

# Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding

Received: 4 April 2022

Accepted: 21 September 2022

Published online: 29 September 2022

 Check for updates


Jiantao Guan<sup>1,2,3,10</sup>, Jintao Zhang<sup>1,4,10</sup>, Dan Gong<sup>1,4,10</sup>, Zhengquan Zhang<sup>2</sup>, Yang Yu<sup>2</sup>, Gaoling Luo<sup>5</sup>, Prakrit Somta<sup>6</sup>, Zheng Hu<sup>1</sup>, Suhua Wang<sup>1</sup>, Xingxing Yuan<sup>7</sup>, Yaowen Zhang<sup>8</sup>, Yanlan Wang<sup>9</sup>, Yanhua Chen<sup>5</sup>, Kularb Laosatit<sup>6</sup>, Xin Chen<sup>7</sup>, Honglin Chen<sup>1</sup>, Aihua Sha<sup>4</sup>, Xuzhen Cheng<sup>1</sup>, Hua Xie<sup>2</sup>  & Lixia Wang<sup>1</sup> 

Rice bean (*Vigna umbellata*) is an underexploited domesticated legume crop consumed for dietary protein in Asia, yet little is known about the genetic diversity of this species. Here, we present a high-quality reference genome for a rice bean landrace (FF25) built using PacBio long-read data and a Hi-C chromatin interaction map, and assess the phylogenetic position and speciation time of rice bean within the *Vigna* genus. We sequence 440 landraces (two core collections), and GWAS based on data for growth sites at three widely divergent latitudes reveal loci associated with flowering and yield. Loci harboring orthologs of *FUL* (*FRUITFULL*), *FT* (*FLOWERING LOCUS T*), and *PRR3* (*PSEUDO-RESPONSE REGULATOR 3*) contribute to the adaptation of rice bean from its low latitude center of origin towards higher latitudes, and the landraces which pyramid early-flowering alleles for these loci display maximally short flowering times. We also demonstrate that copy-number-variation for *VumCYP78A6* can regulate seed-yield traits. Intriguingly, 32 landraces collected from a mountainous region in South-Central China harbor a recently acquired InDel in *TFL1* (*TERMINAL FLOWER1*) affecting stem determinacy; these materials also have exceptionally high values for multiple human-desired traits and could therefore substantially advance breeding efforts to improve rice bean.

The genus *Vigna* is a pan-tropical genus in the family Fabaceae, comprising more than 100 wild species and 10 domesticated species such as cowpea (*Vigna unguiculata*), mung bean (*V. radiata*), and rice bean (*V. umbellata*)<sup>1</sup>. As one of the representative species in the genus *Vigna*, the rice bean is a multipurpose legume and is widely cultivated in South, Southeast, and East Asia<sup>2</sup>. The seeds of rice beans have been consumed for thousands of years as a good source of dietary protein and micronutrients, and these are used as a diuretic in traditional medicine practices<sup>3,4</sup>. Rice bean has also been widely used as a donor parent for interspecific hybridization with other species in the genus *Vigna*<sup>5-7</sup> due to its notable agronomic characteristics including high

grain yield and large biomass potential<sup>2,8</sup>, as well as strong resistance to pests<sup>9-14</sup>, diseases<sup>15</sup>, drought<sup>2,16,17</sup>, water logging<sup>18</sup>, and capacity to grow in poor fertility soils<sup>19</sup>. Thus, as the continually growing population and exacerbated climate changes, rice bean has received increased attention in recent years and has been proposed as one of the potential future smart foods to help to fight hunger and malnutrition in Asia<sup>2,18,20</sup>. However, the lack of a high-quality reference genome for rice beans has hindered the exploration of the genetic basis of these excellent agronomic characteristics and its further genetic improvement.

Current thinking holds that the rice bean originated and was domesticated in tropical regions of South & Southeast Asia, after

A full list of affiliations appears at the end of the paper.  e-mail: [xiehua@baafs.net.cn](mailto:xiehua@baafs.net.cn); [wanglixia03@caas.cn](mailto:wanglixia03@caas.cn)

which it spread to higher latitude regions including China, Japan, and Korea<sup>2,21,22</sup>. There are many rice bean landraces that have, through long-term human and natural selection, become locally adapted to diverse environments. However, as a short-day plant, the yield potential and agricultural utility of rice beans can be strongly affected by photoperiod and temperature conditions<sup>2,23,24</sup>. Moreover, few cultivated rice bean varieties have a determinate stem growth habit that influences the potential grain yield and is also required to support mechanical harvest<sup>6,25</sup>. Landraces have been demonstrated as useful resources for the improvement of diverse crop species<sup>26</sup>, and there are presently two rice bean core collections, one comprising mainly landraces from South & Southeast Asia and the other with a preponderance of Chinese rice bean landraces<sup>22,27,28</sup>. Thus, there are rich germplasm panels available representing the high diversity and broad adaptation of rice beans to both tropical and temperate environments.

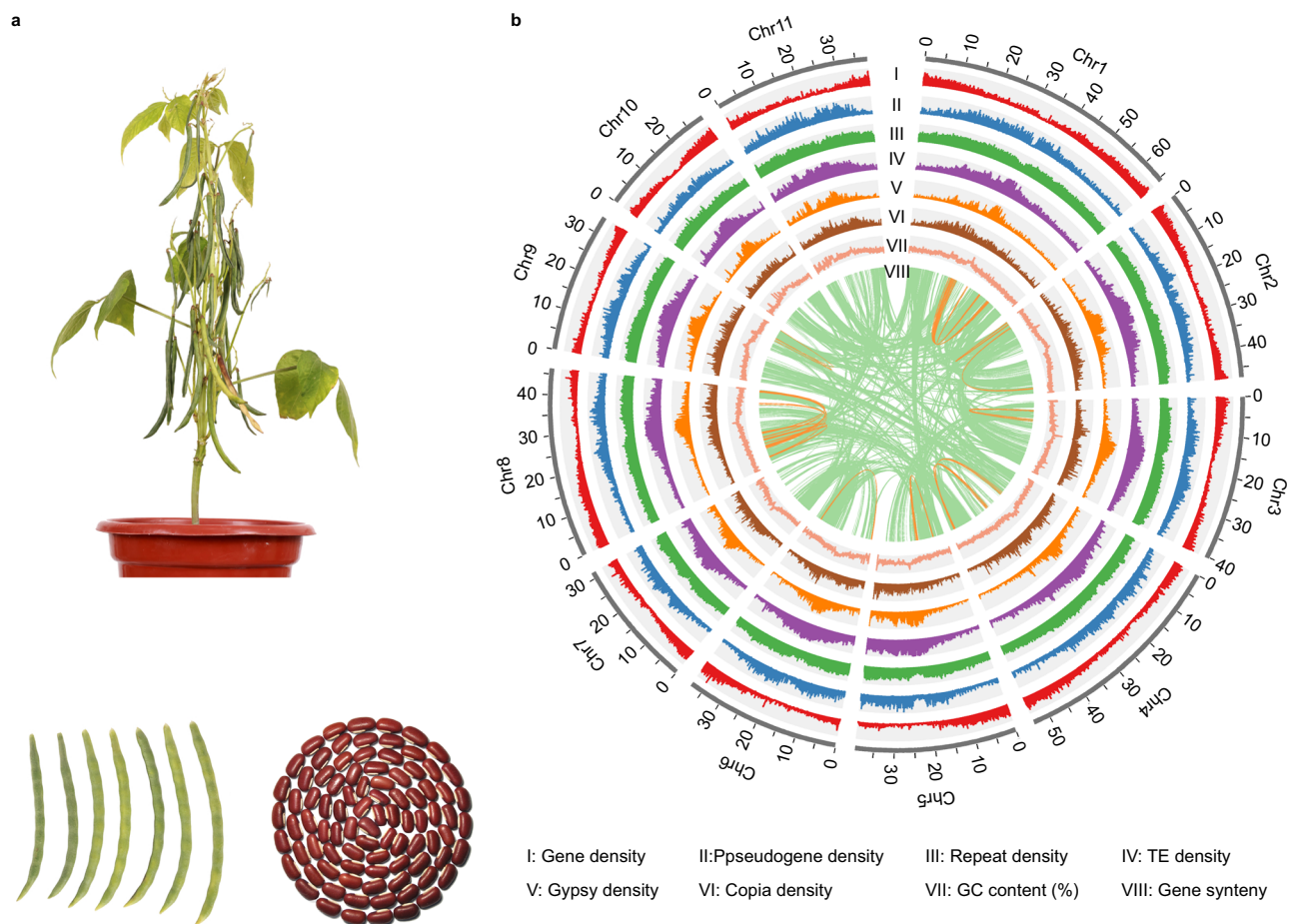
Previous studies have reported several QTLs for adaptation and yield component-related traits using linkage mapping based on biparental populations in rice bean<sup>11,29</sup>. However, the resolution and sensitivity have been limited by the small number of markers and genetic recombination, thus making it difficult to reveal the genetic mechanism of these traits and/or to develop breeding markers<sup>30,31</sup>. Genome-wide association studies (GWAS) have been successfully applied in crops for the efficient identification of favorable alleles/haplotypes or causal variants/genes underlying complex traits as this strategy could simultaneously detect many natural allelic variations using a diverse germplasm panel<sup>32,33</sup>.

Here, we present a high-quality reference genome assembly of rice beans based on the integration of Illumina short-reads, PacBio long-reads, and Hi-C sequencing data. We also construct a genome variation map based on sequencing of 440 diverse rice bean landraces covering two core collections. Subsequent population genomic analyses support the previously proposed origin of rice bean in South & Southeast Asia and revealed genetic bottlenecks that occurred along the northward dispersal of rice bean. GWAS based on phenotypic data for a germplasm diversity panel grown at three sites with widely divergent latitudes helps decipher the genetic basis of traits including flowering time, seed yield, and stem determinacy. Our study also identifies candidate genes and landraces with strong potential as elite germplasm lines that could be used to generate excellent varieties that simultaneously display geographically suitable flowering times, stem determinacy to support mechanized cultivation, and high yields of rice beans.

## Results

### Sequencing and assembly of a reference genome for rice bean

The rice bean landrace FF25—which has red seeds, an erect habit, and wide environmental adaptability—was selected for genome sequencing and de novo assembly of a rice bean reference genome (Fig. 1a). We integrated three sequencing technologies: PacBio single molecule real-time (SMRT) long-read sequencing, Illumina short-read sequencing, and chromosome conformation capture sequencing data (Hi-C) (Supplementary Table 1). The estimated genome size of the FF25 genome was -525.60 Mb based on 17-kmer depth distribution using



**Fig. 1 | FF25 genome assembly.** **a** FF25 plant (Top); FF25 pod (Bottom left); FF25 seeds (Bottom right). **b** Genomic features of the FF25 reference genome. The outer gray track represents the chromosomes of the FF25 genome assembly (with

units in Mb). The densities of features were calculated based on 100 kb window size, with a step size of 10 kb. The inner green and orange links represent the intra- and inter-chromosomal collinear genes, respectively. Photograph credit: LXW (a).

**Table 1 | Summary statistics for the rice bean genome assembly**

Genomic feature	Value
Total assembly size (Mb/%)	475.64/90.49%
Number of contigs	351
Largest contigs (Mb)	32.05
Contig N50 (Mb)	18.26
Sequences anchored to chromosomes (Mb)	465.19/97.80%
Genomic GC content (%)	34.21
Genome Complete BUSCOs <sup>a</sup> (%)	97.3
Protein Complete BUSCOs <sup>a</sup> (%)	96.9
LTR assembly index, LAI	20.30
Repetitive sequences (%)	57.19
Protein-coding genes	26,736
Mean gene length (bp)	3,602
Mean coding sequences/exon/intron length (bp)	1232/238/570

<sup>a</sup>Analysis based on comparisons with the eudicotyledons\_odb10 database.

Illumina short-reads (~106.65×; Supplementary Fig. 1). The PacBio reads (~300.54×) were used to assemble the contigs using Canu v1.9<sup>34</sup> and the highly efficient repeat assembly (HERA) algorithm<sup>35</sup>, which resulted in a 475.64 Mb genome (90.49% of the estimated size) containing 351 contigs, with an N50 of 18.26 Mb (Table 1), thus representing highest quality genome among species of the *Vigna* genus<sup>36–40</sup>. To assign the contigs to different chromosomes, 66 contigs (~465.19 Mb, 97.80% of the original assembly) were anchored to eleven pseudo-chromosomes based on a Hi-C interaction map (Table 1; Supplementary Fig. 2).

We used multiple methods to evaluate the quality of the assembled genome. The mapping and coverage rates of the Illumina short-read data were 99.67% and 99.33%, respectively. We further performed benchmarking universal single-copy orthologs (BUSCO) analysis<sup>41</sup> based on the eudicotyledons\_odb10 dataset, and the result showed that 97.3% of the BUSCO sequences were completely present in the genome assembly, while 0.5% and 2.2% were partially present or missing, respectively (Table 1; Supplementary Table 2). The genome assembly had a high LTR Assembly Index (LAI) score (20.30) (Supplementary Table 2; Supplementary Fig. 3), reaching the level of a gold standard genome assembly according to previously proposed criteria<sup>42</sup>. All of these lines of evidence indicate that our de novo FF25 genome assembly is of high quality.

We used an integrated strategy including evidence-based methods and ab initio gene prediction to annotate the protein-coding gene content of the FF25 genome assembly. A final set of 26,736 protein-coding genes was predicted, of which 26,430 genes (~98.86%) could be assigned to eleven pseudomolecules (Supplementary Table 3). Of these genes, the average lengths of coding sequences, exons, and introns were 1232 base pairs (bp), 238, and 570 bp, respectively (Table 1). The average gene density was one gene per 17.79 Kb, and the genes were unevenly distributed, being more abundant towards the chromosomal ends (Fig. 1b). We also specifically concatenated 2202 transcription factor genes, 9635 pseudogenes, and 3318 noncoding RNA genes comprising 764 transfer RNA genes, 558 ribosomal RNA genes, 714 small nucleolar RNA genes, and 1282 microRNA genes (Fig. 1b; Supplementary Table 4).

