Genome Biology

RESEARCH

Open Access

A special short-wing petal faba genome and genetic dissection of floral and yield-related traits accelerate breeding and improvement of faba bean

Rong Liu[1†], Chaoqin Hu[2†], Dan Gao[3†], Mengwei Li[1†], Xingxing Yuan[4†], Liyang Chen[3], Qin Shu[1], Zonghe Wang[1], Xin Yang[2], Zhengming Dai[2], Haitian Yu[2], Feng Yang[2], Aiqing Zheng[2], Meiyuan Lv[2], Vanika Garg[5], Chengzhi Jiao[3], Hongyan Zhang[6,7], Wanwei Hou[6,7], Changcai Teng[6,7], Xianli Zhou[6,7], Chengzhang Du[8], Chao Xiang[9], Dongxu Xu[10], Yongsheng Tang[11], Annapurna Chitikineni[5], Yinmei Duan[12], Fouad Maalouf[13], Shiv Kumar Agrawal[13], Libin Wei[14], Na Zhao[14], Rutwik Barmukh[5], Xiang Li[15], Dong Wang[16], Hanfeng Ding[16], Yujiao Liu[6*], Xin Chen[4*], Rajeev K. Varshney[5*], Yuhua He[2*], Xuxiao Zong[1*] and Tao Yang[1*]

†Rong Liu, Chaoqin Hu, Dan Gao, Mengwei Li and Xingxing Yuan contributed equally to this work.

*Correspondence: 1996990028@qhu.edu.cn; cx@jaas.ac.cn; rajeev.varshney@murdoch.edu.au; hyh@yaas.org.cn; zongxuxiao@caas.cn; yangtao02@caas.cn

[1] State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Haidian District, Beijing 100081, China
[2] Food Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, Yunnan 650205, China
[4] Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, Jiangsu 210014, China
[5] State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, WA 6150, Australia
[6] State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, Qinghai 810016, China
Full list of author information is available at the end of the article

Abstract

Background: A comprehensive study of the genome and genetics of superior germplasms is fundamental for crop improvement. As a widely adapted protein crop with high yield potential, the improvement in breeding and development of the seeds industry of faba bean have been greatly hindered by its giant genome size and high outcrossing rate.

Results: To fully explore the genomic diversity and genetic basis of important agronomic traits, we first generate a de novo genome assembly and perform annotation of a special short-wing petal faba bean germplasm (VF8137) exhibiting a low outcrossing rate. Comparative genome and pan-genome analyses reveal the genome evolution characteristics and unique pan-genes among the three different faba bean genomes. In addition, the genome diversity of 558 accessions of faba bean germplasm reveals three distinct genetic groups and remarkable genetic differences between the southern and northern germplasms. Genome-wide association analysis identifies several candidate genes associated with adaptation- and yield-related traits. We also identify one candidate gene related to short-wing petals by combining quantitative trait locus mapping and bulked segregant analysis. We further elucidate its function through multiple lines of evidence from functional annotation, sequence variation, expression differences, and protein structure variation.

Conclusions: Our study provides new insights into the genome evolution of Leguminosae and the genomic diversity of faba bean. It offers valuable genomic and genetic resources for breeding and improvement of faba bean.

Keywords: Faba bean, Short-wing petal, Genome assembly, Genomic diversity, GWAS, Yield-related loci, Breeding, Improvement

Background

Faba bean (*Vicia faba* L., 2n = 2x = 12) is a widely adapted cool season legume rich in protein (> 26% in 100 g dry seed) and multiple favorable nutrient elements, such as starch, fiber, iron, zinc, and vitamins, which can be used both as food and feed [1–3]. Faba bean had the highest yield (2.19 Mg ha^{-1}) among legumes after soybean in 2021 (www.fao.org/faostat/en/#data) and was endowed with a significant ecological advantage due to its nitrogen-fixing capacity [4, 5]. However, despite its critical role in human nutrition and health as well as sustainable agriculture [6, 7], genomic studies and molecular breeding research on faba bean lag far behind those on other legume crops, such as soybean [8], common bean [9], pea [10] and chickpea [11, 12], due to its giant genome size (~ 13 Gb) [6]. The recent accomplishment of high-quality faba genome assembly provides an excellent genomic platform for faba bean [13]. However, as one of the earliest domesticated food legumes, faba bean (*Vicia faba* L.) was inferred to have been domesticated in the Near East more than 10,000 years ago [14, 15]. Although the extant wild progenitor of faba bean has not been discovered, rich genetic diversity has accumulated in faba bean germplasm, resulting in significant population genetic structure under long-term domestication and cultivation as well as widespread adaptation [6, 13, 15]. Therefore, genomic analyses using different genotypes with significant genetic divergence and further exploration of genetic diversity are needed for a better understanding of the population differentiation and phenotypic variation in faba bean [16, 17].

Faba bean is partially allogamous and depends on insect pollination, exhibiting a high outcrossing rate (mean 30–60%) [18, 19]. The common mixed self- and outcrossing habit of faba beans, along with the lack of autofertility, results in a slow and costly process for controlling pollination to maintain the desired traits in faba bean breeding [19, 20]. In addition, the yield of faba bean partially depends on the abundance, activity, and pollination efficiency of pollinators, which are strongly affected by the geographical environment and climate change [21, 22]. Consequently, breeders often opt for autofertile lines with limited outcrossing rates in order to strike a balance between these factors and ensure crop productivity [19, 20]. We previously identified a special faba bean germplasm, VF8137, from China with a rare short-wing petal trait accompanied by a low outcrossing rate of less than 5% [23]. This germplasm is highly valuable in faba bean breeding because it serves as a foundational parent for the rapid purification of varieties and the efficient fixation of excellent target traits. However, the genetic basis of short-wing petals and other important agronomic traits is still unclear, becoming a serious obstacle to the development of the molecular breeding of faba beans.

To fully explore the genomic diversity and genetic basis of important agronomic traits, comprehensive genomic and genetic analyses were carried out on the specific faba bean germplasms VF8137 and 558 from global faba bean germplasm accessions in this study. First, we present a high-quality de novo genome assembly of VF8137 via PacBio HiFi sequencing and chromosome conformation capture (Hi-C) sequencing and reveal its genomic evolutionary characteristics as well as unique pangenes among the three

different faba bean genomes. Second, the genome diversity and population genetic structure of 558 accessions of faba bean germplasm were explored. Third, a 3-year and three-location genome-wide association study (GWAS) identified several genetic loci associated with the adaptation and yield-related traits of faba bean. Finally, by integrating genomic resources, quantitative trait locus (QTL) analysis, bulked segregant analysis (BSA), transcriptome sequencing (RNA-Seq), and haplotype analysis, we identified one candidate gene related to short-wing petals and elucidated its function through multiple functional annotations, sequence variations, expression differences, and protein structure variations. These findings will not only lay a solid foundation for understanding the genome evolution of Leguminosae and the genomic diversity of faba bean but also significantly accelerate the breeding, improvement, and development of the seed industry of faba bean.

Results

De novo assembly and annotation of a special short-wing petal faba bean

VF8137 is a special short-wing petal faba bean germplasm found in China (Fig. 1a) that has a low outcrossing rate of less than 5% [23]. In the bee pollination experiment, short-wing petal VF8137 had a significantly lower visiting frequency and duration than germplasms (such as GF45 and TF29) with normal wing petals did, which was probably due to the structural variation in its flower organ (Additional File 1: Fig. S1). To determine the genomic features of this special faba bean germplasm, VF8137 was purified by single-seed descent for five generations to construct a genome assembly. The genome size of VF8137 was estimated to be 11.77 Gb on the basis of 445.12 Gb of Illumina sequence data (~38 × genomic coverage) using *K*-mer analysis, with low heterozygosity (0.49%) and a high proportion of repeat sequences (92.34%) (Additional File 1: Fig. S2). Based on 490.54 Gb of PacBio HiFi reads (~41.57 × genomic coverage) (Additional File 2: Table S1), an 11.81 Gb genome assembly of VF8137 was constructed with a contig N50 size of 10.21 Mb. Using Hi-C scaffolding, 11.07 Gb (93.76%) of the initial assembly was

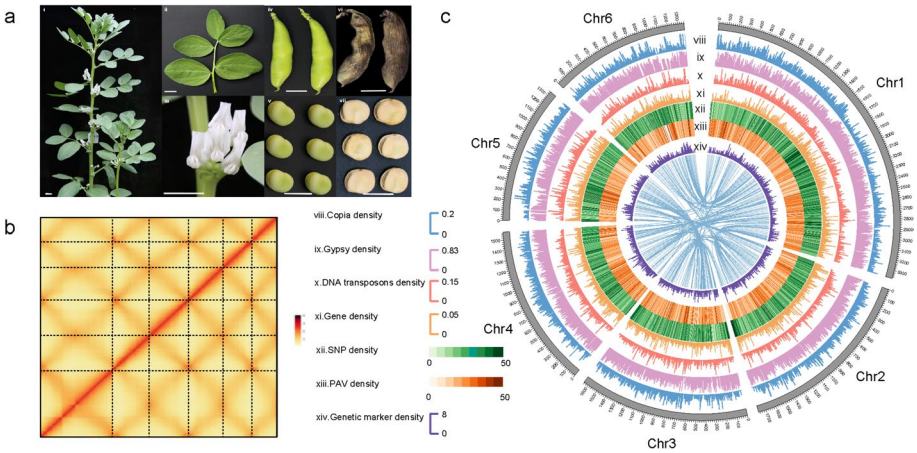

Fig. 1 Morphological and genomic features of the specific faba bean germplasm VF8137. **a** Photos of different tissues. i = plant at the flowering stage, ii = leaf, iii = inflorescence, iv = fresh pods, v = fresh seeds, vi = dry pods, and vii = dry seeds. Scale bars = 2 cm. **b** Hi-C interaction map. **c** Genome features. The circular outer layer with the physical map position and the innermost layer represent the six chromosomes and interchromosomal synteny, respectively. viii = Copia density, ix = Gypsy density, x = DNA transposon density, xi = Gene density, xii = SNP density, and xiii = PAV density, xiv = Genetic marker density. A bin size of 5 Mb is applied for SNP, PAV, and genetic marker density

anchored to six chromosome-level pseudomolecules (Fig. 1b and c, Additional File 2: Table S2). Chr1 of VF8137 had the longest length of 3.38 Gb and was characterized by a satellite (Fig. 1b), and even the shortest Chr5 was 1.34 Gb (Additional File 2: Table S3). Genome features were compared among two assemblies of faba bean and one assembly of *Pisum sativum* as well as *Medicago truncatula*, which revealed a general trend of high repeat density, low gene density, and high GC content in the centromere region, especially for *Medicago truncatula* (Additional File 1: Fig. S3).

