chromatin was fragmented by sonication, 20 μl of supernatant was used as the input for each sample. For subsequent assay, ChIP was performed

using antibodies against the following: H3K4me3 (ABclonal, A2357), H3K9me2 (Abcam, ab1220), H3K27me3 (ABclonal, A2363) and H3K27ac (ABclonal, A7253). ChIP DNA libraries were prepared using a VAHTS Universal DNA Library Prep Kit for MGI (NDM607-01) according to the manufacturer's guidelines. And the ChIP DNA libraries were sequenced using a MGISEQ-T7 system (paired-end 150 bp reads).

After quality control, the paired-end reads were mapped to MH63RS2 and Nip genomes using Bowtie2 (v.2.4.5) with the default parameters[49]. The BAM mapping files were processed with SAMtools v.1.10 and the 'picard MarkDuplicates' in Genome Analysis Toolkit (GATK) (v.4.2.2.0) was used to mark and remove the duplicate reads produced by PCR during library construction[50] and MACS2 (v.2 2.2.7.1) was used for peak calling.

Quantification of endogenous cytokinin

Young panicle (<2 mm) tissues (02428, 02428-TQ^{GY3}, 02428-CP, 02428-OE and 02428$^{LTR\text{-}KO}$) were collected for cytokinin content analysis. The endogenous cytokinin contents were determined by Wuhan Greensword Creation Technology, based on liquid chromatography tandem mass spectrometry analysis according to a previously reported method with minor modifications[51]. The contents were determined by three technical replicates with three independent biological samples.

Protein structure prediction and docking

The structure of GY3 and An-2 was predicted in an online server (https://www.alphafold.ebi.ac.uk/) and the docking with iPRMP using AutoDockTools (v.1.5.7) and vina (v.1.2.0) according to the users' guideline[52–54].

RNA in situ hybridization

To determine the transcripts of *GY3*, *OSH1*, *OsRR6* and *Histone H4*, probes for each gene were transcribed by SP6 and T7 RNA polymerase in conjunction with a DIG-labeled nucleotide mixture (Roche). Sections of 02428 and 02428-*GY3*^{TQ} were placed on the same slides for each probe. The hybridizations were performed at 50 °C for 16 h and the colors were developed and visualized by NBT/BCIP stock solution (Roche).

Data acquisition and *solo-LTR* insertion detection

Two pairs of primers were designed to identify the *solo-LTR* insertion genotype within a diverse global collection of 533 *O. sativa* accessions and 95 parents of commercial *indica* hybrids. The primer pair (CL1F/R) resulted in the generation of a PCR product of 628 bp with the template of the TQ allele without *solo-LTR* insertion, whereas no PCR products were generated for 02428 because the primer pair-defined region spans a 4.5 kb segment due to the *solo-LTR* insertion. The primer pair CL1-2F and CL1R generated a PCR product that was 529 bp in length for 02428 with a *solo-LTR* insertion, whereas no PCR products were generated for TQ because it lacked the complementary sequence of primer CL1-2F designed on the basis of the *solo-LTR* sequence. To detect the *solo*-LTR insertion states in *Oryza* species, the raw sequencing data of 3,010 rice genome accessions[32], *O. glaberrima* accessions[55] and wild rice accessions from OryzaGenome (http://viewer.shigen.info/oryzagenome2detail/about/about.xhtml) were downloaded from the European Bioinformatics Institute (ftp://ftp.sra.ebi.ac.uk/). We used the phylogenetic tree of wild *Oryza* AA species and *O. sativa* from International Oryza Map Alignment Project (IOMAP)[34,56] constructed by the neighbor-joining method based on genome sequence. After quality control, the paired-end reads were mapped to the reference genome Nip using Burrows–Wheeler Aligner (BWA; v.0.7.17) with the default parameters[50]. The BAM mapping files were processed with SAMtools (v.1.10) and 'picard MarkDuplicates' in GATK was used to mark and remove the duplicate reads produced by PCR during library construction[50]. The Manta (v.1.6.0) pipeline[57] was used for calling large structural variants and InDels from the paired-end sequencing reads of each accession.

Introgression analysis

To acquire SNPs for subsequent analysis, the GATK (v.4.2.2.0) VariantFiltration with the default parameters was used to call SNPs. A sliding window of 100 kb and a step length of 50 kb were implemented to calculate the F_{ST} values with VCFtools (v.0.1.16). The ABBA-BABA test was used to search for evidence of 'gene flow' in the form of excess shared derived alleles between two populations (*Jap*$^+$, *Ind*$^+$) and an outgroup (*O. rufipogon*). These populations have a relationship with ((*Jap*$^+$, (*Ind*$^-$, *Ind*$^+$)), *Ruf*). The *D* and f_{dM} values were calculated according to the methods of refs. 58,59, respectively. Each *D* and f_{dM} value was calculated for a sliding window with a size of 200 kb and a step length of 100 kb. The R package ggmap[40] was used to plot the map in Fig. 4. The SNPs were phasing with Beagle (v.5.1)[60] and the haplotype blocks were extracted with plotHaps in vcflib (v.1.0.1)[61] and plotted with R (v.4.0.2).

Statistics and reproducibility

All statistical analyses were performed in R (v.4.0.2). The two-sided paired Student's *t*-test was conducted using function of 't.test' with parameter of 'test.args = two.sided' in R package ggsignif[62]. The multiple comparison was determined with Duncan's test method in the agricolae[63] package of R. The RT–qPCR amplifications were made with three technical replicates for at least three independent biological samples and the rice *ubiquitin* gene (*LOC_Os03g13170*) was used for normalization. RNA in situ hybridization has performed at least three slides for each probe and each slide had continuous sections. Similar transcript patterns observed in most sections of each slide were imaged for presentation.

Reporting summary

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

Data availability

Sequence data from this study can be found in GenBank under accession number PRJEB6180, SRX502298–SRX502317 and SRX502162–SRX502255 for 3,010 diverse accessions, 20 African cultivated rice accessions and 94 accessions of *O. barthii*, respectively. The eChIP–seq sequence data and BED files were deposited in the NCBI Gene Expression Omnibus with accession number GSE231360. The reference genome of MH63RS2 and Nipponbare can be found in Rice Information GateWay (http://rice.hzau.edu.cn/rice_rs2/) and The Rice Annotation Project (https://rapdb.dna.affrc.go.jp/). All data are available in the main text or the Supplementary Information and Supplementary Data. Source data are provided with this paper.

Code availability

All software used in the study are publicly available as described in Methods and Reporting Summary.

References

42. Yu, H., Xie, W., Li, J., Zhou, F. & Zhang, Q. A whole-genome SNP array (RICE6K) for genomic breeding in rice. *Plant Biotechnol. J.* **12**, 28–37 (2014).
43. Ma, X. et al. A Robust CRISPR/Cas9 system for convenient, high-efficiency multiplex genome editing in monocot and dicot plants. *Mol. Plant* **8**, 1274–1284 (2015).
44. Takei, K., Sakakibara, H. & Sugiyama, T. Identification of genes encoding adenylate isopentenyltransferase, a cytokinin biosynthesis enzyme, in *Arabidopsis thaliana*. *J. Biol. Chem.* **276**, 26405–26410 (2001).
45. Zhang, J. et al. Extensive sequence divergence between the reference genomes of two elite indica rice varieties Zhenshan 97 and Minghui 63. *Proc. Natl Acad. Sci. USA* **113**, E5163–E5171 (2016).