Of these predicted protein-coding genes, we found that 96.90% of the BUSCO sequences were completely present (Table 1; Supplementary Table 2). Moreover, the tissue-specific RNA-Seq data confirmed that 85.86% of the predicted protein-coding genes were expressed (FPKM > 1) in at least one of the 6 examined tissues (Supplementary Table 5). And 97.48% of the protein-coding genes were assigned a functional annotation based on five public databases (Supplementary

Table 6). These evaluations collectively support the high accuracy and completeness of our rice bean genome assembly and annotation.

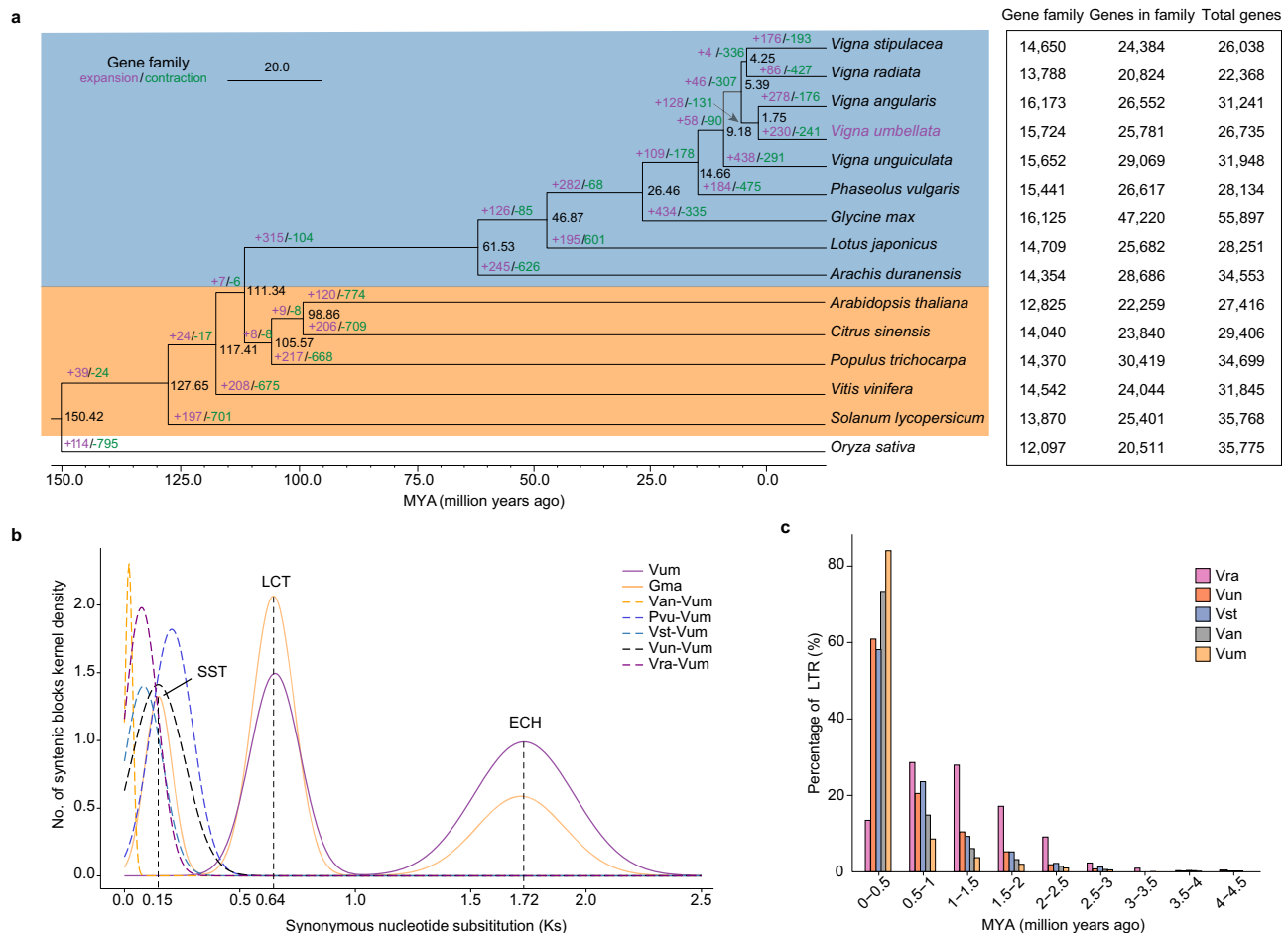
### Phylogenetic position and comparative genomics analyses

To explore the genome evolution of rice bean, genes from the five *Vigna* species (*Vigna stipulacea*, *V. radiata*, *V. angularis*, *V. umbellata*, and *V. unguiculata*), four other legumes (*Phaseolus vulgaris*, *Glycine max*, *Lotus japonicus*, and *Arachis duranensis*), five other eudicots (*Arabidopsis thaliana*, *Citrus sinensis*, *Populus trichocarpa*, *Vitis vinifera*, and *Solanum lycopersicum*), as well as one monocot (*Oryza sativa*) were clustered into 20,736 gene families. Of these, 334 single-copy gene families were used to construct a maximum-likelihood phylogenetic tree (Fig. 2a). This indicated that rice bean is a sister species to adzuki bean (*V. angularis*); they apparently diverged about 1.75 million years ago (MYA), findings in accord with a previous study based on transcriptome data<sup>37</sup>.

This view was also supported by a gene synteny analysis between rice bean and its closely related species in the *Vigna* genus based on protein sequences using the MCScanX program<sup>43</sup>, which revealed that (as expected) rice bean had higher conservation with *V. angularis* in terms of gene structure and order as compared to other *Vigna* species (Supplementary Fig. 4; Supplementary Table 7). Based on the tree, we found that 230 rice bean gene families (comprising 1396 genes) exhibited significant expansions ( $P < 0.01$ ) relative to the MRCA (most recent common ancestor) of rice bean and adzuki bean (Supplementary Data 1). KEGG pathway analysis indicated that these expanded genes were significantly enriched for metabolism pathways such as the phenylpropanoid, sesquiterpenoid, and triterpenoid biosynthesis ( $P < 0.05$ , Fisher's exact test; Supplementary Fig. 5).

Whole-genome duplication (WGD) provides additional genetic material that can be subsequently subjected to divergence, sub-functionalization, and neofunctionalization<sup>44,45</sup>. To investigate WGD events in rice bean, we identified 332 syntenic blocks within its genome (including 8052 homologous genes accounting for ~30.12% of all genes) (Fig. 1b) and estimated synonymous nucleotide substitutions at synonymous sites (Ks) for homologs. The Ks distribution of collinear gene pairs indicated no recent WGD in rice beans; we also observed the expected signals for the ECH event (eudicot-common hexaploidy; Ks = 1.72) and the LCT (legume-common tetraploid; Ks = 0.64) event (Fig. 2b). We estimated the relative time of evolutionary divergence between rice bean and closely related *Vigna* species using the Ks distributions of orthologs based on the known evolutionary time (~13 MYA) of the SST (soybean-specific tetraploid) event in soybean<sup>37,46</sup>. Similar to the very recent speciation time estimated from the maximum-likelihood phylogenetic tree (Fig. 2a), the Ks distribution of rice bean and adzuki bean also showed the smallest peak value at 0.019 (Fig. 2b), corresponding to a divergence time of 1.72 MYA.

Beyond comparisons of orthologous genes, we annotated the repetitive content in the rice bean genome using an integrated pipeline, including de novo repeat identification and homology search methods (see the “Methods” section). We identified that 38.40% of the rice bean genome comprises transposable elements (TEs; Supplementary Data 2). Among the distinct classes of TEs, LTR elements including *Gypsy* and *Copia* elements were the predominant classes; and compared to *Copia* elements (10.41%), *Gypsy* elements (19.85%) occupied relatively larger proportions of genomic sequence in rice bean, which is consistent with earlier reports about other *Vigna* species<sup>37,40</sup>. In addition, we identified full-length LTR elements and performed an insert time analysis for rice bean as well as other four additional *Vigna* species with sequenced genome assemblies. Excepting *Vigna radiata*, more than half of the LTR elements in the other four examined *Vigna* species proliferated at 0 – 0.5 MYA, suggesting that the amplification of LTR elements has largely occurred after speciation (Fig. 2c).



**Fig. 2 | Phylogenetic position and comparative genomics analyses.** **a** Genome evolution and gene family characteristics of *Vigna umbellata* (rice bean) and 13 other dicot species using the monocot plant *Oryza sativa* (rice) as an out-group. This tree was generated using 334 single-copy ortholog families. Black numerical values beside each node show the estimated divergence time of each node (MYA, million years ago) in the phylogenetic tree shown on the left. Blue and orange backgrounds represent *Leguminosae* and non-*Leguminosae* species, respectively. The number of gene families, genes in the family, and the total number of genes are

shown on the right for each species. **b** Density distribution of synonymous nucleotide substitution levels (Ks) of syntenic orthologous (solid curves) and paralogous genes (dashed curves). Vum: *Vigna umbellata*; Gma: *Glycine max*; Van: *V. angularis*; Pvu: *Phaseolus vulgaris*; Vst: *V. stipulacea*; Vun: *V. unguiculata*; Vra: *V. radiata*. **c** Insertion bursts of full-length LTR elements in the genomes of *V. umbellata* and other four *Vigna* species. Source data are provided as a Source Data file.

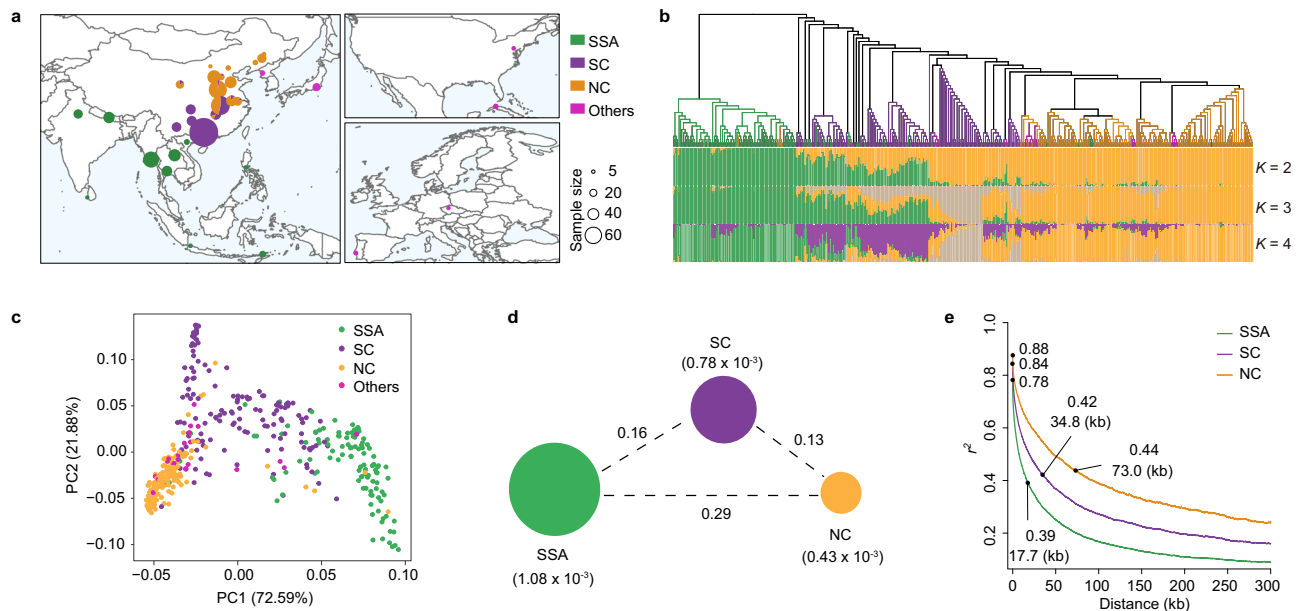
## Population structure and genetic divergence of rice bean landraces

We performed whole genome re-sequencing for a total of 440 rice bean landraces from various geographic regions, including the landraces in the Asia core collection (73) and Chinese core collection (230) using Illumina sequencing technology (Fig. 3a), ultimately generating 5.32 Tb of high-quality sequencing data, with an average depth of  $\sim 24.91\times$  and an average mapping rate of 99.12% based on the newly assembled reference genome (Supplementary Data 3). A final set of 10,525,548 high-quality single-nucleotide polymorphisms (SNPs) and 2,743,289 small insertions and deletions (InDels) were identified. We found 5690 SNPs (0.054%) that caused start codon changes, premature stop codons, or elongated transcripts, while 15,530 InDels (0.57%) lead to frameshift mutations (Supplementary Table 8), proportions similar to other species likely soybean<sup>47</sup>, cucumber<sup>48</sup>, and watermelon<sup>49</sup>.

To infer the population structure, we constructed an SNP-based neighbor-joining (NJ) phylogenetic tree and divided the 440 landraces into three geographical groups: landraces from South & Southeast Asia (SSA), South China (SC; coastline of South China to the Yangtze River), and North China (NC; Yangtze River to North China) (Fig. 3a, b; Supplementary Data 3). This classification was supported by a principal

component analysis (Fig. 3c) as well as a model-based clustering analysis ( $K = 4$ ) conducted using STRUCTURE<sup>50</sup> (Fig. 3b). Notably, the landraces collected from other geographical regions (Japan, Korea, Europe, and America) were spread amongst the SC and NC groups, indicating their close genetic relationship with Chinese landraces or their probable introduction from China<sup>2</sup>. We excluded these landraces from the SC and NC groups in our further analyses.