A total of 11.16 Gb of repetitive sequences were identified, accounting for 94.49% of the VF8137 assembly (Additional File 2: Table S4), which is consistent with the results of K-mer analysis (Additional File 1: Fig. S2) but larger than that of the existing faba assembly for Hedin/2 (9.46 Gb, 78.9% of the genome) [13]. The largest group of TEs was LTR-Gypsy, followed by LTR-Copia, with genome proportions of 74.77% and 14.53%, respectively (Additional File 2: Table S4).

A total of 47,215 protein-coding genes were identified by using a combination of ab initio, homology, and transcriptomic evidence-based predictions (Additional File 1: Fig. S4). Among these, 96.6% were functionally annotated by any of the public databases (see methods, Additional File 2: Tables S5 and S6). Although the number of predicted genes for VF8137 is greater than that for the Hedin/2 and Tiffany assemblies[13], it is comparable to the number of genes reported for other closely related legume species, such as *Pisum sativum* [10] and *Medicago truncatula* [24], which have smaller genome sizes. The average length of the protein-coding genes was 3533.3 bp, with a mean CDS length of 979.0 bp (Additional File 2: Table S5). Compared with those in peas, the average length of genes in faba beans was greater, with a longer mean length of introns but a shorter mean length of CDS [10].

To further assess the quality of the VF8137 assembly, the PacBio HiFi and Illumina reads were mapped back to the genome assembly. This resulted in 99.86% and 99.73% alignment rates, with 99.73% and 99.04% of the assemblies covered by at least 10 reads, respectively (Additional File 2: Table S7). The genome and protein BUSCO completeness of the VF8137 assembly was 98.7% and 95.1%, respectively, which are close to those of the Tiffany assembly but greater than those of the Hedin/2 assembly (Additional File 2: Table S8). Although the consensus quality (48.6) and genome completeness (94.9%) of the VF8137 assembly were slightly lower than those of the Hedin/2 assembly (Additional File 2: Table S9), a greater long terminal repeat (LTR) assembly index (LAI) (12.56) indicated greater LTR completeness in the VF8137 assembly (Additional File 2: Table S10). Taken together, these results indicate a high-quality genome assembly and annotation of the VF8137 germplasm that provides a strong foundation for obtaining biological insights into various aspects of faba beans.

Genome evolution of *V. faba*

To study the genomic evolution of *V. faba*, a phylogenetic tree was constructed by using single-copy genes of 27 plant genomes, including 21 species from Leguminosae and four representative species of core Eudicotyledonous plants (Additional File 2: Table S11). Consistent with previous studies, *V. faba* grouped into the Galegoid branch of Papilionatae in Leguminosae, and the divergence time of faba bean and its closely related species pea was approximately 9.8–15.3 Mya (Fig. 2a). Compared with the Millettoid branch,

the Galegoid branch presented contraction of 2457 gene families (Additional File 1: Fig. S5), which were mainly related to molecular functions such as ABC transporters and the MAPK signaling pathway (Additional File 1: Fig. S6, Additional File 2: Table S12). Similar to those in *Pisum sativum* and *Medicago truncatula*, only two whole-genome duplication (WGD) events occurred during the evolution of *V. faba*: one was a triploidy event (γ event) shared by core Eudicotyledonous plants, and the other was a paleopolyploidy event approximately 55 Mya ago in Papilionatae [25] (Additional File 1: Fig. S7). The genome of *V. faba* showed good collinearity with that of *P. sativum* and *M. truncatula*;

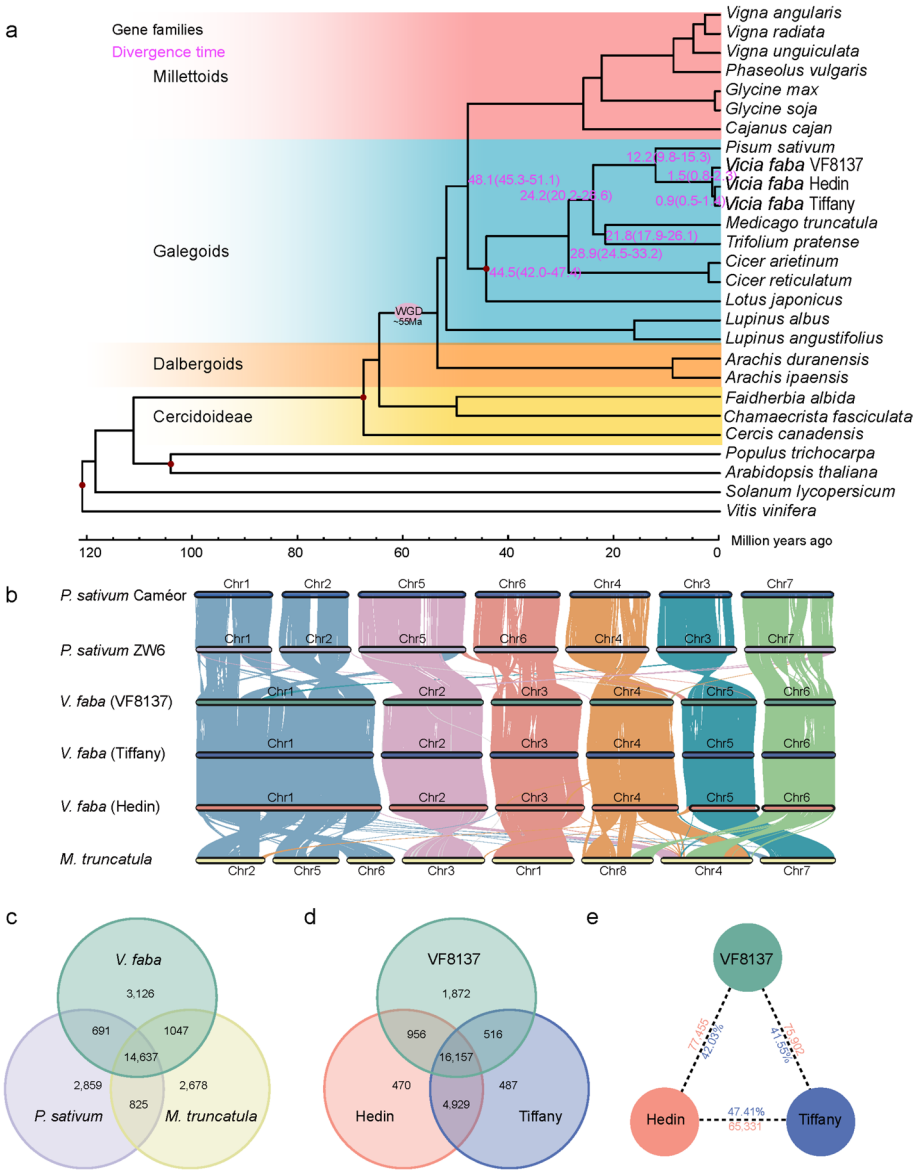

Fig. 2 Genome evolution and pangenome analysis of *Vicia faba*. **a** Phylogenetic relationships of 21 Leguminosae plants and four other dicotyledonous plants. The purple numbers denote the duration of species differentiation in million years. **b** Collinearity among three faba genome assemblies as well as the genome assemblies of *P. sativum* and *M. truncatula*. **c** Gene family analysis of *V. faba*, *P. sativum*, and *M. truncatula*. **d** Gene family analysis of the three faba genome assemblies, namely, VF8137, Hedin, and Tiffany. **e** Genome variation among the three faba genome assemblies. Red and blue represent the PAV number and mapping rate between genomes, respectively

however, large chromosomal rearrangements were observed. For example, compared with *P. sativum*, the longest Chr1 of *V. faba* appears to be formed by the fusion of two chromosomes of *P. sativum* and three chromosomes in the case of *M. truncatula* (Fig. 2b). In addition, comparative analysis of homologous gene families revealed a total of 14,637 gene families shared by *V. faba*, *P. sativum*, and *M. truncatula* and the greatest number of unique gene families in *V. faba* (Fig. 2c).