46. International Rice Genome Sequencing, P. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
47. Krueger, F. & Andrews, S. R. Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications. *Bioinformatics* **27**, 1571–1572 (2011).
48. Zhao, L. et al. Integrative analysis of reference epigenomes in 20 rice varieties. *Nat. Commun.* **11**, 2658 (2020).
49. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
50. Li, H. et al. The sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
51. Liu, Z., Wei, F. & Feng, Y. Q. Determination of cytokinins in plant samples by polymer monolith microextraction coupled with hydrophilic interaction chromatography-tandem mass spectrometry. *Anal. Methods* **2**, 1676–1685 (2010).
52. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
53. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
54. Forli, S. et al. Computational protein–ligand docking and virtual drug screening with the AutoDock suite. *Nat. Protoc.* **11**, 905–919 (2016).
55. Wang, M. et al. The genome sequence of African rice (*Oryza glaberrima*) and evidence for independent domestication. *Nat. Genet.* **46**, 982–988 (2014).
56. Li, W. et al. Draft genomes of two outcrossing wild rice, *Oryza rufipogon* and *O. longistaminata*, reveal genomic features associated with mating-system evolution. *Plant Direct* **4**, e00232 (2020).
57. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
58. Martin, S. H., Davey, J. W. & Jiggins, C. D. Evaluating the use of ABBA-BABA statistics to locate introgressed loci. *Mol. Biol. Evol.* **32**, 244–257 (2015).
59. Malinsky, M. et al. Genomic islands of speciation separate cichlid ecomorphs in an East African crater lake. *Science* **350**, 1493–1498 (2015).
60. Browning, B. L., Zhou, Y. & Browning, S. R. A one-penny imputed genome from next-generation reference panels. *Am. J. Hum. Genet.* **103**, 338–348 (2018).
61. Garrison, E., Kronenberg, Z. N., Dawson, E. T., Pedersen, B. S. & Prins, P. A spectrum of free software tools for processing the VCF variant call format: vcflib, bio-vcf, cyvcf2, hts-nim and slivar. *PLoS Comput. Biol.* **18**, e1009123 (2022).
62. Ahlmann-Eltze, C. & Patil, I. ggsignif: R package for displaying significance brackets for "Ggplot2." Preprint at https://psyarxiv.com/7awm6/ (2021).
63. De Mendiburu, F. Agricolae: Statistical Procedures For Agricultural Research. *R package version 1.3–6* https://CRAN.R-project.org/package=agricolae (2023).

Acknowledgements

We thank B. Han for sharing the clone An-2, M. J. Han for assistance in analyzing the retrotransposon structure and J. B. Wang for his excellent fieldwork. The computations in this paper were performed on the bioinformatics computing platform of the National Key Laboratory of Crop Genetic Improvement, Huazhong Agricultural University. This work was partially supported by funds from the National Natural Science Foundation of China (31821005 to Q.Z., U20A2031 to Y.X., 32061143042 to Y.X. and 31901519 to X.Z.), the Earmarked Fund for China Agriculture Research System (CARS-01 to Q.Z.) and the National Key Laboratory of Crop Genetic Improvement Self-research Program (ZW22B0204 to Y.X.).

Author contributions

Y.X. conceived and supervised this study and reviewed and modified the manuscript. B.W. designed and performed the experiments, collected and analyzed the data and wrote the manuscript. J.M. collected young panicle samples. Hongbo Liu, H.Y. and W.C. measured the cytokinin contents. D.M. and A.S. developed the materials. Z.Z. conducted the eChlp–seq. X.Z. collected the *indica* backbone parents. B.Z. collected the phenotypic data of 533 accessions. Xianghua Li and J.X. organized the laboratory operations. Haiyang Liu, Xingwang Li, W.Y., W.X., P.Y., L.W. and Q.Z. contributed valuable suggestions for the manuscript. All the authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

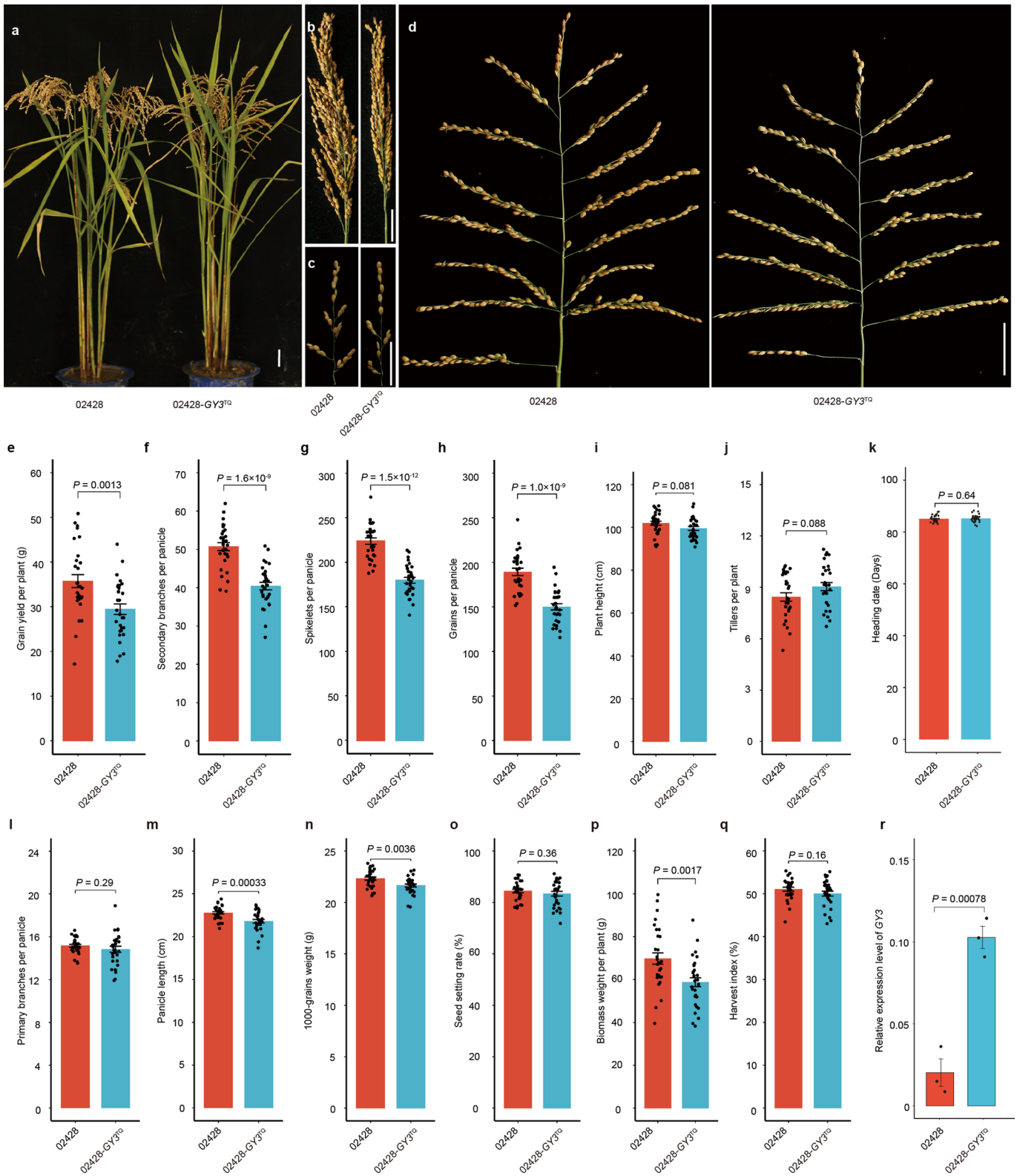
Extended data is available for this paper at https://doi.org/10.1038/s41588-023-01454-3.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41588-023-01454-3.

Correspondence and requests for materials should be addressed to Yongzhong Xing.