To investigate genetic diversity and divergence among the three geographical groups, we calculated the nucleotide diversity ( $\pi$ ) for each group and conducted a pairwise analysis of genetic distances (Fixation index values,  $F_{ST}$ ). The SSA group showed the highest nucleotide diversity ( $1.08 \times 10^{-3}$ ), consistent with the previous results using SSR markers<sup>22</sup> and further supporting the hypothesis that rice beans originated from South & Southeast Asia<sup>2,22</sup>. Compared with the SSA group, gradually decreased nucleotide diversity was observed in the SC group ( $0.78 \times 10^{-3}$ ) and then the NC group ( $0.43 \times 10^{-3}$ ), indicating that sequential bottlenecks ( $\pi_{SSA}/\pi_{SC} = 1.38$ ;  $\pi_{SC}/\pi_{NC} = 1.81$ ) occurred during the northward dispersal of rice bean from the origin center (Fig. 3d). When compared with the SSA group, the  $F_{ST}$  value of the SC group was 0.16, whereas it became higher (0.29) for the NC group, indicating enlarged population differentiation during the northward dispersal (Fig. 3d).



**Fig. 3 | Population structure and genetic divergence of rice bean landraces.** **a** The geographic distributions of 440 rice bean landraces. SSA South & Southeast Asia, SC South China, NC North China. The size and color of each pie chart represent the sample size in a specific geographic location. The map was created using the `map_data()` function in the R package `ggplot2`. **b** Phylogenetic tree and model-based clustering ( $K=2-4$ ) of 440 sequenced landraces. **c** Scores plot from a principal component analysis, supporting the division of the landraces into three geographical groups (SSA, SC, and NC). **d** Summary of nucleotide diversity ( $\pi$ ) and

population divergence ( $F_{ST}$ ) across the three geographical groups. Values in parentheses represent measures of nucleotide diversity of each group, and values between pairs indicate population divergence. **e** Decay of linkage disequilibrium (LD), measured by  $r^2$ , in the three geographical groups. The upper and lower black dots with numerical values in the lines represented maximum and median values of the  $r^2$  and the corresponding physical distances. Source data are provided as a Source Data file.

We further examined linkage disequilibrium (LD) using the measure ( $r^2$ )<sup>51</sup> between pairwise SNP loci in SSA, SC, and NC groups. For the SSA group, the decay of LD with physical distance (i.e., a drop to half of its maximum value) between SNPs occurred at only  $-7.7$  kb ( $r^2 = 0.39$ ), whereas it increased to  $-34.8$  kb ( $r^2 = 0.42$ ) in the SC group and to  $-73.0$  kb ( $r^2 = 0.44$ ) in the NC group (Fig. 3e); these trends are in accord with the observed gradual reduction in genetic diversity in the SC and NC groups. The LD of rice bean landraces was similar to those of outcrossing species such as maize (30 kb)<sup>52</sup> but shorter than those of inbreeding crops like soybean (83 kb)<sup>53</sup>, rice (123 kb and 167 kb in *indica* and *japonica*, respectively)<sup>54</sup>, and foxtail millet ( $-100$  kb)<sup>55</sup>. This finding is consistent with a previous report that rice bean has a fairly high outcrossing rate<sup>22</sup>. Notably, the relatively rapid LD decay in the rice bean landraces may be useful for enhancement of resolution power of association studies to map a narrow candidate QTL interval<sup>56</sup>.

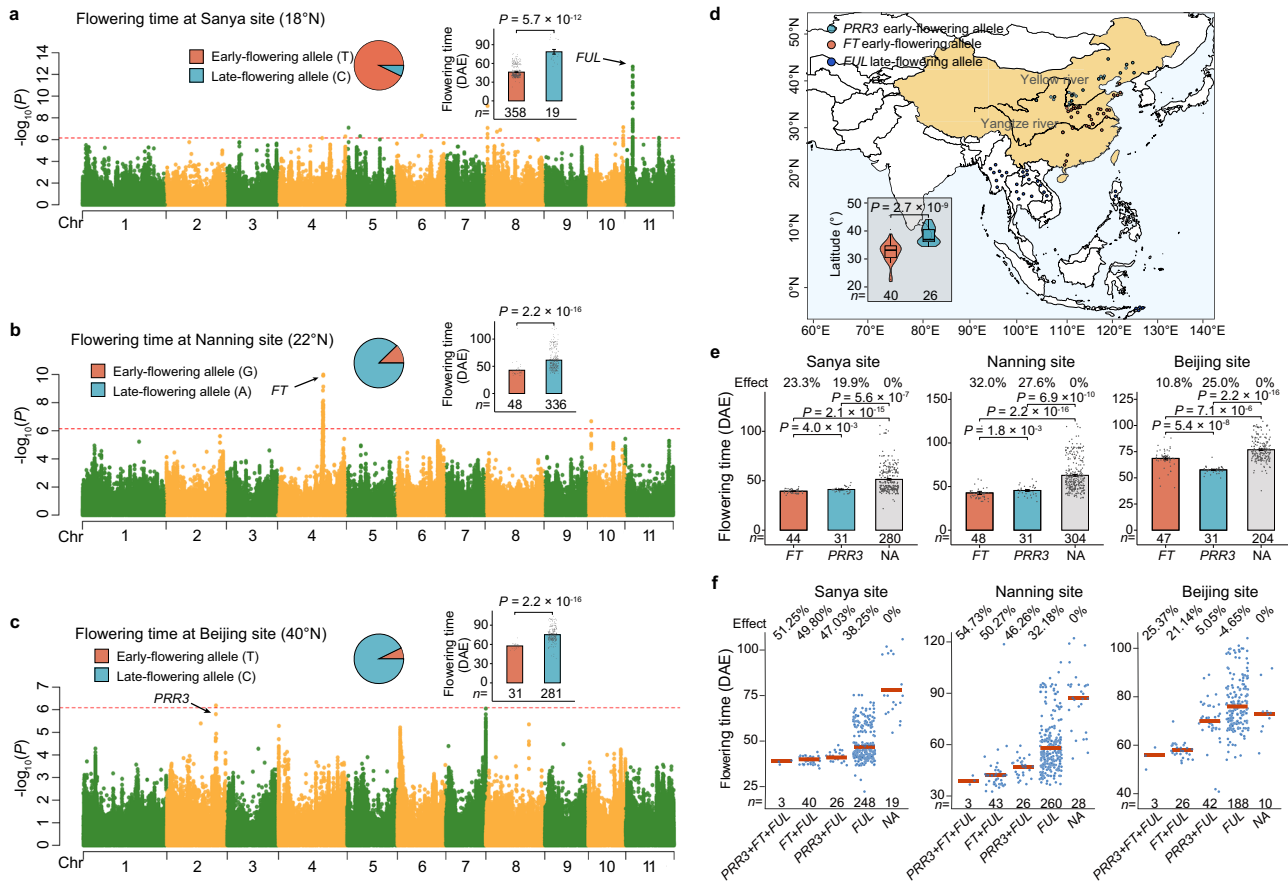
We searched for putatively selective regions with outliers (top 5%) of  $F_{ST}$  over 20-kb windows for the three comparisons (SSA vs. SC, SSA vs. NC, and SC vs. NC). We detected 473, 512, and 444 outlier regions for these three comparisons, respectively occupying 5.67% (26.95 Mb), 5.92% (28.15 Mb), and 5.59% (26.57 Mb) of the genome and including 1894, 1950, and 1296 protein-coding genes (Supplementary Data 4). A MapMan analysis of all the selected genes indicated that these genes were significantly enriched for annotations related to biological processes such as “phytohormone action”, “nutrient uptake”, and “circadian clock system” (Supplementary Fig. 6). Notably, among the genes related to “circadian clock system”, we found four orthologs of reported flowering time genes in *A. thaliana* using FLOWerRing Interactive Database (FLOR-ID<sup>57</sup>), including *TOC1*<sup>58</sup>, *PRR3*<sup>59</sup>, and two *LHY1*<sup>60</sup> genes between the SSC and SC groups, of which one *LHY1*<sup>60</sup> apparently also underwent selection between the SSC and NC groups (Supplementary Fig. 7). These flowering time genes could plausibly have contributed to the adaptation of rice bean landraces to different latitudes.

### The genetic architecture underlying control of flowering at different latitudes

We observed flowering time variation across 440 landraces as grown at three sites with widely divergent latitudes: 22–106 days in Sanya (18°N) in 2020 and 2021, 25–122 days in Nanning (22°N) in 2020 and 2021, and 38–104 days in Beijing (40°N; where some landraces did not bloom before the first frost in the autumn of 2020 and 2021). To explore the genetic basis of the flowering time for rice beans, we performed GWAS for the flowering phenotype data measured in both years at the three sites, which revealed distinct association signals for the different locations (Fig. 4a–c). The repeatedly detected major signal from Sanya was an intergenic region (Chr11: 6,142,933–6,162,249) that was only  $-5$  kb away from a MADS-box gene that is the closest rice bean homolog (*Vum\_11G00418*) of Arabidopsis *FRUITFUL* (*FUL*) (Fig. 4a; Supplementary Figs. 8, 9; Supplementary Table 9), a gene known to control flowering time and reproductive transition<sup>61</sup>.

This GWAS signal explained up to 7.04–14.86% of the flowering time variation across two years (Supplementary Data 5). All the significantly associated SNPs and InDels in this GWAS signal were located in its upstream region ( $>5$  kb) (Supplementary Fig. 10), suggesting that these polymorphisms could influence *FUL* expression to control flowering time. This was further supported by the observation that the expression level of *FUL* (in newly expanded leaves in a panel of 16 diverse rice bean landraces) was strongly negatively correlated ( $R = -0.69$ ,  $P = 2.95 \times 10^{-3}$ ) with flowering time (Supplementary Fig. 11).

For the Nanning site, the repeatedly detected major signal (Chr4: 35,931,101–35,996,258) had a PVE (phenotypic variation explained) value of 6.07–8.23% (Supplementary Fig. 8; Supplementary Data 5; Supplementary Table 9). And the most likely candidate among the five protein-coding genes in this region is a *FLOWERING LOCUS T* (*FT*; *Vum\_04G01668*) ortholog (Fig. 4b; Supplementary Data 6; Supplementary Fig. 12); in many species, *FT* genes function as integrators of diverse signals for controlling of flowering time<sup>62</sup>. We found two significantly associated SNPs around ( $<2$  kb) and within *FT* gene; one SNP



**Fig. 4 | The genetic architecture underlying flowering time control from low to high latitudes.** **a–c** Manhattan plots of GWAS for flowering-time data measured in Sanya (18°N) in 2021 (**a**), Nanning (22°N) in 2020 (**b**), and Beijing (40°N) in 2021 (**c**). Red horizontal dashed line indicated the Bonferroni-corrected significance thresholds of GWAS ( $\alpha = 1$ ). Pie charts represented allelic frequencies of the major associated loci. The bar plots display the flowering time of landraces carrying each allele of the identified major loci. DAE, days after emergence. The number (*n*) of landraces carrying each allele is shown below. **d** The geographical distributions of landraces carrying the early-flowering alleles of *FT* and *PRR3*, and the late-flowering allele of *FUL*, respectively. The map was created using the `map_data()` function in the R package `ggplot2`. The violin plot showed significant differences in latitudes between landraces carrying the early-flowering allele of *FT* (*n* = 40) and *PRR3* (*n* = 26), respectively. In the box plots, central line: median values; bounds of the box: 25th and 75th percentiles; whiskers: 1.5\*IQR (IQR: the interquartile range

between the 25th and 75th percentile). **e** The bar plots show the flowering shortening effects of the early-flowering alleles of *FT* and *PRR3* at each of the three measurement sites. NA indicates landraces carrying neither of these two early-flowering alleles. The number (*n*) of landraces carrying each allele at each of the three measurement sites is shown below. **f** The dot plots show the flowering time shortening effects of early-flowering allelic combinations at each of the three measurement sites. Blue dots represent the landraces categorized according to all of the different allelic combinations found in the 440 sequenced landraces. Red lines indicate the average value of each category. NA indicates landraces carrying no early-flowering alleles. The number of landraces for each category is shown below. The significance was tested with two-sided Wilcoxon tests in (**a–e**). The data in **a–c** and **e** are shown as mean  $\pm$  SE, and the error bars represent SE. Source data are provided as a Source Data file.