Pan-genome analysis of *V. faba*

The annotated genes of the three faba bean genomes, VF8137, Hedin, and Tiffany, were clustered into gene families and compared via homologous alignment. A total of 16,157 gene families were shared by all three genomes, accounting for 63.6% of the total gene families and representing the core gene families of faba bean. In addition, VF8137 exhibited the highest number of unique gene families, which were found to be significantly enriched in mRNA processing, the proton-transporting ATP synthase complex, iron-sulfur cluster binding, and other pathways on the basis of GO enrichment analysis (Fig. 2d, Additional File 1: Fig. S8, Additional File 2: Table S13). In contrast, Hedin/2 and Tiffany showed lower numbers of unique gene families with different GO enrichment pathways than VF8137 did (Fig. 2d, Additional File 1: Fig. S8, Additional File 2: Table S13). Besides, fewer common gene families were observed for VF8137 with Hedin/2 and Tiffany than for Hedin/2 and Tiffany (Fig. 2d). Furthermore, pairwise genomic polymorphism analyses among the three faba bean genomes revealed a large number of SNPs (84.6–97.5 Mb), InDels (55.3–64.0 K) and PAVs (65.3–75.3 K) (Additional File 2: Table S14). The number of genomic variations between VF8137 and Hedin as well as between VF8137 and Tiffany were greater than those between Hedin and Tiffany (Fig. 2e). Notably, the number of SVs between VF8137 and Hedin/2 or Tiffany is dominated by INS and DEL (> 85%), spanning a genome length of approximately 0.9 Gb, whereas a small number of INVs (~ 0.5%) cover a wider range of genome regions close to 1.1 Gb. On the other hand, for the SVs detected between Hedin and Tiffany, INSs and DELs were dominant in both number and length (Fig. 2e; Additional File 2: Table S14). These results indicate that the genetic background of VF8137 is significantly different from that of Hedin and Tiffany [26], and the assembly and annotation of VF8137 provide valuable genomic and genetic resources for pangenomic analysis of *Vicia faba*.

A high proportion of TEs and their biased insertions in the genome of *V. faba*

By comparing the genomes of 25 representative species from Fabaceae and Poaceae (Additional File 2: Table S11), we found that both LTR length and the intron/exon length ratio were highly positively correlated with genome size, and the correlation among species from the same family was even greater (Fig. 3a and b). In addition, the frequency of TEs upstream and downstream of genes was analyzed, and the results showed that the frequencies of DNA TEs and LTR-Copia were greater near the 2 kb upstream and downstream regions of the genes, whereas the frequency of LTR-Gypsy was greater farther from the genes (Fig. 3c). In addition, the number and length of TEs in the intron region were significantly greater than those in the CDS region (Fig. 3d and e). Finally, we compared the five longest homologous genes of *V. faba* and *P. sativum* and found that they did not differ significantly in the number or length of exons, but all of them in faba bean

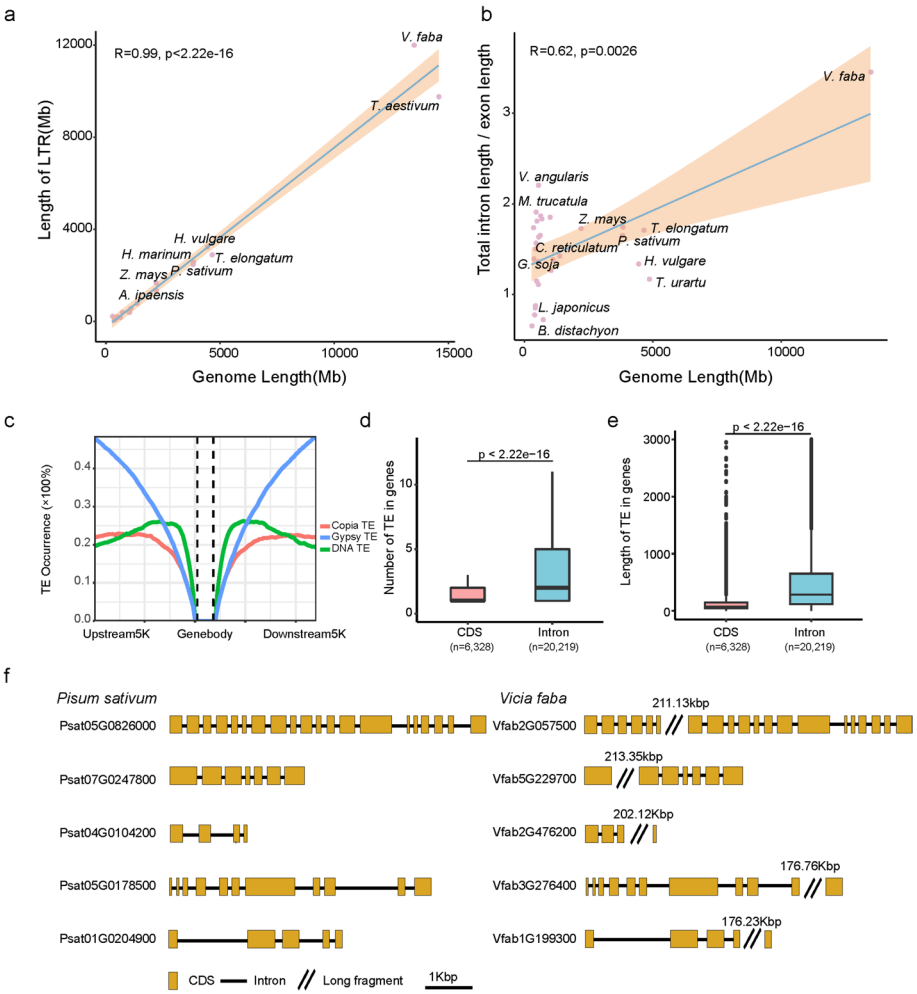

Fig. 3 TE analyses revealed its effect on genome size and gene length. **a** Correlation between LTR length and genome size across Leguminosae and Poaceae. **b** Correlations between intron versus exon length and genome size across Leguminosae and Poaceae. **c** Preference distributions of three main types of TEs upstream and downstream of genes. **d** and **e** Number and length of TEs in the CDS and intron regions. **f** Gene structure of the five longest homologous genes in *P. sativum* and *V. faba*

had transposon insertions larger than 150 kb in the intron region (Fig. 3f). These results suggest that a high proportion of TEs and their preferred insertions in intergenic and intron regions resulted in the expansion of the faba bean genome and an increase in the average gene length.

Genomic polymorphisms and population genetic structure of 558 faba bean germplasms

To investigate genomic polymorphisms in landrace and faba bean cultivars, a total of 1,332,986 high-quality SNPs were identified from a set of 558 worldwide accessions of faba bean germplasm based on our previous development of the Faba_bean_130K targeted next-generation sequencing (TNGS) SNP genotyping platform [27] (Fig. 4a, Additional File 2: Tables S15 and S16). A total of 3.4% and 9.3% of the SNPs were found in exons and introns, respectively, while 11.5% and 75.8% of the SNPs were located in regulatory and intergenic regions, respectively (Additional File 2: Table S16).

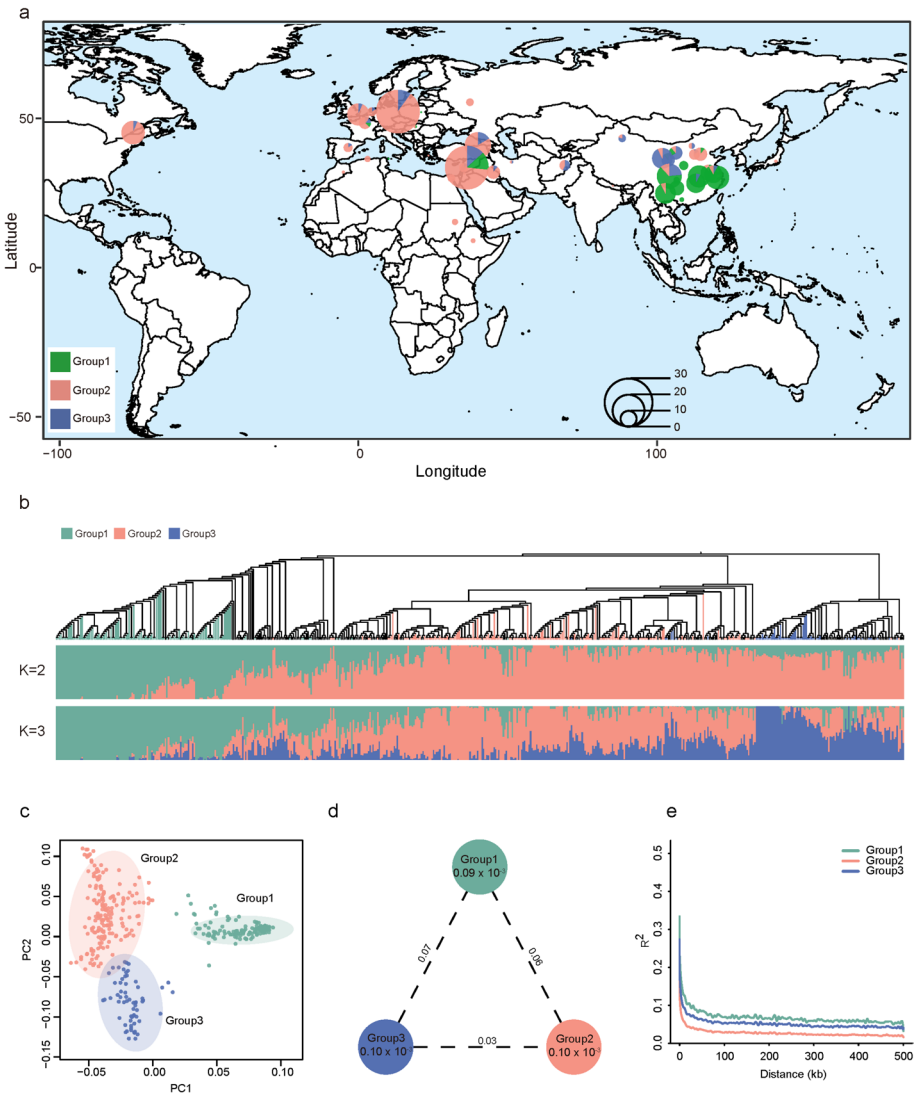

Fig. 4 Population genomic analysis of 558 accessions of faba bean germplasm. **a** Geographical distribution of 558 accessions of faba bean germplasm. Provincial capital coordinates were applied for samples from the same province in China, whereas capital coordinates were used for samples from other countries. **b** Phylogenetic analysis and ADMIXTURE analysis at $K = 2$ and 3. **c** PCA. **d** Population genetic diversity and genetic differentiation between three genetic groups of faba beans. **e** LD decay analysis of three genetic groups. Green, pink, and blue represent group 1, group 2, and group 3, respectively

Through population genetic structure analysis using SNPs, three distinct genetic groups were identified among these 558 faba bean accessions, with some showing admixture (Fig. 4b). This classification was further supported by phylogenetic analysis (Fig. 4b). Furthermore, the PCA separated all the samples into three distinct genetic groups, which corresponded well with the three major genetic groups of the ADMIXTURE results (Fig. 4c). There is obvious geographical differentiation of faba beans, as evidenced by the substantial variation in genetic composition between Asia and Europe as well as North America. Interestingly, PCA was also conducted using SNPs integrated with the variation from the three genomes and the 558 faba bean population, resulting in VF8137 being placed within group 1 and Hedin/2 and Tiffany being classified under

group 2 (Additional File 1: Fig. S9). Such genetic divergence, in turn, aligns with their respective geographic origins. In addition, the faba beans in China are composed of all three genetic groups but are dominated by group 1 (48.9%). The variation pattern in China exhibits geographical differences between northern and southern China, as indicated by the differences between spring sowing and autumn sowing (Fig. 4a).