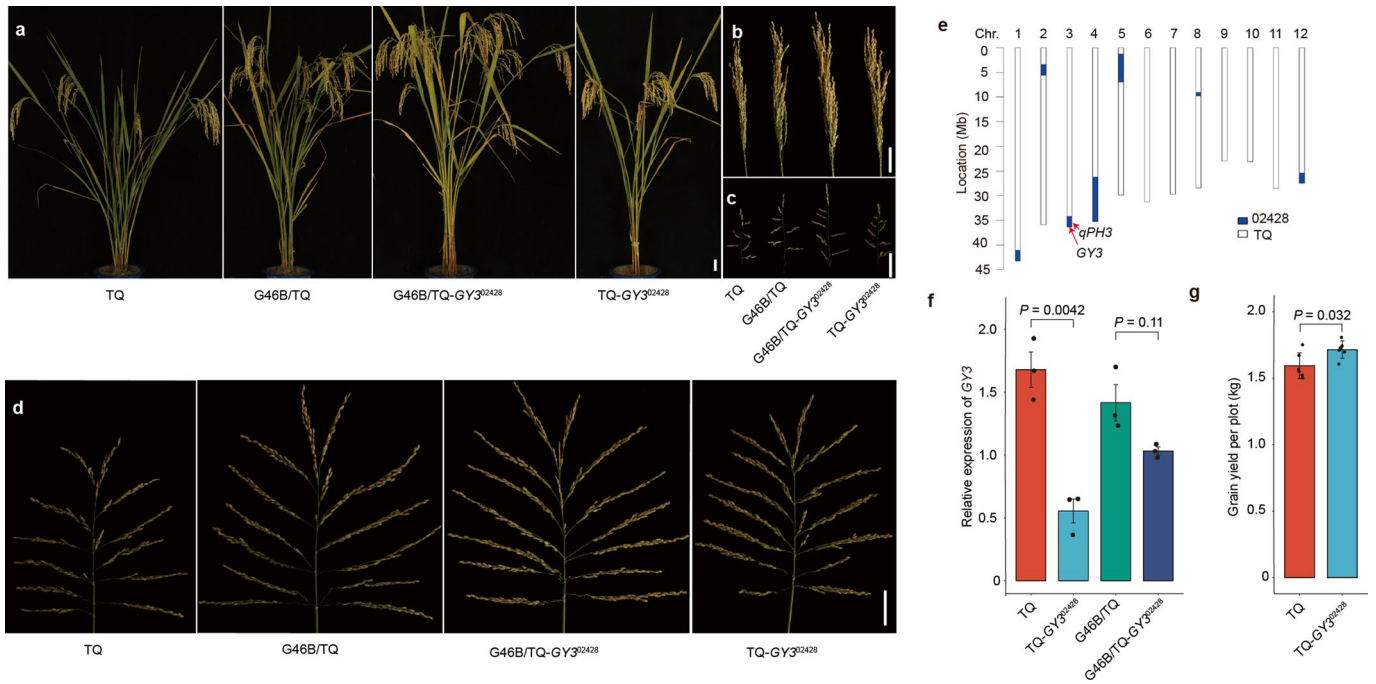
Peer review information *Nature Genetics* thanks Hitoshi Sakakibara, Xiangdong Fu and Yuling Jiao for their contribution to the peer review of this work. Peer reviewer reports are available.

https://doi.org/10.1038/s41588-023-01454-3



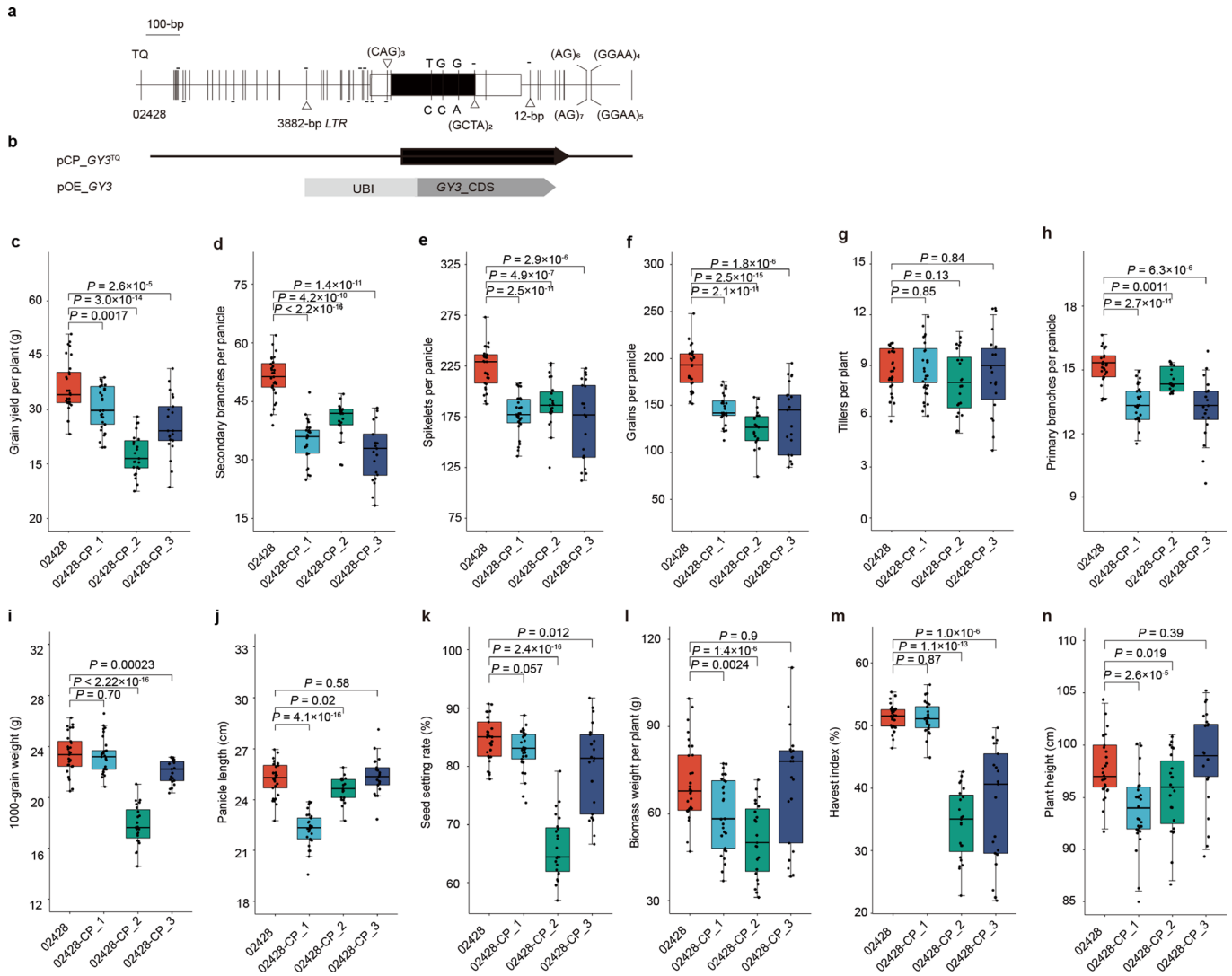
Extended Data Fig. 1 | Plant architecture and yield performance of near-isogenic lines of *GY3*. **a**–**d**, Comparison of whole plant (**a**), panicle (**b**), primary branch (**c**) and panicle architecture (**d**) between 02428 and 02428-*GY3*^{TQ} at maturation stage. Bars, 5 cm. **e**–**q**, Performance of grain yield per plant (**e**), secondary branches per panicle (**f**), spikelets per panicle (**g**), grains per panicle (**h**), plant height (**i**), tillers per plant (**j**), heading date (**k**), primary branches per panicle (**l**), panicle length (**m**), 1000-grains weight (**n**), seed setting rate (**o**), biomass weight per plant (**p**) and harvest index (**q**) between 02428 and 02428-*GY3*^{TQ}. **r**, Relative expression level of *GY3* in leaves at seedling stage. Data are means ± SE (n = 30 for **e**–**q**, n = 3 for **r**). *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