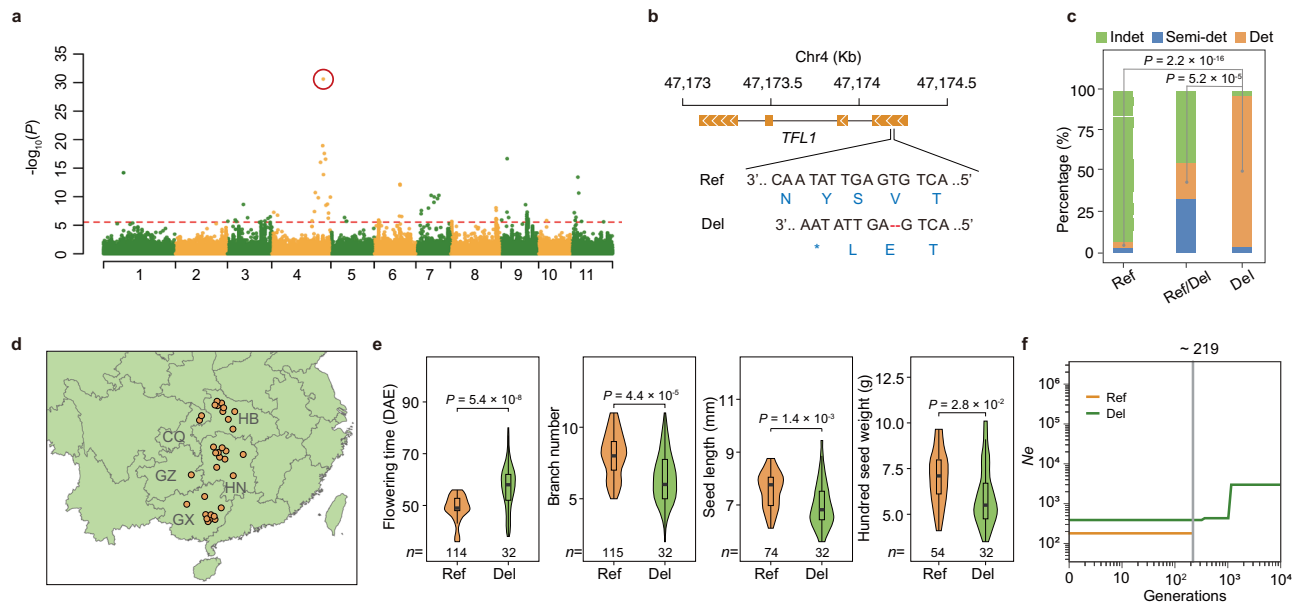
(Chr4:35,950,445) was located upstream (<200 bp) of the transcription start site and another was located in the first intron (Chr4:35,951,311) (Supplementary Fig. 13).

For the Beijing data, we repeatedly detected a peak SNP in the *PSEUDO-RESPONSE REGULATOR 3* (*PRR3*) gene (*Vum\_02G01965*) at Chr2: 38,647,190 (7.30–17.30% of PVE), encoding a nonsynonymous variant (S–F) in the third CDS consisting of the functional PR (pseudo-receiver) domain (Fig. 4c; Supplementary Data 5; Supplementary Figs. 8 and 14; Supplementary Table 9). *PRR3* is an ortholog of the known soybean circadian clock gene *GmTof12/GmPRR3b* that has been previously shown to function as a major flowering time regulatory gene and has been linked to the expansion of soybean into higher latitudes<sup>63,64</sup>. Notably, a similar effect from a single amino acid change (S–L) on flowering time has also been reported for the *GmPRR3b* gene in soybean<sup>63</sup>.

We next explored the potential flowering-time-related impacts of the *FUL*, *FT*, and *PRR3* orthologs in rice beans by classifying the landraces according to their alleles at these three loci. There were two

alleles for *FUL* in the collection, and at the Sanya site, the set of 28 landraces carrying the minor allele (6.73%) displayed significantly ( $P < 0.001$ ) later flowering time (-33 days delayed, a 70.72% increase) than the set of landraces carrying the major allele (Fig. 4a). Note that all of the landraces carrying the late-flowering *FUL* allele were initially collected from low latitude regions (South & Southeast Asia; Fig. 4d). We also found these landraces carrying the late-flowering *FUL* allele also exhibited a significantly higher number of branches than other landraces carrying the early-flowering *FUL* allele (Supplementary Fig. 15), suggesting the probable effect of high yield potential from the late-flowering *FUL* allele.

In contrast, landraces carrying the minor alleles for *FT* (12.31%) and for *PRR3* (7.24%) displayed earlier flowering times, with the average flowering time for the landraces carrying the early-flowering *FT* allele -19 days earlier (a 30.29% reduction) and 18 days earlier (23.36%) for the landraces carrying the *PRR3* minor allele (Fig. 4b, c). There were notable geographical differences among the landraces carrying the early-flowering alleles of *FT* and *PRR3* genes: for *FT* there was a clear



**Fig. 5 | The molecular basis and selection history of stem determinacy in cultivated rice bean.** **a** InDel-based GWAS result from the analysis of data for stem determinacy measured at the Nanning site in 2020. The peak InDel is indicated by the red circle. The red horizontal dashed line indicated the Bonferroni-corrected significance thresholds of GWAS ( $\alpha = 1$ ). **b** A 2-bp causative deletion (the peak InDel) introduced a premature termination codon in the first exon of the *TFL1* gene. Ref reference, Del deletion. **c** The frequency distributions of three types of stem growth habit (indeterminate (Indet), semi-determinate (Semi-det), and determinate (Det)) among three groups comprising landraces carrying the homologous reference alleles (designated as Ref), heterozygous (Ref/Del) or homologous mutation (2-bp deletion) alleles (Del), respectively. Two-sided Fisher's exact tests were used to assess the significance of the differences in the proportion of the determinate type of stem growth habit between landraces carrying Ref and Del alleles and between landraces carrying Ref/Del and Del alleles. **d** The geographical distributions of the

32 landraces carrying Del alleles from Southern-Central China. The map was created using the `map_data()` function in the R package `ggplot2`. HB Hubei province, HN Hunan province, CQ Chongqing province, GZ Guizhou province, GX Guangxi province. **e** There was a significant improvement for the 32 landraces carrying Del alleles compared with the landraces carrying Ref alleles in the SC group for multiple human-desired traits including flowering time (DAE, days after emergence), branch number, seed length, and HGW (hundred seed weight). Significance was tested with two-sided Wilcoxon tests. In the box plots, central line: median values; bounds of the box: 25th and 75th percentiles; whiskers:  $1.5 \times \text{IQR}$  (IQR: the interquartile range between the 25th and 75th percentile). **f** Divergence time of the 32 landraces carrying Del alleles with the landraces carrying Ref alleles in the SC group, inferred using the SMC++ program<sup>66</sup>, under a mutation rate  $\mu = 1.5 \times 10^{-8}$  per site per generation<sup>40</sup>, and a generation time of one year. Source data are provided as a Source Data file.

trend for collection from the region between the Yangtze and Yellow rivers, whereas the landraces harboring the early-flowering *PRR3* allele tended to be from higher latitude regions north of the Yellow River (including Northwest and Northeast China) (Fig. 4d). We also inferred the model of inheritance for these alleles and found that the best models for *FUL*, *FT*, and *PRR3* loci were additive, dominant, and additive, respectively (Supplementary Table 10; see the "Methods" section).

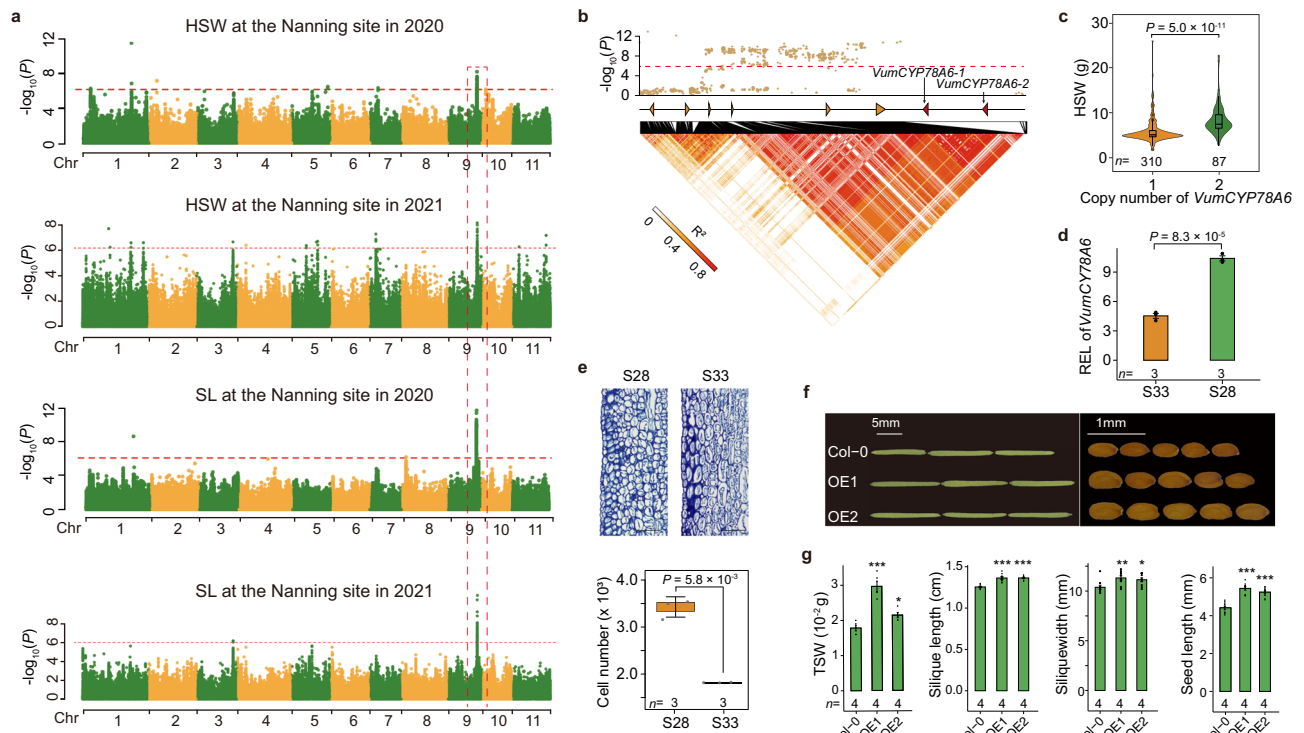
Beyond suggesting that early-flowering alleles for both of these loci have contributed to the adaptation of rice beans to higher latitudes (relative to the tropical origin center), these results indicate potential discrete impacts of the two loci that are sensitive to conditions found in different latitudinal ranges. Offering support for this idea, analysis of phenotype data from the geographically distinct test site revealed differential impacts from the two alleles of interest at the *FT* and *PRR3* loci. That is, at the northernmost site of our study (Beijing), the extent of the flowering time shortening effect was significantly larger among the set of landraces carrying the relevant *PRR3* allele as compared to the set of landraces carrying the relevant *FT* allele (Fig. 4e). Importantly, this trend was reversed at the other two (more southerly) sites: at both Nanning and Sanya, the set of landraces with the early-flowering *FT* allele had the shorter flowering times (Fig. 4e).

We also evaluated the pyramiding effects of the alleles for the *FUL*, *FT*, and *PRR3* loci by comparing the flowering time data in Sanya, Nanning, and Beijing sites among landraces carrying multiple early-flowering allelic combinations. As expected, landraces carrying a relatively higher number of early-flowering alleles invariably exhibited

relatively earlier flowering times (Fig. 4f): a total of three landraces carried all the three early-flowering alleles, and these showed the earliest detected flower times, with the average maximum shortening effects for this set of three landraces being 51.25%, 54.73%, and 25.37% for the Sanya, Nanning, and Beijing sites, respectively (Fig. 4f). It should be noted that this apparently weaker shortening effect at the Beijing site was virtually certainly underestimated, as most of those landraces harboring no early-flowering alleles failed to bloom before the autumn frost. Collectively, these results highlight an opportunity to improve rice bean adaptability for growth in distinct latitudes through breeding efforts to combine the early-flowering alleles for three flowering time-controlling genes.

### The molecular basis and selection history of stem determinacy in cultivated rice bean

The stem determinacy trait is known to strongly influence lodging in legumes<sup>65</sup>. We collected data for stem determinacy traits in 2020 and 2021 for the 440 landraces at the Nanning site. The large majority (>85%) of the landraces exhibited an indeterminate stem growth phenotype (Supplementary Data 3). Notably, this distribution emphasizes that most rice bean landraces do not have the determinate stem growth phenotype that is amenable for mechanized cultivation systems. We performed GWAS analysis of stem determinacy based on the whole genome SNPs data for the germplasm panel and detected a total of 29 and 22 significant signals for stem determinacy in 2020 and 2021, including 7 signals detected repeatedly in both years (Supplementary Fig. 16; Supplementary Data 5; Supplementary Table 9). Among the repeatedly detected signals, the strongest signal was at



**Fig. 6 | Tandem duplication of the *VumCYP78A6* gene associated with seed yield traits. a** GWAS using the 2020 and 2021 Nanning datasets, indicating that the strongest association signals for hundred seed weight (HSW) and seed length (SL) traits all located at Chr9: 29,030,437–29,126,729. **b** Local Manhattan plot of HSW (top), the gene models (middle), and pairwise linkage disequilibrium heat map (bottom) at Chr9: 29,030,437–29,174,247. The two tandemly duplicated *VumCYP78A6* genes (*VumCYP78A6-1* and *VumCYP78A6-2*) are shown with the red dashed triangles. In **a** and **b**, the red horizontal dashed lines indicated the Bonferroni-corrected significance thresholds of GWAS ( $\alpha = 1$ ). **c** The HSW distributions of landraces carrying distinct copy numbers of the *VumCYP78A6* gene. The number ( $n$ ) of landraces carrying distinct copy numbers is shown below. **d** Bar plot showing the relative expression levels of *VumCYP78A6* in the pods at 16 DAP (days after pollination) from the long-seed landrace S28 (carrying two gene copies) and the short-seed landrace S33 (carrying one gene copy). **e** Light microscope images (top)

and cell number per square millimeter (bottom) of the cross-sections of the pod wall for the S28 and S33 landraces at 16 DAP. Scale bar, 100  $\mu\text{m}$ . In the box plots of **c** and **e**, central line: median values; bounds of box: 25th and 75th percentiles; whiskers: 1.5\*IQR (IQR: the interquartile range between the 25th and 75th percentile). **f** Silique (left) and seed (right) morphology of the wild type (Col-0) and two independent *Arabidopsis thaliana* transformants overexpressing the *VumCYP78A6-2* gene (OE1 and OE2). Scale bar: 5 mm for silique and 1 mm for seed. **g** The bar plots of thousand seed weight (TSW), silique length, silique width, and SL for Col-0, OE1, and OE2.  $P$  values are  $1.07 \times 10^{-4}$ ,  $1.11 \times 10^{-2}$ ,  $2.36 \times 10^{-8}$ ,  $4.18 \times 10^{-6}$ ,  $1.44 \times 10^{-2}$ ,  $5.25 \times 10^{-3}$ ,  $1.88 \times 10^{-7}$ , and  $1.94 \times 10^{-6}$ . The significance was tested using the two-sided Student's  $t$ -test in **c**, **d**, **e**, and **g**. \* $P < 0.05$ , \*\* $P < 0.01$ , \*\*\* $P < 0.001$  in (**g**). The data in **d** and **g** are shown as mean  $\pm$  SE. In **d**, **e**, and **g**, the number ( $n$ ) of each independent experiment is shown below. Source data are provided as a Source Data file.