Among the three genetic groups, group 2 and group 3 presented equivalent nucleotide diversity ($\pi = 0.1E-03$), whereas group 1 exhibited a slightly lower level of diversity ($\pi = 0.09E-03$) (Fig. 4d). Furthermore, the genetic differentiation between group 1 and group 3 was the greatest ($F_{ST} = 0.07$), followed by that between group 1 and group 2 ($F_{ST} = 0.06$) and that between group 2 and group 3 ($F_{ST} = 0.03$) (Fig. 4d). In addition, linkage disequilibrium (LD, R^2) was calculated with SNPs for the three genetic groups, and the LD decay distance increased from 1 kb in group 2 to 3.5 kb in group 1 (Fig. 4e).

Selective signals during population differentiation of faba bean

To identify putative selective genome regions involved in the population differentiation of faba bean, the F_{ST} over 20 kb sliding windows were estimated via comparisons of the three genetic groups (group 1 vs. group 2, group 2 vs. group 3, and group 1 vs. group 3). With the first 5% of the F_{ST} value as the threshold, we identified 4466, 3894, and 4210 selective sweeps for these three comparisons, encompassing 3248, 2848, and 2988 protein-coding genes covering 89.3 Mb, 77.9 Mb, and 84.2 Mb of the assembled genome, respectively (Additional File 2: Table S17, S18, and S19). Additionally, we analyzed the number of structural variations (SVs) present in the selective sweep regions between group 1 and group 2, which detected 676 and 706 SVs in the selective sweep regions of VF8137 vs. Hedin/2, and VF8137 vs. Tiffany, respectively (Additional File 2: Table S20). To investigate the functional differentiation between faba beans in China and those outside of China (Fig. 4a), gene ontology (GO) analysis was performed on the 3248 candidate-selected genes between group 1 and group 2. The results showed enrichment of genes involved in biological processes such as protein modification and biological regulation, cellular components of membrane-bound organelles, and molecular functions such as catalytic and kinase activity (Additional File 1: Fig. S10). In addition, the candidate-selected regions of group 1 and group 3, representing the ecological differentiation of faba bean in North and South China, contained several genes homologous to *FT* genes related to the photoperiod pathway in *G. max*, *M. truncatula*, and *P. sativum* [28–30] (Additional File 2: Table S19).

GWAS analysis of ten agronomic traits of faba bean

To explore the morphological variation of faba bean, 3-year and three-site phenotypic identification of nine quantitative traits in Hebei, Qinghai, and Yunnan between 2019 and 2021 as well as identification of the qualitative trait hilum color at Qinghai in 2021 were performed on 558 accessions of faba bean germplasm (Additional File 2: Table S21). A comparison of the character distributions among the different experiments revealed that year and location had strong effects on flowering time, plant height, and pod size (Additional File 1: Fig. S10). In addition, the average flowering time for faba bean germplasms in Yunnan, which represents the autumn sowing area, was longer than those in Hebei and Qinghai, which belong to the northern

spring sowing area of China. The average plant height and pod size in Qinghai were the greatest of all three sites, while the pod and seed size as well as the pod numbers of the fruiting germplasms in Yunnan were the smallest (Additional File 1: Fig. S11, Additional File 2: Table S22).

To determine the genetic basis of the important agronomic traits of faba bean, a GWAS was conducted on 558 accessions of faba bean in combination with phenotyping data obtained from seven experiments and genome-wide SNPs identified from genotyping data. A total of 545 SNPs spanning 222 genes were significantly associated with the ten agronomic traits (Additional File 2: Table S23). For flowering time, 11 significantly associated SNPs in the *Vfab5G146000* gene on Chr5 were detected via linkage decay (LD) block analysis in Yunnan in 2019 (YN19) (Fig. 5a, b, c). A distinct SNP composition and haplotype distribution of this gene were observed in different genetic groups (Fig. 5d, e). In addition, the two different haplotypes of this gene presented significant differences in flowering time (Fig. 5f). Furthermore, accessions with different haplotypes exhibit a geographic distribution trend from north–south and significant latitudinal differences (Fig. 5g, h). BLAST results suggested that it is an ortholog of the *FT* gene known to be related to the photoperiod pathway in *G. max*, *M. truncatula*, and *P. sativum* [28–30] and is also a candidate gene selected between Group 1 and Group 3, representing the ecological differentiation of faba bean in northern and southern China mentioned previously (Additional File 2: Table S19). Notably, collinearity analysis of the *FT* gene in three faba bean genome assemblies revealed significant differences in the upstream, downstream, and intron regions of the key gene *FT* for flowering period adaptability between VF8137 and Hedin/2 or Tiffany (Additional File 1: Fig. S12). Two unique photoperiod-related regulatory elements in the upstream region and a long insertion in the intron region were identified only in the *FT* gene of VF8137, shedding light on the molecular mechanism of adaptive differentiation in faba beans.

Plant height plays a vital role in determining the lodging resistance and yield of plants [31, 32]. In this study, the plant heights of the faba beans at the different test sites over the 3 years showed relatively consistent trends, except for those in the experiments of Qinghai in 2019 (QH19) and YN19; the average plant height was greatest in group 3, followed by group 2 and group 1 (Additional File 1: Fig. S13). A major signal located on Chr1 associated with plant height was repeatedly detected at the Hebei site in 2019 and 2020 (Additional File 1: Fig. S14 a and b). Three candidate genes related to plant height were further identified, and distinct haplotypes of the three genes showed significant differences in plant height (Additional File 1: Fig. S14 c, d, e). *Vfab1G475000* was predicted to be an ortholog of the SWEETIE-like gene in *P. sativum* and *M. truncatula*; this gene is related to intercellular sugar transport and involved in a wide range of physiologically important processes, including plant growth and development, host-pathogen interactions, and abiotic stress responses [33–35]. *Vfab1G475700* was annotated as an ortholog of the SNARE-associated Golgi family protein in *Medicago truncatula*, which was suggested to involve in mediating the post-Golgi trafficking required for auxin-mediated development in *Arabidopsis* [36]. The functions of *Vfab1G475900* are uncharacterized, and further studies are needed to determine its effects on plant height.

Seed size and weight are closely related to crop yield and quality and are regarded as key traits in crop domestication and breeding [37, 38]. A comparison of the