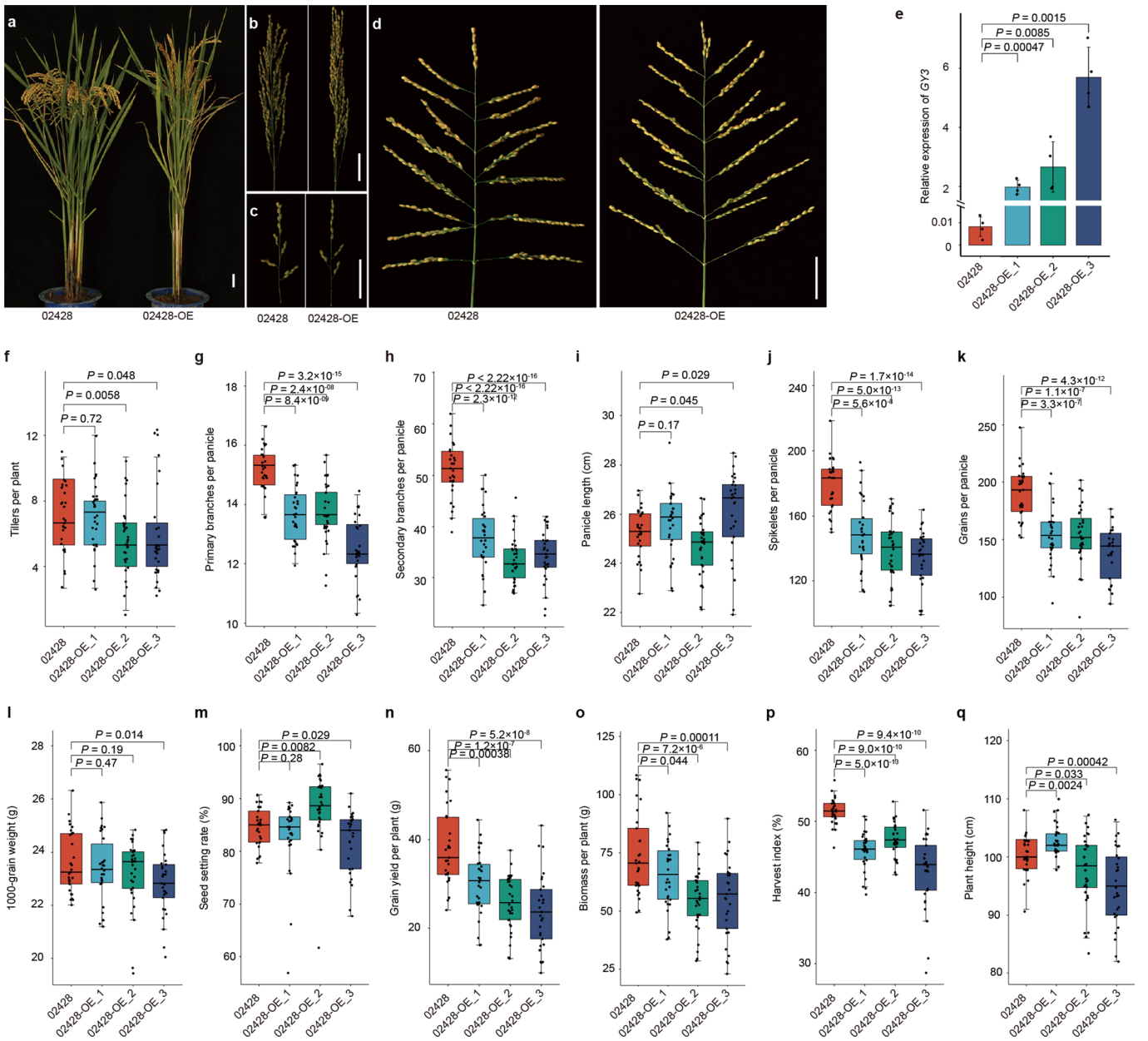
Extended Data Fig. 2 | Improvement of the yield performance of TQ and its related hybrid by *GY3*^{02428}. **a**–**d**, Performances of whole plant (**a**), panicle (**b**), branch (**c**) and panicle architecture (**d**) of TQ, TQ-*GY3*^{02428} and their hybrids with G46B at maturation stage, Bars, 5 cm. **e**, Genetic constitution of TQ-*GY3*^{02428} identified by 6 K SNP array. **f**, Relative expression level of *GY3* in TQ, TQ-*GY3*^{02428} and their hybrids with G46B in leaves at seedling stage. **g**, Grain yield per plot of TQ and TQ-*GY3*^{02428}. Data are mean ± SE (n = 3 for f, n = 6 for g); *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



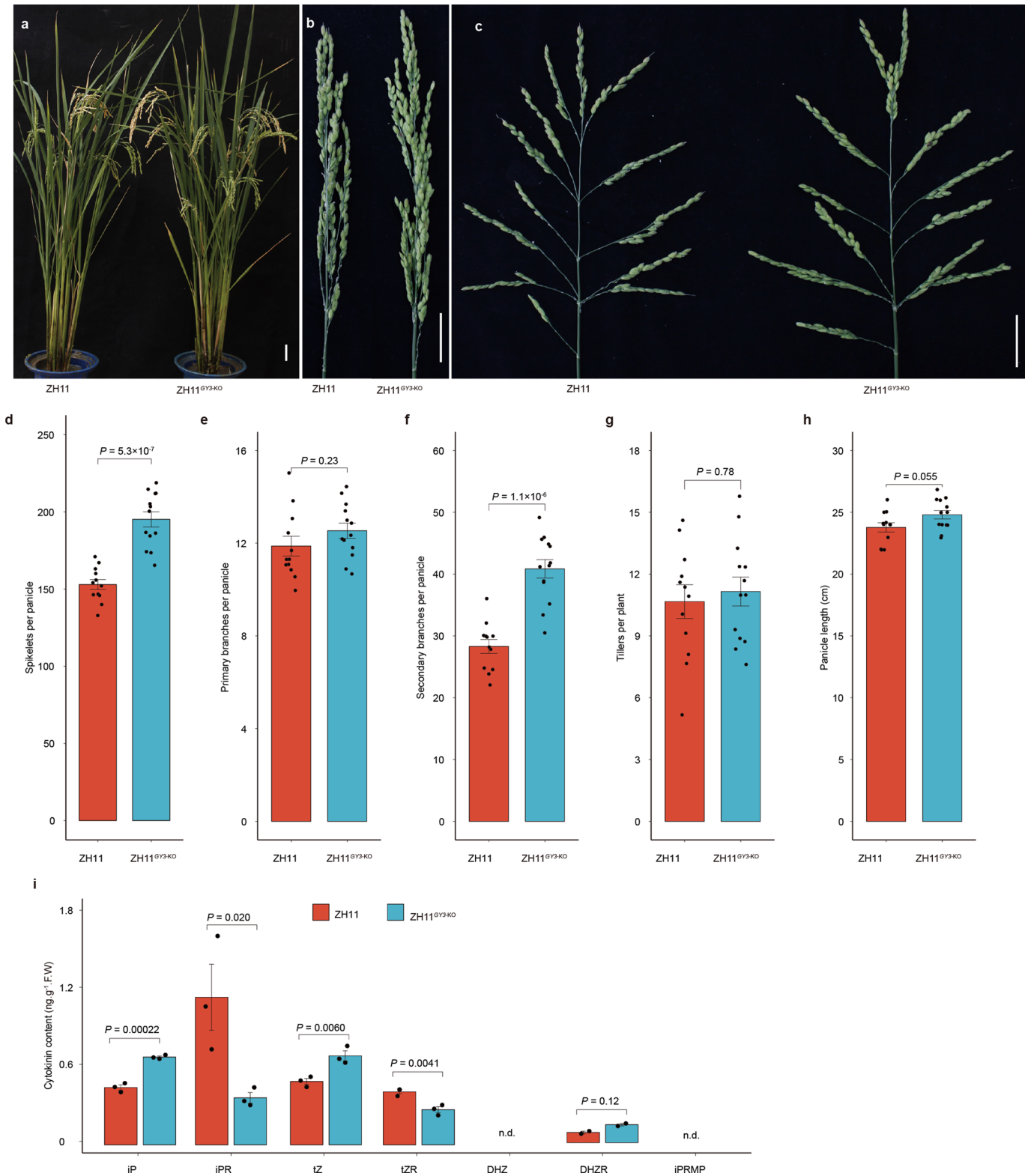
Extended Data Fig. 3 | Natural variations and complementation test of *GY3*. **a**, Gene structure and sequence variations between 02428 and TQ alleles. **b**, Schematic representation of the complementing (CP) and over-expressing (OE) constructs of *GY3*. **c**–**n**, Comparisons of yield-related traits including grain yield per plant (**c**), secondary branches per panicle (**d**), spikelets per panicle (**e**), grains per panicle (**f**), tillers per plant (**g**), primary branches per panicle (**h**),1000-grains weight (**i**), panicle length (**j**), seed setting rate (**k**), biomass per plant (**l**), harvest index (**m**) and plant height (**n**) between 02428 and 02428-CP lines. For box plots in **c**–**n**, the box limits indicate the 25th and 75th percentiles, whiskers further extend by ±1.5 times the interquartile range from the limits of each box, and the center line indicates the median, individual data points are plotted (n = 29, 29, 23 and 21 for 02428, 02428-CP_1, 02428-CP_2 and 02428-CP_3, respectively). *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