Chr4 with high PVE values (17.96–41.43%) but spanned up to -12 Mb genomic region (Chr4: 42,022,544–53,749,059).

The InDel-based GWAS for the 2-year stem determinacy data revealed a significantly associated InDel (2 bp-deletion at Chr4: 47,174,187) with a PVE of 22.01–35.21% positioned within the strongest SNP signal (Fig. 5a; Supplementary Data 5). Gene functional annotation revealed that this InDel apparently leads to premature termination of translation for the first exon in the gene *Vum\_O4G02513*, *TFL1* (*TERMINAL FLOWER1*; Fig. 5b), for which the ortholog in soybean was reported as the *Dt1* locus (*Gmtfl1* gene) controlling stem determinacy<sup>25</sup>. We found that a total of 32 landraces carried the homozygous mutation (2-nt deletion) alleles, which were identified using the sequencing data and were confirmed using Sanger sequencing (Supplementary Fig. 17). These landraces had a significantly higher proportion of determinate growth habit type compared to landraces harboring the reference alleles or the heterozygous alleles (Fig. 5c).

Notably, these 32 landraces were all in the SC group and were originally collected from an adjoining and mountainous area in South-Central China comprising five provinces (Chongqing, Hunan, Hubei, Guizhou, and Guangxi) (Fig. 5d). We also observed that these 32 landraces (represented by the bars with a predominant proportion of beige color in the Supplementary Fig. 18) were genetically distinct from other landraces within the SC group using model-based

clustering ( $K = 4$ ), an inference that was further supported by a moderate level of differentiation ( $F_{ST} = 0.11$ ). Notably, these landraces also displayed desirable agronomic traits including significantly earlier flowering time and significantly increased pod width, seed length, hundred seed weight, and branch number as compared to the other landraces of the SC group (Fig. 5e).

We next estimated the divergence time for these 32 landraces from the other landraces with distinct genetic admixture in the SC group inferred by the model-based clustering analysis (Fig. 3b), and obtained a similar divergence time of -219 and 249 years ago using the SMC++<sup>66</sup> and MSMC2<sup>67</sup> methods, respectively (Fig. 5f; Supplementary Fig. 19). Our results collectively support that the 32 landraces carrying the homologous mutation alleles have been improved by producers in certain mountainous regions in South-Central China for at least 200 years, and suggest that these materials have huge potential for utilization in modern breeding programs seeking a variety of improvement goals.

### Tandem duplication of the *VumCYP78A6* gene associated with seed yield trait

Seed yield traits (including size and weight) have undergone strong selection in the domestication histories and modern breeding programs for legume crops<sup>53,68–70</sup>. We measured the hundred seed weight



(HSW) and seed length (SL) at the Nanning and Sanya sites in both 2020 and 2021. We next performed GWAS analysis to explore the genetic basis of these two traits, and identified one QTL locus significantly associated with the two traits at both examined sites in both examined years (Fig. 6a; Supplementary Fig. 20; Supplementary Table 9); this QTL is positioned at Chr9: 29,030,437–29,126,729, contains six predicted ORFs (Fig. 6b; Supplementary Data 7), and explains 5.99–16.17% of phenotypic variations (with the maximum value for SL at the Nanning site in 2020; Supplementary Data 5).

We next conducted a qPCR analysis for seed tissues of one long-seed landrace (S28) and one short-seed landrace (S33) at the 16 DAP (days after pollination) for the six candidate genes positioned within the aforementioned significantly associated interval on Chr9. Two of these genes showed significant differences in expression level between the two landraces, but neither of them had an obviously relevant functional annotation (Supplementary Fig. 21), which prompted us to explore the potential candidate genes positioned adjacent to this QTL. We found two tandemly repeated genes (*Vum\_09GO1129* and *Vum\_09GO1130*) at -10.45 kb downstream of the QTL (Fig. 6b). Using an in silico detection approach based on read depth information<sup>71</sup>, a copy number variation (CNV) analysis of this gene in the 440 landraces showed that the 87 landraces carrying two copies exhibited significantly higher values for the two examined phenotypes than the 310 landraces carrying only one copy (Fig. 6c; Supplementary Fig. 22). These results suggested that the CNV may represent the causal variant controlling these two seed yield component traits.

The two duplicated genes had identical CDS sequences and were homologous to the *AtCYP78A6* gene (64.65% amino acid sequence identity; Supplementary Fig. 23), which encodes a cytochrome P450 monooxygenase known to function in maternally promoting seed growth by increasing the cell number in the integument of developing Arabidopsis seeds<sup>72</sup>. We, therefore, designated these rice bean genes as *VumCYP78A6-1* (*Vum\_09GO1129*) and *VumCYP78A6-2* (*Vum\_09GO1130*). A qPCR analysis showed that the expression level of *VumCYP78A6* in pod wall tissue at 16 DAP was significantly higher (-2-fold) in S28 than S33 (Fig. 6d). The impact of this CNV on the expression of *VumCYP78A6* was also verified in a larger panel comprising 20 landraces with one copy and 20 landraces with two copies. Specifically, qPCR analysis of the *VumCYP78A6* gene for the first fully expanded trifoliate leaves at 14 days after sowing showed a significantly ( $P < 0.01$ ) higher expression level (-2-fold) in the 20 landraces with two copies than that in 20 landraces with one copy (Supplementary Fig. 24).

We also examined the number of cells in the pod wall at 16 DAP through cytological observation and detected a significantly increased number of cells in S28 compared to S33 (Fig. 6e). Finally, we generated two independent transgenic Arabidopsis lines by overexpressing the *VumCYP78A6-2* gene (Supplementary Fig. 25), both of which displayed significantly increased values for silique length, silique width, seed length, and seed weight (Fig. 6f, g; Supplementary Fig. 26). Viewed collectively, these results support *VumCYP78A6* as a highly probable causal gene underlying seed yield component traits in rice bean.

## Discussion

Rice bean has been proposed as a potential multipurpose legume crop to promote sustainable agriculture and fight hunger in Asia<sup>18,73</sup>. In the present study, we assembled a high-quality landrace FF25 reference genome and developed a valuable genomics resource by re-sequencing 440 rice bean landraces. By combining the high coverage of PacBio long reads and a Hi-C interaction map, our reference genome reached high accuracy and high continuity; this genome provides a valuable resource for future comparative genomics, evolutionary studies, and molecular research. Our rice bean genome assembly still contains 87 gaps, 78 of which have more than one flanking region (100 bp) with a high proportion (>90%) of repeat

sequences, suggesting that most of the gaps were caused by the incomplete assembly of the repeat sequences, which also reported by other studies<sup>74,75</sup>. We also predicted the candidate centromere regions using a previously published method<sup>76</sup> (see the “Methods” section) and found that all the 11 candidate centromere regions contained more than one assembly gap, suggesting none of the centromere sequences was fully assembled (Supplementary Table 11); future efforts using long sequencing reads (likely the ultra-long ONT reads) should help to ‘close these gaps’. Additionally, our phylogenomic analysis clarified trends in the geographical distribution of the 440 rice bean landraces and revealed a bottleneck as well as an obvious “isolation by distance” pattern<sup>77</sup> for landraces during the northward dispersal of rice beans into and throughout China.

Genomic mutations associated with geographical adaptation allow the radiation of crop species to different agro-ecological and cultural environments<sup>78</sup>. Genetic control of flowering time is of great significance in determining the adaptation during the domestication and diversification of many crop species<sup>79,80</sup>. Appropriate timing to flowering is undoubtedly an advantage for survival and/or propagation<sup>81</sup> at distinct latitudes, as this impacts the growth period structure, yield, and quality of crops<sup>82–84</sup>. Studying flowering time is a large research field in plant biology because of its obvious agronomic implications, and studies from multiple species have shown that flowering time is controlled by multigene, highly topologically complex regulatory networks<sup>85–87</sup>. Our study has revealed how genetic alterations of the three known flowering loci—*FUL*, *FT*, and *PRR3*—have apparently supported rice bean’s adaptation during its dispersal across a latitudinal gradient from South to North.

Agronomically, experience with crops including soybean and rice has established that flowering is delayed when a short-day crop species are grown at a high latitude location<sup>84,88</sup>, so it is necessary to reduce the photoperiod sensitivity of such plants to advance flowering time and thus enable productive growth and yield<sup>86</sup>. We found that an early-flowering allele of the *PRR3* gene apparently supports early flowering for landraces from North of the Yellow river. It is notable that studies of barley<sup>89</sup>, soybean<sup>63,64</sup>, and rice<sup>90</sup> have also implicated *PRR* gene family members in high-latitude adaption. Previous studies in rice<sup>91</sup>, cucumber<sup>92</sup>, and soybean<sup>93</sup> have implicated natural variation in the *FT* gene in enhancing adaptation to higher latitudes. We identified an early-flowering allele of the *FT* gene that has apparently contributed to the adaptation of rice beans in the relatively low latitude region between the Yangtze and Yellow rivers. In contrast to these two alleles supporting adaptation to higher latitudes, a late-flowering allele of the *FUL* gene was found to have the potential to increase grain yield by extending the vegetative growth period and generating more branches at low latitude growth sites. Pleiotropy of the *FUL* gene has also been reported for other species including Arabidopsis<sup>94</sup>, tomato<sup>95</sup>, and *Setaria viridis*<sup>96</sup>. Similar to the *FUL* gene, in the short-day model plant soybean, breeding exploitation of the *J* gene (*ELF3*) has enabled the successful deployment of commercial soybean cultivation in tropical regions<sup>97</sup>. It is conceivable that—perhaps similar to successful efforts to variously combine mutations in four *E* loci in soybean<sup>98</sup>—our insights about the differential geographical distributions of alleles for flowering loci could be exploited to develop high-yield rice varieties for growth at low to high latitudes.

Our GWAS analyses helped decipher the genetic basis of stem determinacy in rice beans, detecting that stem determinacy of rice beans is influenced by the *TFL1* gene; this gene has been implicated in determining node termination and node number to control plant height and stem determinacy in many legumes species<sup>25,99–101</sup>. We also found that 32 landraces from Southern-Central China have multiple agronomically desirable traits and have undergone improvement by humans for at least 200 years; these materials should be considered for use as elite parents in rice bean breeding programs. Historically, elite landraces of other crops have been hugely beneficial to modern

breeding<sup>26,102,103</sup>, for example with Taiwanese landraces in rice<sup>104</sup>: the so-called “miracle rice” IR8 with high yield supported the Green Revolution in Asia, and this line harbored a semi-dwarf allele from the Taiwanese landrace Dee-Geo-Woo-Gen<sup>105</sup>. Although QTLs for stem determinacy and seed yield-related traits were detected by our GWAS analyses in one and two environments respectively, further efforts should be made to investigate the robustness of these QTLs in more different environments.

Although the rice bean has been cultivated for thousands of years, to date it has received very little attention from breeders and agricultural scientists. The wealth of resources developed and identified in our study should help to rapidly advance breeding programs seeking to produce excellent varieties that simultaneously display geographically suitable flowering times, stem determinacy to support mechanized cultivation, and high yields through marker-assisted selection.

## Methods

### Plant materials and sequencing

The sequenced rice bean (*Vigna umbellata*) landraces used in this study were obtained from the Center for Crop Germplasm Resources, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. An individual plant of rice bean landrace FF25 growing in a field in Beijing was used for the reference genome construction. The tender leaves were sampled for DNA extraction, and tissues including root, tender leaves, tender stem, flower, pod, and seed were harvested and immediately frozen in liquid nitrogen. Samples were stored at  $-80^{\circ}\text{C}$  prior to DNA or RNA extraction.

The high-quality genomic DNA from tender leaves was extracted and purified using DNeasy Plant Maxi Kits (Qiagen, Germany). The DNA concentration was measured using a NanoDrop spectrophotometer (Thermo Fisher Scientific, USA) and a Qubit 2.0 Fluorometer (Invitrogen, USA). Illumina short-read data were obtained using the Illumina NovaSeq platform, which generated a total of 338.19 million paired-end reads, with a total length of 50.73 Gb (Supplementary Table 1). Single-Molecule Real-Time (SMRT) cells were sequenced on the PacBio Sequel platform (Pacific Biosciences, USA), generating a total of 10.99 million reads with a total length of 142.95 Gb. Hi-C libraries were constructed from tender leaves using the Illumina NovaSeq platform. This allowed us to generate a total of 465.99 million paired-end reads and 69.90 Gb of sequencing data.