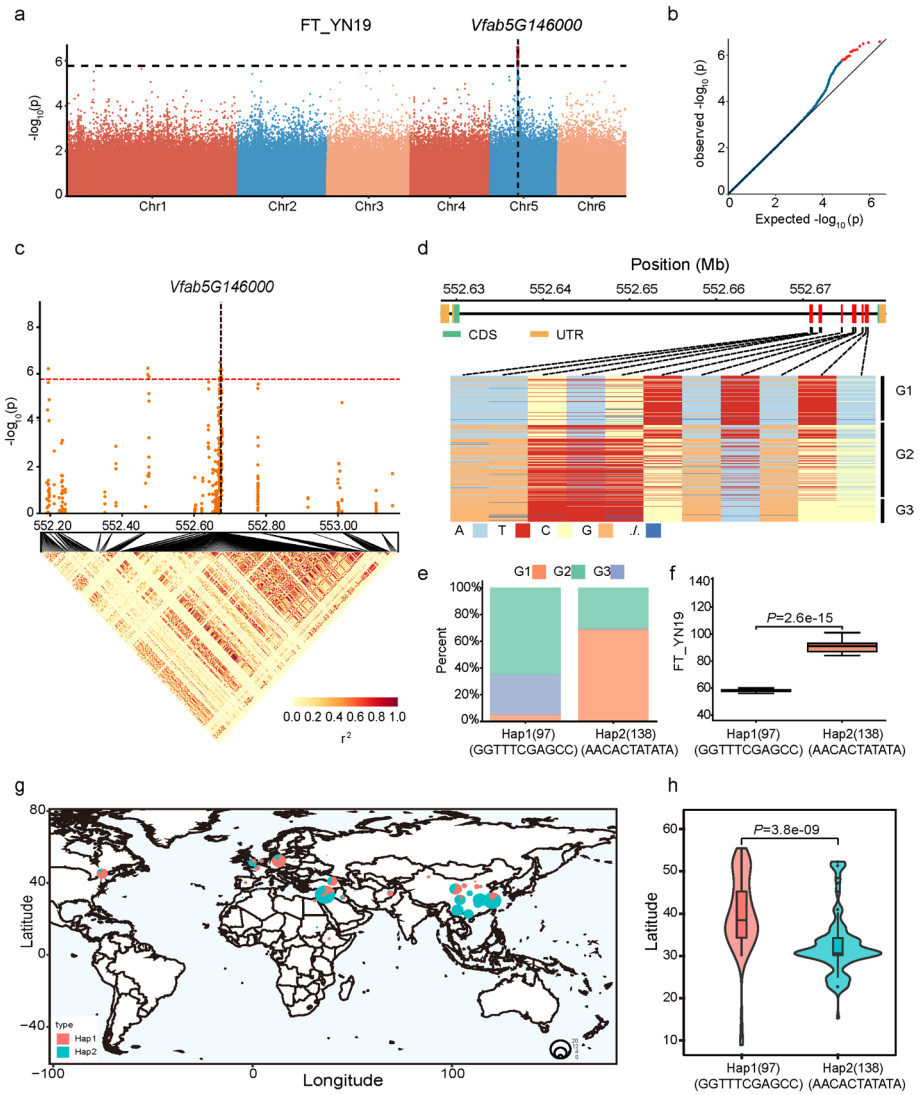

Fig. 5 GWAS and genetic analyses of candidate genes associated with flowering time. **a** SNPs significantly associated with flowering time in YN19. **b** QQ plot of flowering time according to GWAS. **c** LD block analysis for significant SNPs significantly associated with FT_YN19. **d** Allelic composition of the candidate gene *Vfab5G14600* (*FT*) controlling flowering time. **e** Haplotype composition of *Vfab5G14600* (*FT*) in different genetic groups of 558 accessions of faba bean germplasm. **f** Phenotypic differences in flowering time among distinct haplotypes of *Vfab5G14600* (*FT*). **g** Geographic distribution of 558 accessions of faba bean germplasm with different haplotypes of *Vfab5G14600* (*FT*). **h** Latitudinal differences in samples with different haplotypes of *Vfab5G14600* (*FT*)

one-hundred-seed weight (100SW) of different experimental sites across 3 years revealed that the average value of 100SW in group 3 was greater than that in group 1 and group 2. Considering the geographical differentiation of different genetic groups, the difference in 100SW may be related to geographical environment adaptability, especially for those faba bean germplasms from China whose geographic information is clear (Additional File 1: Fig. S15). A similar trend was observed for the seed size traits, including seed length (SL), seed width (SW), and seed thickness (ST) (Additional File 2: Table S22). GWAS analysis repeatedly revealed two loci on Chr1 (Chr1: 557,845,769–557,845,769

and Chr1: 862,330,442–862330442) that were significantly associated with SL, SW, and 100SW in four and three experiments (Additional File 1: Fig. S16 and S17, Additional File 2: Table S23). The genotypes of the two associated SNPs differed significantly among 100SW, SL, and SW in several experiments (Additional File 1: Fig. S16 and S17). These polymorphisms were further shown to be located upstream of two candidate genes, *Vfab1G141200* and *Vfab1G792100* (< 100 bp), suggesting that these polymorphisms may influence the ability of these two genes to control seed size and weight. *Vfab1G141200* was annotated as a PRA1 family protein encoding Rab receptors preferentially expressed in developing tissues in *Arabidopsis thaliana* [39]; this protein was also reported to be involved in the trafficking of intercellular vesicles and related to cell wall remodeling during fruit maturation [40, 41]. *Vfab1G792100* was annotated as a member of the 2OG-Fe(II) oxygenase superfamily and plays diverse roles in plant biological processes, including DNA demethylation, plant hormone biosynthesis, and the synthesis of various specialized metabolites. It is also reported to be involved in the biosynthesis of the phytohormone gibberellins (GAs), which are important regulators of seed development [38, 42].

In addition, one significant signal on Chr3 was repeatedly detected to be associated with pod size and was ~ 14 kb away from the long chain acyl-CoA synthetase 6 gene (Additional File 1: Fig. S18, Additional File 2: Table S23), which is involved in regulating lipid synthesis and degradation and has multiple roles in plant development and stress resistance [43]. Furthermore, a *PPO* gene demonstrated to control hilum color (HC) in a previous study [13] was also associated with HC in this study (*Vfab1G537500*), confirming that the GWAS results are reliable in this study (Additional File 1: Fig. S19, Additional File 2: Table S23).

Genetic dissection of short-wing petals in VF8137 plants via the combination of QTL mapping and BSA analysis

To explore the genetic basis of important agronomic traits based on the high-quality reference genome of VF8137, we reanalyzed some previous phenotypic and genotyping data from an F_2 population consisting of 167 individuals derived from a cross between VF8137 (short wing-petal) and H3712 (normal wing-petal)[44] (Additional File 1: Fig. S20, Additional File 2: Table S24). The population was genotyped using the 130 K TNGS genotyping platform developed by our group [27, 44], and a total of 255,925 high-quality SNP markers for 167 F_2 lines were clustered into 3369 bin markers for genetic map construction. A high-density genetic linkage map was constructed into six linkage groups spanning a total length of 1876.92 cM with an average marker spacing of 0.56 cM (Additional File 1: Fig. S21, Additional File 2: Table S25). In addition, it exhibited exact consistency with the genome assembly of VF8137, which provides new evidence for the accuracy of the genome assembly of VF8137 (Additional File 1: Fig. S22). Subsequently, QTL analysis was performed for 12 out of 16 agronomic traits (Additional File 2: Table S26), and nine QTLs were found to be associated with seven agronomic traits, with LOD values ranging from 4.23 to 78.99 (Fig. 6a, Additional File 1: Fig. S23, Additional File 2: Table S27). Among the 10 QTLs, one QTL on chromosome 3 associated with wing petal length (WL) was named wing-petal length 3 (WL3), which presented the highest LOD (78.99) and PVE (87.9%),

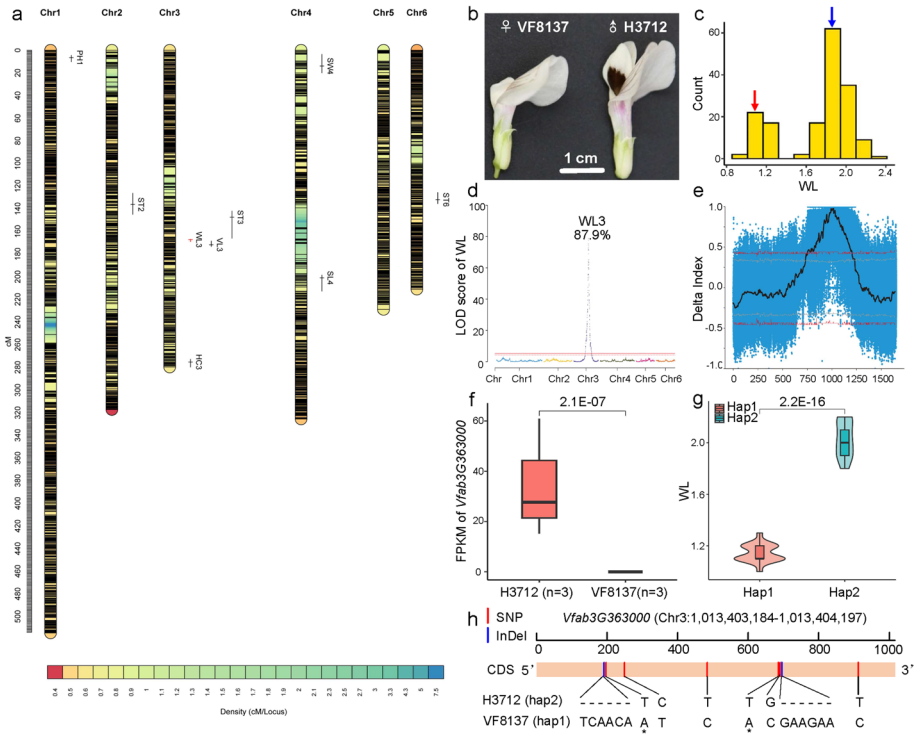

Fig. 6 Genetic dissection of short-wing petals of VF8137 via combined QTL mapping and BSA analysis.**a** QTL analysis detected nine QTLs associated with seven agronomic traits in faba bean. **b** Differential flowers of the parents VF8137 and H3712 in the F_2 population. **c** Phenotype distribution of wing petal length (WL) in the F_2 population. **d** WL3, which is associated with wing petal length, has large LOD and PVE values. **e** BSA analysis identified genome regions related to WL on Chr3. **f** FPKM values of the candidate gene (*Vfab3G363000*) in the flower tissue of the parents VF8137 and H3712. **g** WLs of the two haplotypes of *Vfab3G363000*. **h** Genotypes of *Vfab3G363000* in the parents VF8137 and H3712

with sharp QTL peaks in the 15.79 Mb genomic region (Fig. 6b, c, d; Additional File 2: Table S27).

In addition, a bulked segregant analysis based on whole-genome resequencing (BSA-Seq) was conducted on the parents of the F_2 population and two bulks of short-wing and normal-wing samples (Additional File 2: Table S28). A total of 36,912,642 SNPs and 816,684 InDels were screened for further BSA analysis, which identified three genomic regions on Chr3 exceeding the 99% confidence level threshold. Among these regions, one region (BSA3) was further screened for a delta index of 1, which partially overlapped with WL3 (Fig. 6e, Additional File 1: Fig. S24, Additional File 2: Table S29). Furthermore, using transcriptome data of flower tissues from the two parental materials, 30 differentially expressed genes (DEGs) were identified in the genomic regions of WL3 and BSA3 (Additional File 2: Table S30). One candidate gene (*Vfab3G363000*) with no expression in the flowers of VF8137 (Fig. 6f) was found to be a homolog of the *FAF*-like gene in *Arabidopsis thaliana*, which is known to regulate meristem size and flower development pathways [45]. Notably, there was also a significant difference in wing petal length between the mutant and wild type in the F_2 population (Fig. 6g). Compared with H3712 (normal wing petal), this gene in VF8137 (short wing petal) has two nonsynonymous SNPs and two InDels in the exonic region

(Fig. 6h), which probably results in protein structural variation (Additional File 1: Fig. S25) and zero expression in the flowers of VF8137.