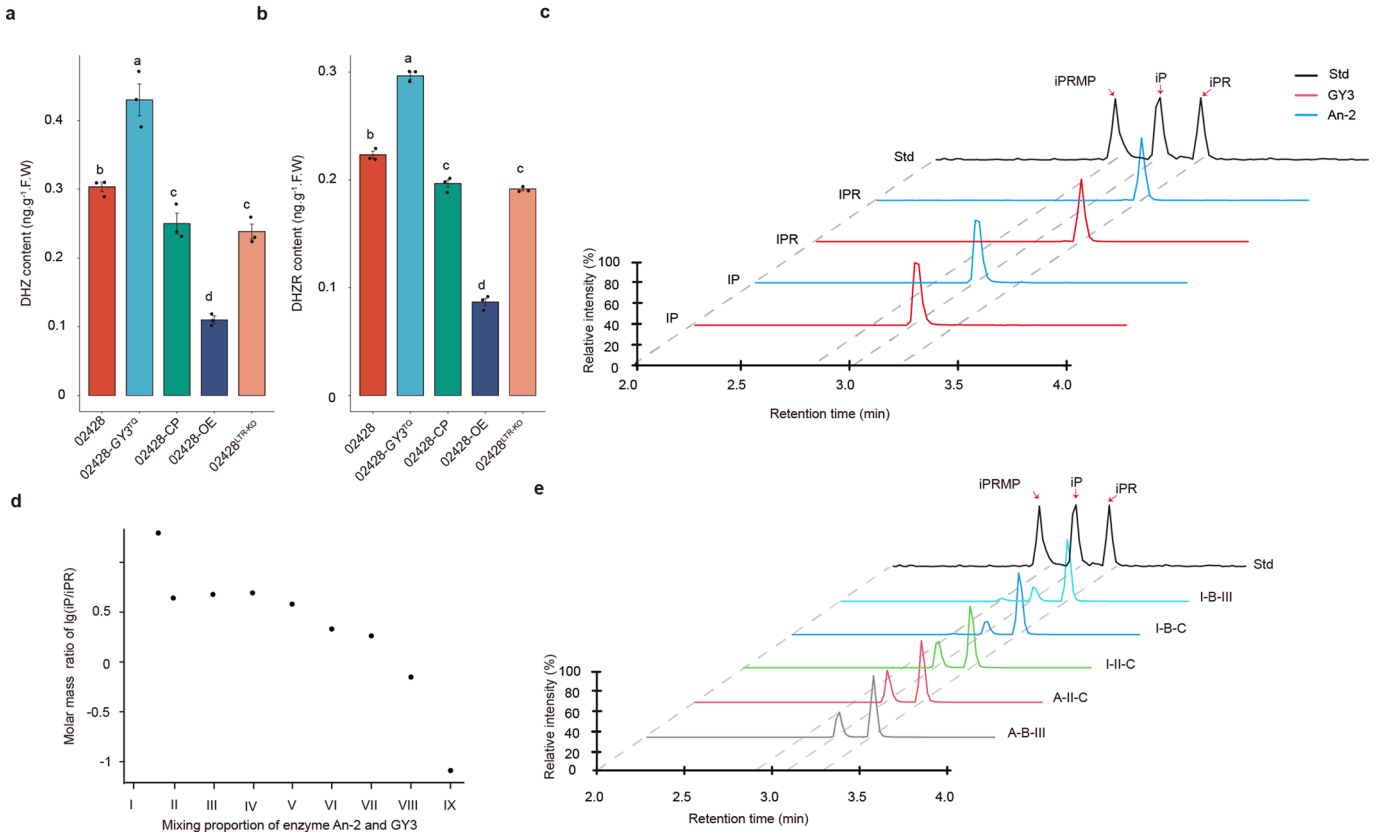
Extended Data Fig. 4 | Performances of the over-expressor of *GY3*. **a**–**d**, Comparisons of whole plant (**a**), panicle (**b**), primary branch (**c**) and panicle architecture (**d**) between 02428 and 02428-OE plant. Bars, 5 cm. (**e**) Expression level of *GY3* in the leaves of 02428 and 02428-OE plants at seedling stage. Data are means ± SE (n = 4). **f**–**q**, Performances of tillers per plant (**f**), primary branches per panicle (**g**), secondary branches per panicle (**h**), panicle length (**i**), spikelets per panicle (**j**), grains per panicle (**k**) 1000-grains per plant (**l**), seed setting rate (**m**), grain yield per plant (**n**), biomass per plant (**o**), harvest index (**p**) and plant height (**q**) of 02428 and 02428-OE plants. For box plots in **f**–**q**, the box limits indicate the 25th and 75th percentiles, whiskers further extend by ±1.5 times the interquartile range from the limits of each box, and the center line indicates the median, individual data points are plotted (n = 29, 30, 32 and 29 for 02428, 02428-OE_1, 02428-OE_2 and 02428-OE_3, respectively). *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