Each of the 440 landraces was planted at different sites for 2 years: (1) Beijing site ( $40.23^{\circ}\text{N}$ ,  $116.56^{\circ}\text{E}$ ) with sowing date in the middle of June 2020 and 2021; (2) Nanning site ( $23.15^{\circ}\text{N}$ ,  $108.28^{\circ}\text{E}$ ) with sowing date in the middle of July 2020 and 2021; (3) Sanya site ( $18.38^{\circ}\text{N}$ ,  $109.21^{\circ}\text{E}$ ) with sowing date in the middle of November 2019 and 2020. Supplementary Fig. 27 presents the day length (per day) during the -5-month growth period for the three sites. The day length differs obviously among the three sites (but was very similar between the two observation years). The average day lengths of Beijing during the first 4 months (during which all the landraces opened the first flower) was the longest (13.94 and 13.93 h) in both 2020 and 2021, followed by the Nanning site (12.52 and 12.53 h) and the Sanya site (11.28 and 11.28 h). Note that the phenotypes of the landraces grown at the Sanya site were measured the next year (i.e., 2020 and 2021; when the landraces were harvested); thus, the time of phenotypic data was designated as 2020 and 2021. For the plantings, 20 seeds of each landrace were sown in two rows (10 plants per row). Phenotypes in all three environments were investigated following the “Descriptors and data standards [*Vigna umbellata* (Thunb.) Ohwi & Ohashi]”<sup>106</sup>. Briefly, flowering time was recorded as the number of days after emergence (DAE) when the first flower opened. The main stem type was classified as indeterminate, semi-determinate, or determinate according to the growth state of the plants by observation<sup>106</sup> on five healthy individuals randomly selected from each plot for each landrace. The seed morphological

traits (seed length and hundred seed weight) for each landrace at each site were measured after harvest using automatic seed counting and analyzing instrument (Model SC-G, Hangzhou Wanshen Detection Technology Co., Ltd., Hangzhou, China, <http://www.wseen.com/>)<sup>107</sup>. The pod morphological traits (pod length and pod width) were measured using a vernier caliper with at least five healthy individuals for each landrace at each site after harvest.

### Genome assembly and quality assessment

In order to estimate the genome size of rice beans, the Illumina short reads were recruited to determine the *K*-mer distributions using GCE v1.0.2 (<https://github.com/fanagislab/GCE>). The PacBio long-read data were de novo assembled into PacBio contigs using Canu v1.9<sup>34</sup>, and then the contigs were extended without the introduction of any gaps using the highly efficient repeat assembly (HERA) method<sup>35</sup>, generating a total of 351 contigs with an N50 value of 18.26 Mb (Table 1). The Illumina short-read data was used for error-correcting of the contigs using Pilon<sup>108</sup>. Subsequently, to anchor the contigs into chromosomes, we aligned the Hi-C sequencing data into these contigs using Juicer v1.8.9<sup>109</sup>. The contigs were finally linked into 11 distinct chromosomes by 3D-DNA v180922<sup>110</sup>.

The Illumina short-read data were also used to evaluate assembly accuracy and completeness using BWA-MEM v0.7.17-r118896<sup>111</sup>. The completeness of the genome assembly and the gene annotations were assessed with a plant database composed of 2121 conserved plant genes (eudicotyledons\_odb10) using BUSCO v3.0.297<sup>41</sup>.

### Repeats and gene annotation

The annotation of transposable elements was performed using RepeatMasker (<http://www.repeatmasker.org>). The repeat libraries included the RepBase-20170127 and a de novo repeat library created using RepeatModeler (<http://www.repeatmasker.org>) (with the parameter -LTRStruct). We analyzed the density distribution of the top-50 most abundant repeat subfamilies in 100 kb windows (using RepeatMasker), and used BEDtools<sup>112</sup> to merge the results with the parameter ‘-d 100000’. The rnd-6\_family-604 subfamily (a 217-bp repeat) was identified as a centromere-specific repeat (Supplementary Fig. 28). The candidate centromere regions were also predicted according to the density distribution of this centromere-specific repeat (Supplementary Table 11). The LTRharvest<sup>113</sup> and the LTR\_FINDER<sup>114</sup> programs were used to identify intact LTRs in the genomes of five *Vigna* species (*V. stipulacea*, *V. radiata*, *V. angularis*, *V. umbellata*, and *V. unguiculata*). LTR insertion times were estimated according to the formula  $T = d/2m$  ( $d$ , the nucleotide distance for each pair of LTRs;  $m$ , the nucleotide substitution rate =  $1.64 \times 10^{-8}$ ).

Protein-coding genes were predicted using three different strategies: ab initio prediction, homology-based prediction, and transcript-based prediction. We used augustus<sup>115</sup> and SNAP<sup>116</sup> for ab initio predictions, and exonerate<sup>117</sup> was used for homology-based predictions. For transcript-based predictions, the RNA-Seq clean reads of tissues including root, tender leaves, tender stem, flower, pod, and seed were mapped to the genome assembly using HISAT2<sup>118</sup>. The mapping reads were assembled into transcripts using StringTie<sup>119</sup>. The transcripts were used for gene structure prediction using TransDecoder (<http://transdecoder.github.io>) and GeneMarkS-T<sup>120</sup>. These clean reads were also de novo assembled using Trinity<sup>121</sup> and the assembled transcripts were subsequently used for gene prediction using PASA<sup>122</sup>. Finally, EvidenceModeler (EVM) v1.1.1<sup>123</sup> was used to integrate the prediction results obtained by the above three methods (codon length  $\geq 150$  bp) to produce high-confidence gene models.

Ribosomal RNAs (rRNAs) were identified using RNAMmer<sup>124</sup> with default parameters. Reliable tRNA structures were detected using tRNAscan-SEM v1.23<sup>125</sup>. Non-coding RNAs containing miRNA and snoRNA features were annotated using INFERNAL<sup>126</sup> with default parameters. Pseudogenes were identified using the published

pipeline<sup>127</sup>. The transcription factors and transcription regulators were annotated using iTAK v18.12<sup>128</sup> with default parameters.

### Gene families and phylogenetic analysis

We used OrthoFinder v2.3.9<sup>129</sup> to identify shared gene families between rice beans and 13 other plant species, including five *Vigna* species (*V. stipulacea*, *V. radiata*, *V. angularis*, *V. umbellata*, and *V. unguiculata*), four other legumes (*Phaseolus vulgaris*, *Glycine max*, *Lotus japonicus*, and *Arachis duranensis*), five other eudicots (*Arabidopsis thaliana*, *Citrus sinensis*, *Populus trichocarpa*, *Vitis vinifera*, and *Solanum lycopersicum*), and one monocot (*Oryza sativa*). Based on the protein sequences of 334 single-copy ortholog families, the phylogenetic relationships among these species were estimated using RAxML v8.2.12<sup>130</sup>. Divergence times were estimated by the MCMCTree program embedded in PAML v4.9<sup>131</sup>. We measured the expansion and contraction of orthologous gene families based on a maximum likelihood tree using CAFE v4.2 (<https://github.com/hahnlab/CAFE>).

### KEGG enrichment analysis

The R package ClusterProfiler v3.18.0<sup>132</sup> was used to perform KEGG enrichment analysis. KEGG terms showing adjusted *P* values < 0.05 were considered significantly enriched.

### Comparative genomics and Ks analysis

Gene synteny analysis was performed using MCScanX<sup>43</sup> and BLASTP<sup>133</sup> (-evalue < 1e-10, -v 5, -b 5) to determine the pairwise similarity among the protein sequences of *Glycine max*, *Phaseolus vulgaris*, and five *Vigna* species (*V. stipulacea*, *V. radiata*, *V. angularis*, *V. umbellata*, and *V. unguiculata*). The synteny figure was plotted using the NGenomeSyn program (<https://github.com/hewm2008/NGenomeSyn>). Synonymous nucleotide substitutions on synonymous sites (Ks) were estimated using the WGDtool (<https://github.com/SunPengChuan/wgdi>) with default parameters.

### SNP and small InDel calling

We sequenced the genomes for 440 rice bean landraces with an average depth of 24.91× using the Illumina NovaSeq platform (Supplementary Data 3). The quality control for the raw sequencing data was performed using fastp v0.20.1<sup>134</sup> with default settings. The high-quality short reads were aligned to the genome using BWA-MEM v0.7.17-r118896<sup>111</sup>; PCR duplicates were removed using Picard v1.118 (<http://broadinstitute.github.io/picard/>); SNPs and InDels were identified using HaplotypeCaller of the Genome Analysis Toolkit (GATK) v4.1.5.0<sup>135</sup>, and were subsequently filtered ('QD < 2.0 || FS > 60.0 || MQ < 40.0 || MQRankSum < -12.5 || ReadPosRankSum < -8.0' for SNPs, and 'QD < 2.0 || FS > 200.0 || ReadPosRankSum < -20.0' for InDels)<sup>49</sup>. Non-biallelic SNPs/InDels with a read depth < 5 were removed from further analyses.

### Phylogenetic and population structure analyses

A total of 1,400,862 SNPs with a minor allele frequency (MAF) ≥ 0.05 and missing rate ≤ 50% were used to build a maximum likelihood phylogenetic tree using TreeBeST v1.9.2<sup>136</sup>, as well as to perform principal component analyses (PCA) using the smartPCA program embedded in the Eigensoft package v7.2.1<sup>137</sup>. The  $\pi$  and  $F_{ST}$  values were calculated using VCFtools v0.1.17<sup>138</sup> based on the same SNP set. Population structure was investigated based on 20,000 randomly selected SNPs using STRUCTURE v2.3.4<sup>50</sup> with 100,000 iterations of burning and 200,000 iterations of MCMC, and evaluating each *K* from 2 to 4.

### Divergence time estimation

MSMC2 v2.1.1<sup>67</sup> was used to infer the divergence times of stem determinacy landraces carrying homologous deletion mutation alleles with other landraces carrying the homologous reference alleles in the SC

group. To improve reliability, genome regions were masked with the SNPable tool (<http://lh3lh3.users.sourceforge.net/snpable.shtml>) when the coverage depth was < 15× after removing reads with mapping quality < 20. First, we split the reference genome into overlapping 35-mers and then mapped these 35-mers back to the reference genome using BWA<sup>139</sup> (bwa aln -R 1000000 -O 3 -E 3). Only regions where the majority of 35-mers were uniquely mapped and without mismatch were retained for further analysis. We selected the top 10 samples in each population with the highest coverage after masking. The 8 most frequent haplotypes were randomly selected from the 10 samples in order to infer the demographic history of each population. We repeated this procedure 20 times. Scaled times were converted to years by assuming a generation time of 1 year and a mutation rate of  $\mu = 1.5 \times 10^{-8}$  per site per generation<sup>140</sup>. We also used the SMC++ v1.15.2<sup>66</sup>, which does not rely on haplotype phase information, to estimate the divergence times (using the same generation time and mutation rate).

### Linkage disequilibrium

To estimate and compare the patterns of linkage disequilibrium (LD) decay in each population, we computed the mean squared correlation coefficient ( $r^2$ ) values between any two SNPs within 300 kb using PopLDdecay v3.41<sup>141</sup>.

### GWAS analysis

We retained SNPs with a MAF ≥ 0.05 and a missing rate ≤ 50% to perform GWAS analysis. After imputation using Beagle v4.1<sup>142</sup> with default parameters, the GWAS analysis was performed based on a linear mixed model using the program Fast-LMM v2.06.20130802<sup>143</sup>. The *P* value threshold for significance was estimated as  $1/n$  (where *n* corresponds to the SNP number). The phenotypic variance that was explained by each SNP was estimated following the below previously reported method<sup>144</sup>:

$$PVE = \frac{2\hat{\beta}^2 \times MAF \times (1 - MAF)}{2\hat{\beta}^2 \times MAF \times (1 - MAF) + (se(\hat{\beta}))^2 \times 2N \times MAF \times (1 - MAF)} \quad (1)$$

where  $\hat{\beta}$  and MAF is the effect size estimate and minor allele frequency for the SNP, *N* is the sample size, and  $se(\hat{\beta})$  is the standard error of effect size for the SNP.

### Inference of the inheritance model for alleles of the flowering time genes

To infer the most likely inheritance model of alleles for the flowering-time related loci (*FUL*, *FT*, and *PRR3*), we used the R package "SNPassoc"<sup>145</sup> to perform association analysis of the alleles based on several genetic models (co-dominant, dominant, recessive, over-dominant, or additive). The model with the smallest Akaike information criteria (AIC) value was identified as the best fitting genetic model.

### Histological analysis

Cross-sections of pod walls from S28 and S33 landraces were analyzed by light microscopy (BX51; Olympus). The pod wall tissues were sampled at 16 DAP and immediately fixed with FAA: glacial acetic, 38% form aldehyde, 70% ethanol (1:1:18), and then dehydrated through a standard ethanol series. The pod wall tissues were embedded in Paraplast Plus tissue-embedding medium (Sigma-Aldrich), sectioned at 8 mm using a microtome (RM2235, Leica Microsystems), and then stained with toluidine blue. The cell numbers in the cross-sections were measured using Olympus Stream software. The analysis was based on at least three biological replicates.

### RNA extraction and qPCR analysis

qPCR analysis was used to quantify the relative expression levels of the *FUL* gene in newly expanded leaves in a panel of 16 diverse landraces, the expression levels of the *VumCYP78A6* gene in the seed and pod tissues of the S28 and S33 landraces, and the expression levels of the *VumCYP78A6-2* gene in the primary inflorescence stems of the transgenic Arabidopsis plants. Total RNA was extracted using Trelief™ RNAPrep Pure Plant Kits (Polysaccharides & Polyphenolics-rich) (Tsingke, China). First-strand cDNA was synthesized using a PrimeScript™ RT Reagent Kit with gDNA Eraser (Takara, Japan). Quantitative PCR was performed using TSINGKE Master qPCR Mix (SYBR Greenwith UDG) (Tsingke, China), on a StepOnePlus™ Real-Time PCR System (Applied Biosystems, USA) following the manufacturer's instructions. cDNA transcript levels were normalized to those of the reference gene *ACTIN* using the 2<sup>-ΔΔCT</sup> method<sup>46</sup>. PCR reactions were performed in triplicate for each biological replicate; three or more biological replicates were assessed. Primers were designed to span an intron in order to avoid the amplification of genomic DNA and are shown in Supplementary Table 12.