Discussion

With advancements in sequencing technologies and genome assemblies, large-scale resequencing and pangenome projects have gained momentum in various crop species [46], including rice [47, 48], wheat [49], maize [50, 51], soybean [8], and chickpea [11, 12]. Faba bean is one of the most important legume crops, having a high yield potential comparable to that of soybean, as well as high nutritional value and biological nitrogen fixation capacity [3]. However, research on the genome of faba bean is still in its early stages, whereas pan-genome research remains unexplored. To explore the genomic diversity of faba bean and the genomic characteristics of specific germplasms in-depth, we constructed a high-quality genome assembly and annotation of the special germplasm VF8137 from China, which has unique morphological characteristics (Fig. 1a, Additional File 1: Fig. S1). Compared with previously released genome assemblies of faba bean [13], the genome assembly VF8137 constructed in this study exhibits a significant improvement in assembly integrity, as evidenced by higher values of contig N50 and N90 along with a reduced contig number (Additional File 2: Table 2). Additionally, more repetitive sequences and a greater number of rigorously screened genes were annotated in the genome assembly VF8137 (Additional File 2: Table S4, S5), which provides valuable genetic resources for comparative genomic and functional genomic studies of faba bean and other legumes. Furthermore, multiple lines of evidence revealed significant genetic differentiation between VF8137 and Hedin/2 or Tiffany. First, on the basis of the three faba genomes, we identified a large number of genome-wide SNPs, InDels, SVs and PAVs as well as unique genes in VF8137 (Fig. 2b, d, e; Additional File 2: Tables S13, S14). Second, by integrating with variation from the three genomes and an additional panel of 558 faba bean germplasms, VF8137 was classified into different genetic groups from Hedin/2 or Tiffany, which also corresponded to its geographic origin (Additional File 1: Fig. S9). Third, the upstream, downstream, and intron regions of the key gene *FT* for flowering period adaptability exhibited significant differences between VF8137 and Hedin/2 or Tiffany (Additional File 1: Fig. S12). All of these findings indicate the significant role of new genomes from different populations with remarkable genetic differences in crop domestication and improvement, exploration of germplasm resource diversity, and functional genomics research [52, 53].

Previous research has revealed a significant population genetic structure and geographical differentiation within faba beans, with European, African, and Asian accessions representing separate genetic groups [26, 54, 55]. Additionally, Chinese faba bean germplasms have been found to possess a unique genetic background compared with those from other regions [26, 54]. Using a total of 1,332,986 high-quality genome-wide SNPs, three genetic groups associated with geographic origin were identified from a set of 558 accessions of global faba bean germplasm (Fig. 4). Consistent with previous studies [26, 54, 55], a clear genetic divergence of Chinese and European faba bean germplasms has been confirmed with large-scale genomic diversity analyses. As anticipated, VF8137 from China and Hedin/2 from Europe were classified into group 1 and group 2, respectively (Additional File 1: Fig. S8). Additionally, a large number of SVs detected

between VF8137 and Hedin/2 or Tiffany were also found to be located within the selective sweep regions between group 1 and group 2 (Additional File 2: Table S20), indicating a strong correlation between these SVs and the observed population differentiation in faba beans. Furthermore, the ecological differentiation of spring and autumn sowing areas in northern and southern China has been repeatedly verified [54], which is similar to the differentiation of faba bean germplasms in northern and southern Europe [55]. Moreover, on the basis of 3 years of phenotypic observations of 558 accessions of global faba bean germplasm in different ecological regions, it was found that north–south ecological differentiation is also reflected by important agronomic traits, such as flowering time, plant height, and seed size (Fig. 5, Additional File 1: Fig. S11, S13, S15). This phenomenon of north–south differentiation has also been observed in other legume crop species, such as common bean and rice bean [9, 56], which is mainly related to photoperiodic flowering adaptation. However, due to the lack of ancestral genotypes, it is difficult to determine whether the population differentiation of faba bean originated from different ancestors or was due to adaptation after domestication.

Genome-wide association studies (GWAS) have been successfully applied in staple food crops and legume crops for the gene mining of important traits via the use of diverse germplasm panels [9, 48, 56, 57]. Previous studies have identified QTLs for several agronomic traits in faba bean based on biparental populations, MAGIC populations, or a limited number of accessions [55, 58, 59]. Owing to the small number of markers and limited genetic recombination, the genetic basis of important agronomic traits has still not been fully described. Here, a GWAS was conducted on 558 accessions of global faba bean germplasm in combination with phenotyping data obtained from seven experiments at three locations over 3 years and a total of 1,332,986 high-quality genome-wide SNPs to identify 545 SNPs involving 222 genes significantly associated with 10 agronomic traits (Additional File 2: Table S23). Eight of these genes were repeatedly detected across multiple environments or years, indicating that these candidate genes are strongly associated with plant height, pod length, seed length, and seed weight (Additional File 2: Table S23). These yield-related loci and genes will lay a good foundation for breeding and improvement of high-yield faba beans.

As a key agronomic trait, flowering time has an important impact on crop adaptation and yield, and determining the mechanism of flowering time is highly important for molecular crop breeding [60–67]. Photoperiodic flowering regulation has been well studied, and complex signaling networks have been identified in *Arabidopsis thaliana*, rice, soybean, and many other crop species [63–66]. Although the photoperiodic regulatory pathway of soybean is different from that of *Arabidopsis thaliana* and rice [67, 68], the *FT* gene tends to be conserved in most plants, functioning as a molecular trigger of flowering downstream of photoperiodic regulatory networks[30, 69, 70]. In this study, we identified a QTL significantly associated with flowering time harboring an ortholog of the *FT* gene (*Vfab5G146000*). This gene contains 11 SNPs in the intron region, forming two major haplotypes (Hap 1 and Hap 2) in all the samples. Group 1, which is mainly distributed in southern China, is dominated by Hap 1, whereas group 3, which represents northern China, is dominated by Hap 2. In addition, the two haplotypes also exhibited significant differences in flowering time. Interestingly, this gene was also found to be located in the region selected for North–South population differentiation. The above

results revealed the molecular mechanism of North–South differentiation in faba beans to a certain extent and could be exploited to develop wide adaptability of faba beans in China.

Unlike other mainly self-pollinated legume crops, faba bean has a high outcrossing rate because of its well-developed nectary, which can attract pollinators such as bees [19, 22, 71]. However, a high outcrossing rate affects the consistency and stability of faba bean varieties as well as yield due to pollinator abundance and activity, restricting the development of the faba bean breeding process and the seed industry [19, 22, 71]. Considering the above problems as well as the difficulty in hybrid breeding of faba bean, VF8137, which possesses a specific short-wing petal germplasm along with a low outcrossing rate is almost an ideal parent for eliminating dependence on pollinators and avoiding yield fluctuations caused by pollinators [72], providing a new mode of self-breeding for faba bean. According to our field observations, compared with that of flowers with normal-wing petals, the rate at which bees visit flowers with short-wing petals is almost zero (Additional File 1: Fig. S1). It was speculated that the shortened wing length increased the difficulty of frontal flower visitation, thus reducing the outcrossing rate. In this study, by integrating genomic resources, quantitative trait locus (QTL) analysis, bulked segregant analysis (BSA), transcriptome sequencing (RNA-Seq), and haplotype analysis, we identified one candidate gene related to short-wing petals and elucidated its function through multiple functional annotations, sequence variations, expression differences, and protein structure variations (Fig. 6). Based on these results, short-wing petals will become a visible characteristic of varieties with low outcrossing rates in faba beans, and molecular markers of short-wing petal genes provide technical support for the efficient screening of varieties with low outcrossing rates in faba beans. The application of these results will not only promote the rapid fixation of excellent target traits and shorten the purification cycle of varieties but also greatly improve the breeding efficiency and accelerate the molecular breeding and development of the seed industry of faba beans.

Conclusions

In summary, the high-quality genome assembly of VF8137, genomic diversity of global faba bean germplasms, and floral and yield trait-related genetic loci in faba bean presented in this study not only reveals new insights into the genome evolution of Leguminosae and faba bean but also provides valuable genomic and genetic resources for accelerating the breeding and improvement of faba bean [7, 13].

Methods

Sampling and genome sequencing

The Chinese-specific faba bean germplasm VF8137, which has special short-wing petals, was purified by single-seed descent for five generations. Genomic DNA was extracted from young leaves of VF8137 plants using a DNeasy Plant Mini Kit (Qiagen, Germantown, CA, USA). A 15 kb library was constructed and sequenced using the PacBio Sequel II platform (Pacific Biosciences, Menlo Park, CA, USA), which generated a total of 490.54 Gb of CCS HiFi reads with an N50 size of 16.65 kb using ccs software (v3.0.0) (https://github.com/Pacific Biosciences/University/, –min-passes 3 –min-length 10,000 –max-length 1,000,000 –min-rq 0.99). In addition, Hi-C libraries were constructed using

fresh leaf tissue fixed in 1% formaldehyde to extract chromatin, which was subsequently digested using the DpnII restriction enzyme (New England Biolabs, USA). After quality control, the Hi-C libraries were sequenced on an Illumina NovaSeq Xten platform (Illumina, San Diego, CA, USA), and a total of 632.04 Gb of Hi-C data was obtained. Furthermore, a total of 652.06 Gb of Illumina sequence data were produced using the NovaSeq Xten platform (Illumina, San Diego, CA, USA). For RNA-seq, seven samples from different tissues, including roots, stems, leaves, buds, flowers, fresh pods, and fresh seeds, were collected at the flowering and pod-setting stages and stored at − 80 °C. Total RNA was isolated using the RNAprep Pure Plant Kit (TIANGEN, Beijing, China) to construct RNA-seq libraries, which were subsequently sequenced on the NovaSeq Xten platform (Illumina, San Diego, CA, USA). A total of 73.57 Gb paired-end reads were generated for the seven RNA-seq libraries (Additional File 2: Table S1).