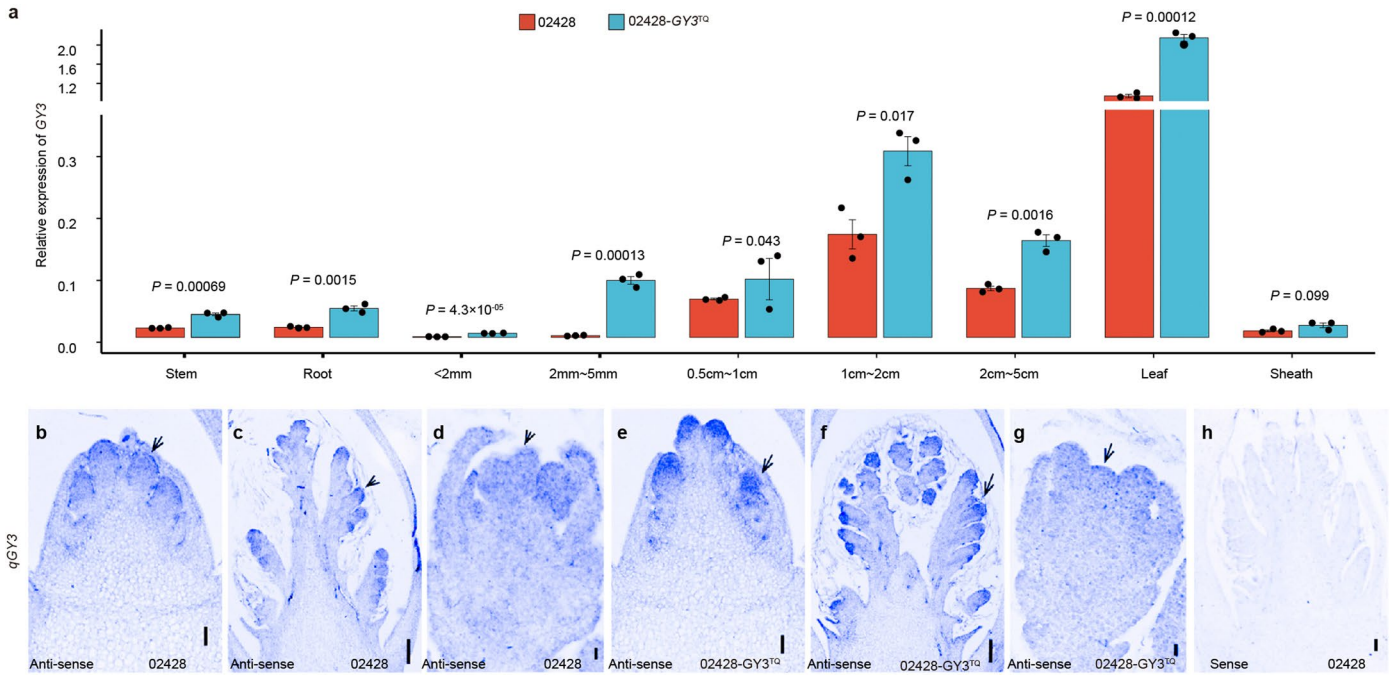
Extended Data Fig. 5 | Performances of loss-of-function of *GY3* in Zhonghua 11 (ZH11) background. a–**d**, Comparisons of whole plant (**a**), panicle (**b**), and panicle architecture (**c**) between ZH11 and ZH11$^{GY3\text{-}KO}$ plant. Bars, 5 cm. **d**–**h**, Performance of spikelets per panicle (**d**), primary branches per panicle (**e**), secondary branches per panicle (**f**), tillers per plant (**g**), panicle length (**h**) between ZH11 (n = 12) and ZH11$^{GY3\text{-}KO}$ (n = 13) T0 plants with varied insertions or deletions resulted *GY3* loss-of-function. Data are means ± SE. (**i**), The cytokinin contents in young leaves of ZH11 and ZH11$^{GY3\text{-}KO}$ plants. Data are means ± SE (n = 3). *P* values were calculated by two-sided paired Student's *t*-test. n.d. represents not detected.

https://doi.org/10.1038/s41588-023-01454-3



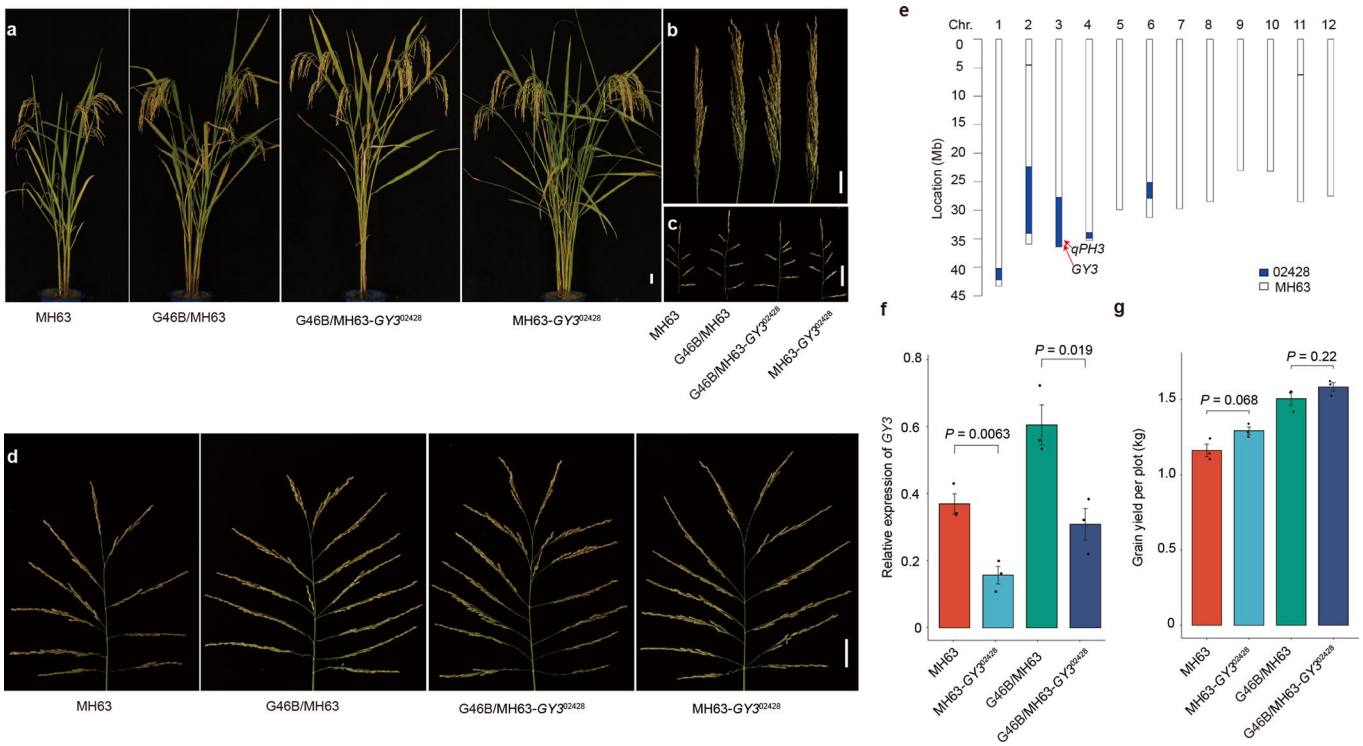
Extended Data Fig. 6 | Cytokinin content in young panicles and enzyme-catalysed reactions. a, **b**, Comparisons of cytokinin contents including DHZ (**a**) and DHZR (**b**) in young panicles of 02428, 02428-*GY3*^{TQ}, 02428-CP, 02428-OE, and 02428$^{LTR\text{-}KO}$. Data are means ± SE (three biological and three technical replicates). Different letters upon the bars indicated significant difference at $P < 0.01$ *via* Duncan test (**a** and **b**). **c**, Detection of the enzyme-catalysed reaction products of GY3 (red) and An-2 (blue) to iP and iPR, respectively. Std, standards of iPRMP, iP and iPR. **d**, Enzymatic product molar mass ratio of iP to iPR in An-2 (I), GY3 (IX) and mixtures with a molar mass ratio of An-2 to GY3 from 27:1 to 1:27 in 3-fold increments (from II to VIII with 27:1, 9:1, 3:1, 1:1, 1:3, 1:9, 1:27, respectively). **e**, Detection the enzyme-catalysed reaction products of newly recombinant enzymes (A-B-III, A-II-C, I-II-C, I-B-C, I-B-III). Std, standards of iPRMP, iP and iPR.