### Arabidopsis transformation

The total RNA of the pod tissue from the FF25 landrace was extracted and reverse transcription was performed. The full coding sequence of the *VumCYP78A6-2* gene (*Vum\_09GO1130*) was amplified and cloned into the pEasy-T1 vector. The binary vector pCambia3301 was used to subclone the gene for overexpression. The construct was individually introduced into *Agrobacterium tumefaciens* strain GV3101 and transformed into the Arabidopsis ecotype Columbia (Col-0) using the floral dip method<sup>47</sup>. Relative expression levels of the *VumCYP78A6-2* gene in primary inflorescence stems of 2-week-old T1 transgenic plants were measured with qPCR, and two lines with relatively high *VumCYP78A6-2* expression were selected for further analyses. All phenotypes were measured for T3 homozygote plants. Primers are shown in Supplementary Table 12.

### Reporting summary

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

### Data availability

Data supporting the findings of this work are available within the paper and its Supplementary Information files. A reporting summary for this Article is available as a Supplementary Information file. The datasets and plant materials generated and analyzed during the current study are available from the corresponding author upon request. All datasets reported in this study have been deposited in the National Center for Biotechnology Information (NCBI) with the following accession IDs: FF25 genome assembly, [JALEER000000000](https://www.ncbi.nlm.nih.gov/assembly/JALEER000000000/); Raw data for FF25 genome assembly, [PRJNA819955](https://www.ncbi.nlm.nih.gov/assembly/PRJNA819955/); Raw data for genome sequencing of 440 landraces, [PRJNA803965](https://www.ncbi.nlm.nih.gov/assembly/PRJNA803965/). The annotation files including predicted CDS and protein sequences generated for FF25 genome assembly have been deposited at Figshare [<https://doi.org/10.6084/m9.figshare.19420058>]. The online tools and database used in this paper include: Pfam [<http://pfam.xfam.org/>], InterPro [<https://www.ebi.ac.uk/interpro/>], NR [<https://www.ncbi.nlm.nih.gov/refseq/about/nonredundantproteins/>], GO [<http://geneontology.org/>], KEGG [<https://www.genome.jp/kegg/>], FLOWer Interactive Database [<http://www.phytosystems.ulg.ac.be/florid/>]. Source data are provided with this paper.

### Code availability

The scripts used for the analyses can be freely and openly accessed at GitHub [<https://github.com/guanjiantao-caas/Code-for-rice-bean>].

### References

1. Takahashi, Y. et al. Novel genetic resources in the genus *Vigna* unveiled from gene bank accessions. *PLoS ONE* **11**, e0147568 (2016).
2. Pattanayak, A. et al. Rice bean: a lesser known pulse with well-recognized potential. *Planta* **250**, 873–890 (2019).
3. Pattanayak, A. et al. Diversity analysis of rice bean (*Vigna umbellata* (Thunb.) Ohwi and Ohashi) collections from North Eastern India using morpho-agronomic traits. *Sci. Hortic.* **242**, 170–180 (2018).
4. Tomooka, N., Vaughan, D. A., Moss, H. & Maxted, N. *The Asian Vigna: Genus Vigna Subgenus Ceratotropis Genetic Resources* (Kluwer Academic Publishers, Dordrecht, The Netherlands, 2003).
5. Bhanu, A. N., Singh, M. N. & Srivastava, K. Efficient hybridization procedure for better pod setting in inter-specific crosses involving *Vigna* species. *Adv. Plants Agric. Res.* **8**, 101–105 (2018).
6. Singh, I., Sandhu, J. S., Gupta, S. K. & Singh, S. Introgression of productivity and other desirable traits from rice bean (*Vigna umbellata*) into black gram (*Vigna mungo*). *Plant Breed.* **132**, 401–406 (2013).
7. Dana, S. & Karmakar, P. G. *Plant Breeding Reviews* Ch. 2 (Timber Press, Inc., Portland, OR, 1990).
8. Smartt, J. *Grain Legumes: Evolution and Genetic Resources* (Cambridge University Press, Cambridge, 1990).
9. Kaur, H., Gill, R. S. & Kaur, R. Correlation between biophysical seed characteristics of rice bean, *Vigna umbellata* (Fabaceae: Faboideae: Phaseoleae) and the development of *Callosobruchus maculatus* (Coleoptera: Chrysomelidae: Bruchinae). *J. Stored Prod. Res.* **83**, 9–13 (2019).
10. Cheema, H. K., Gill, R. K. & Singh, P. Screening of rice bean genotypes against major insect pests and avoidable yield losses. *Agric. Res. J.* **56**, 675–682 (2019).
11. Kashiwaba, K., Tomooka, N., Kaga, A., Han, O. K. & Vaughan, D. A. Characterization of resistance to three bruchid species (*Callosobruchus* spp., Coleoptera, Bruchidae) in cultivated rice bean (*Vigna umbellata*). *J. Econ. Entomol.* **96**, 207–213 (2003).
12. Somta, P. et al. Development of an interspecific *Vigna* linkage map between *Vigna umbellata* (Thunb.) Ohwi & Ohashi and *V. nakashimae* (Ohwi) Ohwi & Ohashi and its use in analysis of bruchid resistance and comparative genomics. *Plant Breed.* **125**, 77–84 (2006).
13. Venkataramana, P. B. et al. Mapping QTL for bruchid resistance in rice bean (*Vigna umbellata*). *Euphytica* **207**, 135–147 (2016).
14. Tomooka, N., Kashiwaba, K., Vaughan, D. A., Ishimoto, M. & Egawa, Y. The effectiveness of evaluating wild species, searching for sources of resistance to bruchid beetle in the genus *Vigna* subspecies. *Caratotropis*. *Euphytica* **115**, 27–41 (2000).
15. Arora, R. K., Chandel, K. P. S., Joshi, B. S. & Pant, K. C. Rice bean: tribal pulse of eastern India. *Econ. Bot.* **34**, 260–263 (1980).
16. Atta, K., Chettri, P. & Pal, A. K. Physiological and biochemical changes under salinity and drought stress in rice bean [*Vigna umbellata* (Thunb.) Ohwi and Ohashi] seedlings. *Int. J. Environ. Clim. Change* **10**, 58–64 (2020).
17. Wanek, W. & Richter, A. Biosynthesis and accumulation of D-ononitol in *Vigna umbellata* in response to drought stress. *Physiol. Plant.* **101**, 416–424 (1997).
18. Asha, R. K., Koundinya, A., Das, A. & Chattopadhyay, S. B. A review on an underutilised multipurpose legume: rice bean. *Acta Hortic.* **1241**, 57–64 (2019).
19. Nandeshwar, B. C. & De, D. K. Screening of rice bean (*Vigna umbellata* (Thunb.) Ohwi and Ohashi) accessions at early seedling stage for NaCl tolerance under controlled condition. *Curr. J. Appl. Sci. Technol.* **40**, 71–79 (2021).
20. Dhillon, P. K. & Tanwar, B. Rice bean: a healthy and cost-effective alternative for crop and food diversity. *Food Secur.* **10**, 525–535 (2018).

21. Seehalak, W. et al. Genetic diversity of the *Vigna* germplasm from Thailand and neighboring regions revealed by AFLP analysis. *Genet. Resour. Crop Evol.* **53**, 1043–1059 (2006).
22. Tian, J. et al. Genetic diversity of the rice bean (*Vigna umbellata*) gene pool as assessed by SSR markers. *Genome* **56**, 717–727 (2013).
23. Gupta, S. et al. Genetic parameters of selection and stability and identification of divergent parents for hybridization in rice bean (*Vigna umbellata* Thunb. (Ohwi and Ohashi)) in India. *J. Agric. Sci.* **147**, 581–588 (2009).
24. Craufurd, P. Q. & Wheeler, T. R. Climate change and the flowering time of annual crops. *J. Exp. Bot.* **60**, 2529–2539 (2009).
25. Tian, Z. et al. Artificial selection for determinate growth habit in soybean. *Proc. Natl Acad. Sci. USA* **107**, 8563–8568 (2010).
26. Dwivedi, S. L. et al. Landrace germplasm for improving yield and abiotic stress adaptation. *Trends Plant Sci.* **21**, 31–42 (2016).
27. Wang, L. et al. Analysis of simple sequence repeats in rice bean (*Vigna umbellata*) using an SSR-enriched library. *Crop J.* **4**, 40–47 (2016).
28. Wang, L., Cheng, X. & Wang, S. Genetic diversity analysis and a core collection construction of rice bean (*Vigna umbellata*) in China. *J. Plant Genet. Resour.* **15**, 242–247 (2014).
29. Isemura, T., Kaga, A., Tomooka, N., Shimizu, T. & Vaughan, D. A. The genetics of domestication of rice bean, *Vigna umbellata*. *Ann. Bot.* **106**, 927–944 (2010).
30. Alqudah, A. M., Sallam, A., Stephen Baenziger, P. & Borner, A. GWAS: fast-forwarding gene identification and characterization in temperate cereals: lessons from Barley—a review. *J. Adv. Res.* **22**, 119–135 (2020).
31. Yu, Y. et al. Population-scale peach genome analyses unravel selection patterns and biochemical basis underlying fruit flavor. *Nat. Commun.* **12**, 3604 (2021).
32. Xiao, Y., Liu, H., Wu, L., Warburton, M. & Yan, J. Genome-wide association studies in Maize: praise and stargaze. *Mol. Plant* **10**, 359–374 (2017).
33. Guan, J. et al. Genome structure variation analyses of peach reveal population dynamics and a 1.67 Mb causal inversion for fruit shape. *Genome Biol.* **22**, 13 (2021).
34. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).
35. Du, H. & Liang, C. Assembly of chromosome-scale contigs by efficiently resolving repetitive sequences with long reads. *Nat. Commun.* **10**, 5360 (2019).
36. Yang, K. et al. Genome sequencing of adzuki bean (*Vigna angularis*) provides insight into high starch and low fat accumulation and domestication. *Proc. Natl Acad. Sci. USA* **112**, 13213–13218 (2015).
37. Kang, Y. J. et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. *Nat. Commun.* **5**, 5443 (2014).
38. Schmutz, J. et al. A reference genome for common bean and genome-wide analysis of dual domestications. *Nat. Genet.* **46**, 707–713 (2014).
39. Lonardi, S. et al. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). *Plant J.* **98**, 767–782 (2019).
40. Pootakham, W. et al. A chromosome-scale assembly of the black gram (*Vigna mungo*) genome. *Mol. Ecol. Resour.* **21**, 238–250 (2021).
41. Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
42. Ou, S., Chen, J. & Jiang, N. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**, e126 (2018).
43. Wang, Y. et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res.* **40**, e49 (2012).
44. Van de Peer, Y., Maere, S. & Meyer, A. The evolutionary significance of ancient genome duplications. *Nat. Rev. Genet.* **10**, 725–732 (2009).
45. Lynch, M. & Conery, J. S. The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155 (2000).
46. Wang, J. et al. Hierarchically aligning 10 legume genomes establishes a family-level genomics platform. *Plant Physiol.* **174**, 284–300 (2017).
47. Liu, Y. et al. Pan-genome of wild and cultivated soybeans. *Cell* **182**, 162–176.e13 (2020).
48. Qi, J. et al. A genomic variation map provides insights into the genetic basis of cucumber domestication and diversity. *Nat. Genet.* **45**, 1510–1515 (2013).
49. Guo, S. et al. Resequencing of 414 cultivated and wild watermelon accessions identifies selection for fruit quality traits. *Nat. Genet.* **51**, 1616–1623 (2019).
50. Hubisz, M. J., Falush, D., Stephens, M. & Pritchard, J. K. Inferring weak population structure with the assistance of sample group information. *Mol. Ecol. Resour.* **9**, 1322–1332 (2009).
51. Hill, W. G. & Robertson, A. Linkage disequilibrium in finite populations. *Theor. Appl. Genet.* **38**, 226–231 (1968).
52. Hufford, M. B. et al. Comparative population genomics of maize domestication and improvement. *Nat. Genet.* **44**, 808–811 (2012).
53. Zhou, Z. et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. *Nat. Biotechnol.* **33**, 408–414 (2015).
54. Huang, X. et al. Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat. Genet.* **42**, 961–967 (2010).
55. Jia, G. et al. A haplotype map of genomic variations and genome-wide association studies of agronomic traits in foxtail millet (*Setaria italica*). *Nat. Genet.* **45**, 957–961 (2013).
56. Remington, D. L. et al. Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc. Natl Acad. Sci. USA* **98**, 11479–11484 (2001).
57. Bouché, F., Lobet, G., Tocquin, P. & Périlleux, C. FLOR-ID: an interactive database of flowering-time gene networks in *Arabidopsis thaliana*. *Nucleic Acids Res.* **44**, D1167–D1171 (2016).
58. Somers, D. E., Webb, A. A., Pearson, M. & Kay, S. A. The short-period mutant, *toc1-1*, alters circadian clock regulation of multiple outputs throughout development in *Arabidopsis thaliana*. *Development* **125**, 485–494 (1998).
59. Murakami-Kojima, M., Nakamichi, N., Yamashino, T. & Mizuno, T. The APRR3 component of the clock-associated APRR1/TOC1 quintet is phosphorylated by a novel protein kinase belonging to the WNK family, the gene for which is also transcribed rhythmically in *Arabidopsis thaliana*. *Plant Cell Physiol.* **43**, 675–683 (2002).
60. Schaffer, R. et al. The *late elongated hypocotyl* mutation of *Arabidopsis* disrupts circadian rhythms and the photoperiodic control of flowering. *Cell* **93**, 1219–1229 (1998).
61. Balanza, V., Martínez-Fernández, I. & Ferrandiz, C. Sequential action of *FRUITFULL* as a modulator of the activity of the floral regulators *SVP* and *SOC1*. *J. Exp. Bot.* **65**, 1193–1203 (2014).
62. Corbesier, L. et al. FT protein movement contributes to long-distance signaling in floral induction of *Arabidopsis*. *Science* **316**, 1030–1033 (2007).
63. Li, C. et al. A domestication-associated gene *GmPRR3b* regulates the circadian clock and flowering time in soybean. *Mol. Plant* **13**, 745–759 (2020).
64. Lu, S. et al. Stepwise selection on homeologous *PRR* genes controlling flowering and maturity during soybean domestication. *Nat. Genet.* **52**, 428–436 (2020).