Genome size estimation

The genome size of the 17-mer was estimated by K-mer frequency analysis with Jellyfish (v2.2.6) [73] using a total of 445.12 Gb of Illumina sequence data (~ 38X). Based on the K-mer distribution, the genome size, heterozygosity ratio, and percentage of repeat sequences (Additional File 1: Fig. S2) were estimated using a GCE (ftp://ftp.genomics.org.cn/pub/gce).

Genome assembly and assessment

A total of 490.54 Gb (~ 41.57 ×) of HiFi reads were assembled into contigs using hifiasm [74, 75] with the parameters "-l 2 -k 51 -w 51". The contigs were then anchored to chromosomes with the aid of Hi-C data (Additional File 2: Table S3). Hi-C reads were aligned to contigs using HICUP (v0.7.3) [76], yielding an alignment BAM file. The contigs were subsequently clustered using the ALLHiC [77] algorithm (ALLHiC_partition -e GATC -k 7). Finally, the assembled genome was manually corrected with Juicebox Assembly Tools (JBAT) (v1.11.08) [78].

In addition, several methods have been applied to assess the genome assembly quality of faba beans. First, the filtered Illumina reads and PacBio HiFi reads were realigned to the genome assembly VF8137 for mapping statistics (Additional File 2: Table S7) using BWA-MEM (v0.7.15) [79] and Minimap2 (v2.1) [80], respectively. Second, benchmarking universal single-copy orthologs (BUSCO) (v5.0.0) [81] were used to assess the genome completeness of the three genome assemblies of faba beans based on the Embryophyta Plant database (Additional File 2: Table S8). Third, Merqury (v1.3) analysis was performed to evaluate the K-mer completeness and heterozygosity of the three faba bean genome assemblies (Additional File 2: Table S9) [82]. Finally, the LTR assembly index (LAI) was used to determine the completeness of the full-length long terminal repeat retrotransposon via a standard process [83–86] (Additional File 2: Table S10).

Genome annotation

Repeat element prediction was conducted with de novo and homology-based approaches (Additional File 2: Table S4). For de novo annotation, LTR_FINDER_parallel (v1.0.7) [86], RepeatScout (http://www.repeatmasker.org/), RepeatModeler (http://www.repeatmasker.org/RepeatModeler.html) and RepeatMasker (http://repeatmasker.org/)

were used to construct a VF8137-specific repeat library by identifying repeat families and masking repetitive sequences in the genome assembly VF8137. For homolog evidence, repetitive sequences were predicted using RepeatProteinMask (http://www.repeatmasker.org/) based on alignment searches against the RepBase database (http://www.girinst.org/repbase).

A comprehensive strategy combining homology-based, transcriptome-based, and ab initio prediction was employed to annotate the protein-coding genes. The protein sequences of *Pisum sativum*, *Cicer arietinum*, *Lotus japonicus*, *Cajanus cajan*, *Glycine max*, and *Medicago truncatula* were aligned to the genome assembly VF8137 using BLAST (E-value < $1e^{-5}$) [87], and the hits were joined by Solar software [88]. The exact gene structure of the corresponding genomic regions for each WUblast hit was predicted by GeneWise [89] and denoted as a homology-based prediction gene set (Homoset). The RNA-seq data were mapped to the genome assembly VF8137 using TopHat (v2.0.8) [90] and Cufflinks (v2.1.1) [91] and then used to assemble the transcripts into gene models (Cufflinks-set). In addition, the RNA-seq data were also assembled by Trinity (v2.6.6) [92, 93], and gene models from Trinity-assembled transcripts were predicted by PASA [94] to obtain the PASA Trinity set. Gene models created by PASA were used as training data to predict coding regions in the repeat-masked genome for ab initio gene prediction programs, including Augustus (v2.5.5) [95], Genscan (v1.0) [96], Geneid [97], GlimmerHMM (v3.0.1) [98], and SNAP [99]. Finally, EVidenceModeler (EVM) [100] was used to combine all the gene models from the Homoset, Cufflinks-set, PASA Trinity set, and ab initio programs into a nonredundant set of protein-coding genes (Additional File 2: Table S5).

Gene function annotation was performed using BLAST [87] (E-value < $1e^{-5}$) for homology-based searches against the NR, KEGG [101], and SwissProt databases [102], as well as InterProScan (v5.0) [103] (-f TSV -dp -gotermes -iprlookup) for potential functions against the InterPro and GO databases (Additional File 2: Table S6).

Genome evolution

To elucidate the genome evolution of *Vicia faba* and its closely related species, 27 plant genomes were selected for comparative analyses, including 21 species from Leguminosae and four representative species of core Eudicotyledonous plants (Additional File 2: Table S11) [104–134]. The genes of each species were filtered to retain the longest transcript for each gene and to exclude protein sequences of less than 30 amino acids. The protein sequence similarity of the 27 plant genomes was obtained through BLASTP (v2.5.0 +) [135] with an E value less than $1e^{-5}$. After that, the protein datasets of all the species were clustered into paralogous and orthologous groups using OrthoMCL [136] with an inflation parameter of 1.5 (http://orthomcl.org/orthomcl/).

The sequences of single-copy orthologous genes from the 27 plant genomes were aligned together using MUSCLE (v5) [137] software. A maximum-likelihood tree was subsequently constructed using RAxML (v8.0.19) [138] with 100 bootstrap replicates. Divergence times between different species were estimated under a relaxed clock model using the Mcmctree program in PAML (v4.9e) [139] (JC69 model, nsample = 1,000,000, burnin = 10,000). Six calibration points were obtained from the TimeTree database (http://www.timetree.org/), including the divergence times between *Arabidopsis*

thaliana and *Populus trichocarpa* (102–112 Mya), *Chamaecrista fasciculata* and *Faidherbia albida* (11–61 Mya), *V. faba* and *P. sativum* (11–31 Mya), *M. truncatula* and *Trifolium pretense* (10–37 Mya), *and Vigna radiata* and *Vigna angularis* (2–3 Mya).

To investigate the expansion and contraction of gene families, we used the likelihood model originally implemented in the software package CAFE (v3.1) [140] to infer gene gain and loss rates among the genomes of the most recent common ancestor (MRCA). The topology and branch length of the phylogenetic tree were taken into account to infer the significance of gene family size changes in each branch. KEGG enrichment analysis was performed for candidate genes from the specific gene families using KOBAS (v2.0) [141] with filtering criteria of $P < 0.05$ and FDR < 0.05.

To detect syntenic blocks for *V. faba*/*M. truncatula* and *V. faba*/*P. sativum*, the protein sequences in one genome were searched against another genome using BLASTP (v2.5.0 +) (*E*-value < $1e^{-5}$) to identify the orthology of the best hit. The results were then subjected to MCScanX [142] to determine syntenic blocks. In addition, 4DTv (fourfold degenerate synonymous sites of the third codon) and Ks were calculated for syntenic segments.

Variation calling and annotation

To determine the population genetic structure of the landrace and cultivars of faba bean, a set of 558 worldwide accessions of faba bean germplasm were genotyped using the Faba_bean_130K TNGS SNP genotyping platform [27] (Additional File 2: Table S15), including 410 accessions used in a previous study [143]. A total of 3151 Gb of clean data were obtained after removing adapters and low-quality 150 bp paired-end Illumina raw reads using fastp (v0.23.2) [144], and two accessions were excluded from further analyses because of low data volume. The clean reads were mapped to the reference genome of VF8137 using BWA-MEM (v0.7.15) [79] with the command "mem -t 4 -k 32 -M". Duplicated reads were removed using SAMtools (v0.1.1) [145] to reduce mismatches generated by PCR amplification. Variant calling was performed using the Sentieon DNAseq software package [146], and identified variants were filtered using VCFtools (v0.1.15) [147] with the following criteria: (1) only two alleles, (2) missing rate < 0.2, (3) minor allele individuals ≥ 5.0, and (4) minor allele frequency ≤ 0.05. Finally, a total of 1,332,986 SNPs were retained for annotation using ANNOVAR [148], with an average SNP detection rate of 91% for the 558 accessions of faba bean germplasm (Additional File 2: Table S15 and S16).

Population genetic analyses of 558 faba bean germplasm accessions

To investigate the population variation pattern of 558 accessions of faba bean germplasm, we first estimated the population genetic structure using the program ADMIXTURE (v1.23) [149] with K values ranging from 2 to 10. The best clustering was determined by the *K* values with the maximized marginal likelihood and the smallest CV error. Individuals whose *q* value of the primary genetic component was less than 60% were identified as having admixture. In addition, an individual-based neighbor-joining (NJ) tree was constructed based on the p-distance by the software TreeBest (v1.9.2) [150] with 1000 bootstrap replications. In addition, principal component analysis (PCA) was conducted using GCTA (v1.93.2 beta) [151] to obtain the genetic relationship matrix (GRM) with

the parameter "–make-grm," and the top three principal components were estimated with the parameter "–pca3."

Nucleotide diversity (π) and the fixation index (F_{ST}) between populations were investigated using VCFtools (v0.1.15) [147] with a 20 kb window and a step size of 10 kb based on the best clustering result of population genetic structure. The threshold was set at the first 5% of F_{ST} value in the selective sweep identification.