https://doi.org/10.1038/s41588-023-01454-3



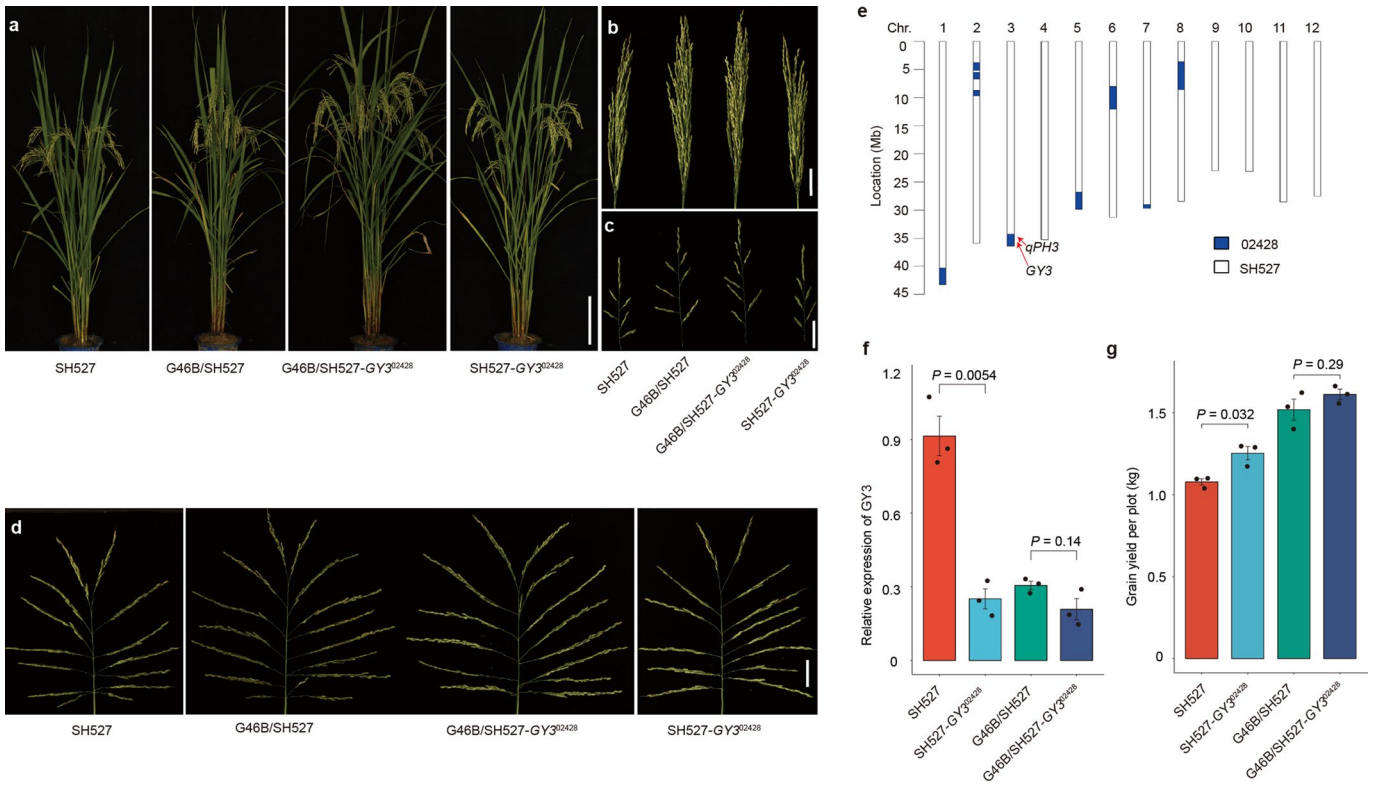
Extended Data Fig. 7 | Expression profiles of *GY3* in young panicle. a, The expression patterns of *GY3* in 02428 and 02428-*GY3*^{TQ} plants. Data are means ± SE (n = 3). *P* values were calculated by two-sided paired Student's *t*-test. **b**–**g**, RNA *in situ* hybridization of *GY3* by anti-sense probes in inflorescence meristems at primary branch initiating stage (**b** and **e**), secondary branching formation stage (**c** and **f**) and floral meristem developing stage (**d** and **g**) between 02428 (**b**, **c** and **d**) and 02428-*GY3*^{TQ} (**e**, **f** and **g**). (**h**), the signals of sense probe of *GY3* in 02428 were used as the negative control. Bars, 20 μm.

https://doi.org/10.1038/s41588-023-01454-3



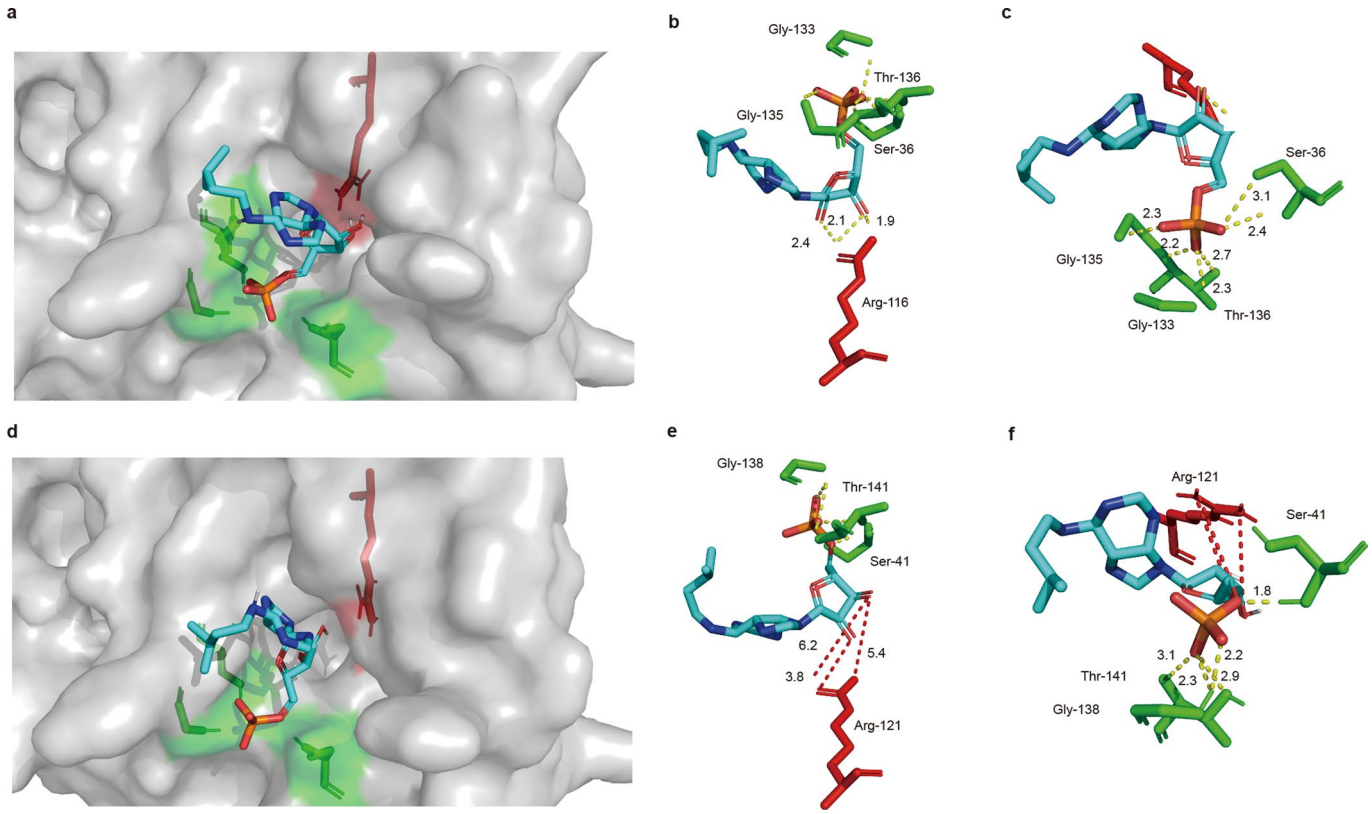
Extended Data Fig. 8 | Improvements of the yield performances of MH63 and its related hybrids by *GY3*^{02428}. **a**–**d**, Performances of whole plant (**a**), panicle (**b**), branch (**c**) and panicle architecture (**d**) of MH63, MH63-*GY3*^{02428} and their hybrids crossed with G46B at maturation stage. Bars, 5 cm. **e**, Genetic components of MH63-*GY3*^{02428} identified by the 6 K SNP array. **f**, Relative expression level of *GY3* in leaves at seedling stage. **g**, Comparison of grain yield per plot (each plot contained 40 plants) of MH63, MH63-*GY3*^{02428} and their hybrids crossed with G46B. Data are mean ± SE (n = 3). *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