65. Yue, L. et al. FT5a interferes with the Dt1-AP1 feedback loop to control flowering time and shoot determinacy in soybean. *J. Integr. Plant Biol.* **63**, 1004–1020 (2021).
66. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
67. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
68. Wang, E. et al. Control of rice grain-filling and yield by a gene with a potential signature of domestication. *Nat. Genet.* **40**, 1370–1374 (2008).
69. Abbo, S. et al. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. *Trends Plant Sci.* **19**, 351–360 (2014).
70. Kaga, A., Isemura, T., Tomooka, N. & Vaughan, D. A. The genetics of domestication of the azuki bean (*Vigna angularis*). *Genetics* **178**, 1013–1036 (2008).
71. Klambauer, G. et al. Cn.MOPS: mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Res.* **40**, e69 (2012).
72. Fang, W., Wang, Z., Cui, R., Li, J. & Li, Y. Maternal control of seed size by EOD3/CYP78A6 in *Arabidopsis thaliana*. *Plant J.* **70**, 929–939 (2012).
73. Siddique, K. H. M., Li, X. & Gruber, K. Rediscovering Asia's forgotten crops to fight chronic and hidden hunger. *Nat. Plants* **7**, 116–122 (2021).
74. Nurk, S. et al. The complete sequence of a human genome. *Science* **376**, 44–53 (2022).
75. Peona, V. et al. Identifying the causes and consequences of assembly gaps using a multiplatform genome assembly of a bird-of-paradise. *Mol. Ecol. Resour.* **21**, 263–286 (2021).
76. Su, X. et al. A high-continuity and annotated tomato reference genome. *BMC Genom.* **22**, 898 (2021).
77. Wright, S. Size of population and breeding structure in relation to evolution. *Science* **87**, 430–431 (1938).
78. Meyer, R. S. & Purugganan, M. D. Evolution of crop species: genetics of domestication and diversification. *Nat. Rev. Genet.* **14**, 840–852 (2013).
79. Olsen, K. M. & Wendel, J. F. Crop plants as models for understanding plant adaptation and diversification. *Front. Plant Sci.* **4**, 290 (2013).
80. Gaudinier, A. & Blackman, B. K. Evolutionary processes from the perspective of flowering time diversity. *N. Phytol.* **225**, 1883–1898 (2020).
81. Austen, E. J., Rowe, L., Stinchcombe, J. R. & Forrester, J. R. K. Explaining the apparent paradox of persistent selection for early flowering. *N. Phytol.* **215**, 929–934 (2017).
82. Lin, X., Liu, B., Weller, J. L., Abe, J. & Kong, F. Molecular mechanisms for the photoperiodic regulation of flowering in soybean. *J. Integr. Plant Biol.* **63**, 981–994 (2021).
83. Blackman, B. K. Changing responses to changing seasons: natural variation in the plasticity of flowering time. *Plant Physiol.* **173**, 16–26 (2017).
84. Izawa, T. Adaptation of flowering-time by natural and artificial selection in *Arabidopsis* and rice. *J. Exp. Bot.* **58**, 3091–3097 (2007).
85. Srikanth, A. & Schmid, M. Regulation of flowering time: all roads lead to Rome. *Cell Mol. Life Sci.* **68**, 2013–2037 (2011).
86. Blümel, M., Dally, N. & Jung, C. Flowering time regulation in crops—what did we learn from *Arabidopsis*? *Curr. Opin. Biotechnol.* **32**, 121–129 (2015).
87. Song, Y. H., Ito, S. & Imaizumi, T. Flowering time regulation: photoperiod- and temperature-sensing in leaves. *Trends Plant Sci.* **18**, 575–583 (2013).
88. Cao, D. et al. Molecular mechanisms of flowering under long days and stem growth habit in soybean. *J. Exp. Bot.* **68**, 1873–1884 (2017).
89. Jones, H. et al. Population-based resequencing reveals that the flowering time adaptation of cultivated barley originated east of the fertile crescent. *Mol. Biol. Evol.* **25**, 2211–2219 (2008).
90. Koo, B. H. et al. Natural variation in OsPRR37 regulates heading date and contributes to rice cultivation at a wide range of latitudes. *Mol. Plant* **6**, 1877–1888 (2013).
91. Ogiso-Tanaka, E. et al. Natural variation of the RICE FLOWERING LOCUS T 1 contributes to flowering time divergence in rice. *PLoS ONE* **8**, e75959 (2013).
92. Wang, S. et al. FLOWERING LOCUS T improves cucumber adaptation to higher latitudes. *Plant Physiol.* **182**, 908–918 (2020).
93. Chen, L. et al. Soybean adaption to high-latitude regions is associated with natural variations of GmFT2b, an ortholog of FLOWERING LOCUS T. *Plant Cell Environ.* **43**, 934–944 (2020).
94. Berner, M. et al. FRUITFULL controls SAUR10 expression and regulates Arabidopsis growth and architecture. *J. Exp. Bot.* **68**, 3391–3403 (2017).
95. Jiang, X. et al. FRUITFULL-like genes regulate flowering time and inflorescence architecture in tomato. *Plant Cell* **34**, 1002–1019 (2022).
96. Yang, J. et al. The SvFUL2 transcription factor is required for inflorescence determinacy and timely flowering in *Setaria viridis*. *Plant Physiol.* **187**, 1202–1220 (2021).
97. Lu, S. et al. Natural variation at the soybean J locus improves adaptation to the tropics and enhances yield. *Nat. Genet.* **49**, 773–779 (2017).
98. Liu, L. et al. Allele combinations of maturity genes E1-E4 affect adaptation of soybean to diverse geographic regions and farming systems in China. *PLoS ONE* **15**, e0235397 (2020).
99. Kwak, M., Toro, O., Debouck, D. G. & Gepts, P. Multiple origins of the determinate growth habit in domesticated common bean (*Phaseolus vulgaris*). *Ann. Bot.* **110**, 1573–1580 (2012).
100. Li, S. et al. Parallel domestication with a broad mutational spectrum of determinate stem growth habit in leguminous crops. *Plant J.* **96**, 761–771 (2018).
101. Liu, B. et al. The soybean stem growth habit gene Dt1 is an ortholog of Arabidopsis TERMINAL FLOWER1. *Plant Physiol.* **153**, 198–210 (2010).
102. Lopes, M. S. et al. Exploiting genetic diversity from landraces in wheat breeding for adaptation to climate change. *J. Exp. Bot.* **66**, 3477–3486 (2015).
103. Newton, A. C. et al. Cereal landraces for sustainable agriculture: a review. *Agron. Sustain. Dev.* **30**, 237–269 (2010).
104. Hour, A. L. et al. Genetic diversity of landraces and improved varieties of rice (*Oryza sativa* L.) in Taiwan. *Rice* **13**, 82 (2020).
105. Evenson, R. E. & Gollin, D. Assessing the impact of the green revolution, 1960 to 2000. *Science* **300**, 758–762 (2003).
106. Cheng, X., Wang, S. & Wang, L. *Descriptors and Data Standards [Vigna umbellata (Thunb.) Ohwi & Ohashi]* (China Agriculture Press, Beijing, 2006) (in Chinese).
107. Zhang, J., Song, Q., Cregan, P. B. & Jiang, G. Genome-wide association study, genomic prediction and marker-assisted selection for seed weight in soybean (*Glycine max*). *Theor. Appl. Genet.* **129**, 117–130 (2016).
108. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
109. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).

110. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using Hi-C yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
111. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. ArXiv e-Prints (2013).
112. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
113. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinform* **9**, 18 (2008).
114. Xu, Z. & Wang, H. LTR\_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).
115. Stanke, M., Steinkamp, R., Waack, S. & Morgenstern, B. AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res.* **32**, W309–W312 (2004).
116. Korf, I. Gene finding in novel genomes. *BMC Bioinform.* **5**, 59 (2004).
117. Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *Bmc. Bioinform.* **6**, 31 (2005).
118. Kim, D., Paggi, J. M., Park, C., Bennett, C. & Salzberg, S. L. Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat. Biotechnol.* **37**, 907–915 (2019).
119. Pertea, M. et al. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat. Biotechnol.* **33**, 290–295 (2015).
120. Tang, S., Lomsadze, A. & Borodovsky, M. Identification of protein coding regions in RNA transcripts. *Nucleic Acids Res.* **43**, e78 (2015).
121. Grabherr, M. G. et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
122. Haas, B. J. et al. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Res.* **31**, 5654–5666 (2003).
123. Haas, B. J. et al. Automated eukaryotic gene structure annotation using EvidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biol.* **9**, R7 (2008).
124. Lagesen, K. et al. RNAMmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res.* **35**, 3100–3108 (2007).
125. Lowe, T. M. & Eddy, S. R. TRNAScan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).
126. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1:100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).
127. Zou, C. et al. Evolutionary and expression signatures of pseudo-genes in *Arabidopsis* and rice. *Plant Physiol.* **151**, 3–15 (2009).
128. Zheng, Y. et al. ITAK: a program for genome-wide prediction and classification of plant transcription factors, transcriptional regulators, and protein kinases. *Mol. Plant* **9**, 1667–1670 (2016).
129. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).
130. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
131. Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).
132. Yu, G., Wang, L., Han, Y. & He, Q. ClusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS* **16**, 284–287 (2012).
133. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
134. Chen, S., Zhou, Y., Chen, Y. & Gu, J. Fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
135. McKenna, A. et al. The Genome Analysis Toolkit: a mapreduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
136. Vilella, A. J. et al. EnsemblCompara GeneTrees: complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* **19**, 327–335 (2009).
137. Price, A. L. et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
138. Danecek, P. et al. The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
139. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
140. Kim, M. S. et al. The patterns of deleterious mutations during the domestication of soybean. *Nat. Commun.* **12**, 97 (2021).
141. Zhang, C., Dong, S., Xu, J., He, W. & Yang, T. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. *Bioinformatics* **35**, 1786–1788 (2019).
142. Browning, B. L. & Browning, S. R. Genotype imputation with millions of reference samples. *Am. J. Hum. Genet.* **98**, 116–126 (2016).
143. Lippert, C. et al. FaST linear mixed models for genome-wide association studies. *Nat. Methods* **8**, 833–835 (2011).
144. Shim, H. et al. A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians. *PLoS ONE* **10**, e0120758 (2015).
145. González, J. R. et al. SNPAssoc: an R package to perform whole genome association studies. *Bioinformatics* **23**, 654–655 (2007).
146. Livak, K. J. & Schmittgen, T. D. Analysis of relative gene expression data using real-time quantitative PCR and the 2<sup>-ΔΔC<sub>T</sub></sup> method. *Methods* **25**, 402–408 (2001).
147. Clough, S. J. & Bent, A. F. Floral dip: a simplified method for agrobacterium-mediated transformation of *Arabidopsis thaliana*. *Plant J.* **16**, 735–743 (1998).

## Acknowledgements

This work was supported by the National Key Research and Development Program of China (grant nos. 2019YFD1001303 and 2019YFD1001300), the China Agriculture Research System of MOF and MARA (CARS-08), the Agricultural Science and Technology Innovation Program (ASTIP) from Chinese Academy of Agricultural Sciences, the Program of Protection of Crop Germplasm Resources in China (grant nos. 2019NWB036-07 and 19200385-6), the Third National General Survey and Collection of Crop Germplasm Resources (grant no. 19200354), and the Program Management Unit for Human Resources and Institutional Development, Research and Innovation of Thailand (grant no. B16F640185).

## Author contributions

L.X.W. and H.X. designed the research. L.X.W., P.S., K.L., X.Z.C., S.H.W., H.L.C., and Y.L.W. provided materials and information. J.T.Z., G.L.L., S.H.W., Y.H.C., Y.W. Z., X.X.Y., X.C., and A.H.S. contributed to phenotyping. J.T.G. and Z.Q.Z. conducted genome assembly, gene annotation, and population analyses. D.G., Z.H., and Y.Y. performed the experiments. J.T.G. and L.X.W. wrote and revised the manuscript with input and comments from the other authors. All authors read and approved the manuscript.

## 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/s41467-022-33515-2>.

**Correspondence** and requests for materials should be addressed to Hua Xie or Lixia Wang.

**Peer review information** *Nature Communications* thanks Fanjiang Kong and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permission 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 4.0 International License, which permits use, sharing, adaptation, 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 license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license 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 license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

---

<sup>1</sup>Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Beijing, China. <sup>2</sup>Institute of Biotechnology, Beijing Academy of Agriculture and Forestry Sciences, Beijing, China. <sup>3</sup>Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing, China. <sup>4</sup>College of Agriculture, Yangtze University, Jingzhou, China. <sup>5</sup>Institute of Rice Research, Guangxi Academy of Agricultural Sciences, Nanning, China. <sup>6</sup>Department of Agronomy, Faculty of Agriculture at Kamphaeng Saen, Kasetsart University, Nakhon Pathom, Thailand. <sup>7</sup>Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, China. <sup>8</sup>College of Agriculture, Shanxi Agricultural University, Taiyuan, China. <sup>9</sup>Crop Research Institute of Hunan Province, Changsha, China. <sup>10</sup>These authors contributed equally: Jiantao Guan, Jintao Zhang, Dan Gong. ✉ e-mail: [xiehua@baafs.net.cn](mailto:xiehua@baafs.net.cn); [wanglixia03@caas.cn](mailto:wanglixia03@caas.cn)