Extended Data Fig. 9 | Improvements of the yield performances of SH527 and its related hybrids by *GY3*^{02428}. **a**–**d**, Performances of whole plant (**a**), panicle (**b**), branch (**c**) and panicle architecture (**d**) of SH527, SH527-*GY3*^{02428} and their hybrids crossed with G46B at maturation stage. Bars, 5 cm. **e**, Genetic components of SH527-*GY3*^{02428} identified by the 6 K SNP array. **f**, Relative expression level of *GY3* in leaves at seedling stage. **g**, Comparisons of grain yield per plot (each plot contained 40 plants) of SH527, SH527-*GY3*^{02428} and their hybrids crossed with G46B. Data are mean ± SE (n = 3); *P* values were calculated by two-sided paired Student's *t*-test.

https://doi.org/10.1038/s41588-023-01454-3



Extended Data Fig. 10 | Structure analysis of An-2 and GY3. Illustration of the substrate-binding cavities of An-2 (**a**) and GY3 (**d**) with docking its substrate iPRMP, the structure of An-2 and GY3 are predicted by AlphaFold. (**b** and **e**) show the hydrogen bond and its lengths between riboside of iPRMP and its interaction amino acid residues in An-2 (**b**) and GY3 (**e**). (**c** and **f**) show the hydrogen bond and its lengths between the phosphate of iPRMP and its interaction amino acid residues in An-2 (**c**) and GY3 (**f**). Residues involved in enzyme catalysis, AMP-binding residues show as red and green sticks, respectively. The yellow dash lines represent the hydrogen bond and the number adjacent to each bond represent hydrogen bond lengths. The red dash lines (**e** and **f**) represent lengths between atoms in GY3 catalysis residue for the corresponding hydrogen bond formation in An-2.

nature portfolio

Corresponding author(s): Yongzhong Xing

Last updated by author(s): Jan 17, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	the gene expression levels were performed on QuantStudio 6 Flex Real-Time PCR System (Applied Biosystems, CA, USA); enzymatic reaction products were measured with a LC/MS system (Agilent 6520 Accurate-Mass Q-TOF LC/MS; Agilent Zorbax C18, 4.6 mm 350 mm); the eChIP-seq were performed on MGI (NDM607-01).
Data analysis	ESPript 3.0 software for LOG protein sequence aligned; Fastp (version 0.20.1) for data quality evaluation; Bismark (v0.23.1) for WGBS reads mapping; Bowtie2 (v2.4.5) for eChIP-seq mapping; SAMtools v1.10 for BAM files process; "picard MarkDuplicates" in GATK (v4.2.2.0) used for mark and remove the duplicate reads; MACS2 (v2 2.2.7.1) used for eChIP-seq peak calling; AutoDockTools-(v1.5.7) and vina (v.1.2.0) used for docking; Burrows–Wheeler Aligner (BWA; v0.7.17) used for paired-end reads mapping; Manta (v1.6.0) used for calling large structural variants and InDels; Beagle (v5.1) used for SNPs phasing; vcflib (v1.0.1) used for haplotype blocks extracted; R (v4.0.2) used for statistical analyses and figures draw.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Sequence data from this study can be found in GenBank under accession number PRJEB6180, SRX502298–SRX502317 and SRX502162–SRX502255 for 3,010 diverse accessions, 20 African cultivated rice accessions and 94 accessions of Oryza barthii, respectively. The eChip-seq sequence data and bed files were deposited in the NCBI Gene Expression Omnibus with accession number GSE231360. The reference genome of MH63RS2 and Nipponbare can be found in Rice Information GateWay (http://rice.hzau.edu.cn/rice_rs2/) and The Rice Annotation Project (https://rapdb.dna.affrc.go.jp/). All data are available in the main text or the supplementary materials.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not involved.
Population characteristics	not involved.
Recruitment	not involved.
Ethics oversight	not involved.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size of all experimental was determined based on the feasibility of sampling and the need of statistical analysis. All replicates represent biologically independent samples. No sample size calculation was performed.
Data exclusions	No data was excluded from the analyses.
Replication	Conclusions drawn in this study are based on at least three independent experiments. The number of biological replicates in each experiment is indicated in the figure legends. For each materials, the border rows were removed and the middle plants were selected to conduct phenotyping.
Randomization	We always randomly selected the samples based on the genotype for experiments.
Blinding	For all experiments, we always performed three or more than three replicates or measurements without human interference, therefore, we did not perform blinding analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used: H3K4me3 (ABclonal, A2357), H3K9me2 (Abcam, ab1220), H3K27me3 (ABclonal, A2363), and H3K27ac (ABclonal, A7253).

Validation: The H3K4me3 (ABclonal, A2357), H3K27me3 (ABclonal, A2363), and H3K27ac (ABclonal, A7253) antibodies were bought from ABclonal (https://abclonal.com/). The H3K9me2 (Abcam, ab1220) antibody was bought from Abcam(https://www.abcam.cn/products/primary-antibodies/histone-h3-di-methyl-k9-antibody-mabcam-1220-chip-grade-ab1220.html). Furthermore, the validation of each antibody can be found in Zhao et, al. (Zhao, L. et al. Integrative analysis of reference epigenomes in 20 rice varieties. Nat Commun 11, 2658 (2020). https://doi.org:10.1038/s41467-020-16457-5).

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links
May remain private before publication.

Data deposited at https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE231360；

Files in database submission

a list of data list as below: Wb10_1.fq.gz; E100059645_L01_65_1.fq.gz; Wb9_1.fq.gz; E100059645_L01_66_1.fq.gz; Wb1_1.fq.gz; E100059645_L01_57_1.fq.gz; Wb2_1.fq.gz; E100059645_L01_58_1.fq.gz; Wb3_1.fq.gz; E100059645_L01_59_1.fq.gz; Wb4_1.fq.gz; E100059645_L01_60_1.fq.gz; Wb7_1.fq.gz; E100059645_L01_63_1.fq.gz; Wb8_1.fq.gz; E100059645_L01_64_1.fq.gz; Wb5_1.fq.gz; E100059645_L01_61_1.fq.gz; Wb6_1.fq.gz; E100059645_L01_62_1.fq.gz; Wb10_2.fq.gz; E100059645_L01_65_2.fq.gz; Wb9_2.fq.gz; E100059645_L01_66_2.fq.gz; Wb1_2.fq.gz; E100059645_L01_57_2.fq.gz; Wb2_2.fq.gz; E100059645_L01_58_2.fq.gz; Wb3_2.fq.gz; E100059645_L01_59_2.fq.gz; Wb4_2.fq.gz; E100059645_L01_60_2.fq.gz; Wb7_2.fq.gz; E100059645_L01_63_2.fq.gz; Wb8_2.fq.gz; E100059645_L01_64_2.fq.gz; Wb5_2.fq.gz; E100059645_L01_61_2.fq.gz; Wb6_2.fq.gz; E100059645_L01_62_2.fq.gz

Genome browser session
(e.g. UCSC)

Not involved

Methodology

Replicates: Each sample was conducted two independent replicates.

Sequencing depth: More than x10 depth for each sample

Antibodies: H3K4me3 (ABclonal, A2357), H3K9me2 (Abcam, ab1220), H3K27me3 (A2363), and H3K27ac (ABclonal, A7253).

Peak calling parameters: The peak calling used the default parameters

Data quality: The peaks called in 02428_ H3K4me3 (25165 peaks and 26256 peaks), 02428_H3K9me2 (26783 peaks and 24837 peaks), 02428_H3K27me3 (14389 peaks and 13303 peaks), 02428_H3K27ac (29195 peaks and 28242 peaks),NIL-GY3_ H3K4me3 (26990 peaks and 27109 peaks), NIL-GY3_H3K9me2 (50904 peaks and 52004 peaks), NIL-GY3_H3K27me3 (16515 peaks and 15867 peaks), NIL-GY3_H3K27ac (35258 peaks and 34494peaks) at FDR 5% .

Software: Bowtie2 (v2.4.5) for eChIP-seq mapping; SAMtools v1.10 for BAM files process; “picard MarkDuplicates” in GATK (v4.2.2.0) used for mark and remove the duplicate reads; MACS2 (v2 2.2.7.1) used for eChIP-seq peak calling