The squared correlation coefficient (r^2) between pairwise SNPs and the pattern of linkage disequilibrium (LD) of different genetic groups in faba bean was estimated using the PopLDdecay [152] pipeline with the following parameters: –MaxDist 500 –MAF 0.05 –Miss 0.2.

Genome-wide association study

Three-year and three-location phenotyping of 10 agronomic traits was conducted on 558 accessions of faba bean germplasm from a diverse geographic origins (Additional File 2: Tables S14 and S20). Ten seeds of each accession were grown in rows 40 cm apart in the fields of Qinghai (N 36.72°, E 101.75°), Hebei (N 41.68°, E 115.66°), and Kunming (N 40.17°, E 115.24°) in China from 2019 to 2021. The plants were sown in April and harvested in September every year at the Qinghai and Hebei sites in the spring-sowing area, whereas they were grown in October each year and harvested in April of the next year at the Yunnan site in the autumn-sowing area. Five plants of each accession were surveyed for 10 agronomic traits based on the descriptions and methods detailed in a previous study [153]. Briefly, flowering time was recorded as days from the sowing date to the date of first flowering. The dry pod length and width were measured with five individuals of each accession using a Vernier caliper during the maturity stage at each site. The seed traits, including seed size and 100-seed weight, were measured after harvest via an automatic seed counting and analysis instrument (Model SC-G, Hangzhou Wanshen Detection Technology Co., Ltd., Hangzhou, China, http://www.wseen.com/).

Statistical analyses, including ANOVA and Pearson correlation analysis, of the agronomic traits were performed and visualized using R (v3.5.3) (https://www.r-project.org/). In combination with the genotyping data, a genome-wide association study (GWAS) was performed for ten agronomic traits using EMMAX software [154]. The threshold of significance was estimated as $-\mathrm{Log}_{10}$ (1/Total # of SNPs) based on Bonferroni correction and 584,294 independent SNPs were used in the GWAS analysis, resulting in a threshold *P* value of 5.77 for significant association signals [155, 156]. The kinship matrix of pairwise genetic similarities calculated by EMMAX was used as the variance–covariance matrix of the random effects.

Pollination experiment

To investigate the effect of variation in wing petal length on the flower visitation rate of bees, three varieties of faba bean, namely, VF8137 (short wing petal), GF45 (normal wing petal with black spot), and TF29 (normal wing petal without black spot), were selected for bee pollination experiments in a screen house in Zhangjiakou (N41.67°, E115.65°), Hebei Province, China, in 2023. Twenty plants per row of each variety were grown, and this process was repeated three times. After all three varieties entered the flowering stage, the existing flowers were removed, and a box of bees with a population density of

8000–10,000 bees was placed in the screen house. Observations were made from 8:00 am to 18:00 pm using three video cameras (ORDRO HDR-AC5 PLUS) to record the number of bees visiting the flowers of the three varieties over 3 days [157]. Statistical analysis of bee visit frequency and time was performed using R (v3.5.3) (https://www.r-project.org/).

Genetic linkage map construction and QTL mapping

A biparental population was developed consisting of 167 F_2 individuals from a cross between VF8137 (short-wing petal) and H3712 (normal-wing petal), and 16 agronomic traits, including two qualitative traits and 14 quantitative traits, was developed surveyed in 2020 [44] (Additional File 2: Table S24). Correlation analysis of different traits was performed using SPSS version 16.0 (SPSS Inc., Chicago, IL, USA), and four traits that were strongly correlated with other traits were excluded from further QTL analysis (Supplementary Data 8). The frequency distribution of wing petal length (WL) in the F_2 population was determined using R (v3.5.3) (https://www.r-project.org/).

Genotyping data obtained through targeted next-generation sequencing were mapped to the reference genome VF8137 using BWA-MEM (v0.7.15) [79]. SNP calling was performed using GATK 4 (https://gatk.broadinstitute.org) with default parameters, and the raw SNPs were filtered with a two-step filtration method as previously described. The ABH-format mapping data file was prepared for R/qtl [158] to select suitable markers and construct a genetic linkage map using the Perl script run_pipline.pl in Tassel (v 5.2.40) [159]. After that, SNPbinner [160] was used to calculate breakpoints and construct genotype bins (a -min ratio of 0.01 in the "crosspoints" command and a -min bin size of 5000 in the "bins" command). QTL analysis was conducted using R/qtl with the interval mapping method to identify QTLs with logarithm of odds (LOD) values higher than the threshold, which were estimated via 1000 permutations at $\alpha = 0.05$ and $\alpha = 0.01$ for each trait using R/qtl (Additional File 2: Table S27). LinkageMapView [161] and CMplot (https://github.com/YinLiLin/CMplot) were used to visualize the results of the genetic map and QTL analysis.

BSA analysis

To investigate the genetic basis of WL, bulked segregant analysis based on whole-genome sequencing (BSA-Seq) was conducted on the parents of the F_2 population and two pools of 30 individuals with short-wing petals and 30 individuals with normal-wing petals. Genomic DNA was extracted from dry leaves of the two parents and 60 individuals of the F_2 population using the DNeasy Plant Mini Kit (Qiagen, Germantown, CA, USA). DNA samples from the two pools were mixed with DNA samples from 30 individuals with different WLs. Four BSA libraries were constructed and sequenced on the NovaSeq Xten platform (Illumina, San Diego, CA, USA), yielding a total of 1135.11 Gb paired-end raw reads (Additional File 2: Table S28). Variant calling was performed using the Sentieon DNAseq software package [146], and two-step filtration was applied as described previously. The homozygous SNPs/InDels between two parents were extracted from the final VCF files, and the read depth information for the homozygous SNPs/InDels above in the offspring pools was obtained to calculate the SNP/InDel index [162]. The genotype of one parent was used as the reference, and the read number for

this parent's genotype or the other's genotype in the offspring pool was counted to calculate the ratio of the different read numbers to the total number as the SNP/InDel index of the site. Variations with SNP/InDel indices less than 0.3 in both pools were filtered out. Sliding window methods were used to determine the SNP/InDel index of the whole genome with a 10 Mb window size and a 1 Mb step size. The difference in the SNP/InDel indices of the two pools was calculated as the delta SNP/InDel index and visualized using R (v3.5.3) (https://www.r-project.org/).

Candidate gene analysis

To identify candidate genes associated with WL, six samples from the flower tissues of VF8137 and H3712 (three replicates for each genotype) were collected on the day before flowering and stored at − 80 °C. Total RNA was isolated using the RNAprep Pure Plant Kit (TIANGEN, Beijing, China) to construct RNA-seq libraries, which were subsequently sequenced on the NovaSeq Xten platform (Illumina, San Diego, CA, USA). The gene expression levels were calculated via the same methods as HISAT2 (v2.0.5) [163], StringTie (vl.3.3b) [164], and featureCounts (vl.0-p3) [165] as previously described. Then, differential expression analysis of the genes was performed using DESeq2 [166] with a standard workflow. The resulting *P* values were adjusted using the Benjamini-Hochberg (BH) method [167] to control the false discovery rate. Genes with an adjusted *P* value less than 0.05 were considered differentially expressed genes.

To verify the distribution of different genotypes of the candidate gene (*Vfab3G363000*) in the F_2 population and its relationship with WL, two primer pairs for *Vfab3G363000* were designed through Primer-BLAST (https://www.ncbi.nlm.nih.gov/tools/primer-blast/) (Additional File 2: Table S31) to obtain gene sequences from the two parents and 167 F_2 individuals via an Applied Biosystem™ 3730XL DNA sequencer (Thermo Fisher Scientific Inc., Massachusetts, USA). The sequences were assembled, aligned, and refined in turn with ContigExpress (Informax Inc., North Bethesda, MD), ClustalX (v1.83) [168], and BioEdit (v7.2.5) [169]. Haplotypes were generated by DnaSP (v6) [170] for further analysis. SWISS-MODEL [171] and PSIPRED (v4.0) [172, 173] were used to predict the three-dimensional structure and secondary structure of the two *Vfab3G363000 proteins* from VF8137 and H3712, respectively.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-025-03532-7.

Additional file 1. Fig. S1 Differences in the visiting frequency (a) and time (b) of bees for three varieties of faba beans with different wing petal lengths. *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$; ****, $P < 0.0001$. Fig. S2 Depth distribution of 17-mers in the genome of VF8137 according to K-mer analysis. Fig. S3 Genome features of two assemblies of faba bean including VF8137 (a) and Hedin/2 (b) and one assembly of *Pisum sativum* (c) and *Medicago truncatula* (d). Bin size of 5 Mb, 1Mb and 500 Kb were applied for two genome assemblies of faba bean including VF8137 and Hedin/2, *Pisum sativum* and *Medicago truncatula*, respectively. Fig. S4 Gene annotation pipeline for genome assembly VF8137 of Faba bean. Fig. S5 Expansion and contraction of gene families during the evolution of Leguminosae plants. Green and red indicate expanded and contracted gene families, respectively. Fig. S6 KEGG enrichment of the contracted gene family in Galegoids (a) and Millettoids (b). The circle color and size indicate padj values and the number of genes involved, respectively. Padj value less than 0.05 is considered statistically significant. Fig. S7 WGD estimation for *V. faba*, *P. sativum* and *M. truncatula* using 4dTv distance (a) and Ks analysis (b). Fig. S8 GO analysis of the unique genes identified in one of the three genome assemblies VF8137 (a), Hedin/2 (b), and Tiffany (c). The circle color and size indicate padj values and the number of genes involved, respectively. Padj value less than 0.05 is considered statistically significant. Fig. S9 PCA analysis using SNPs integrating with the variation from the three genome assemblies of VF8137, Hedin/2, and Tiffany and the 558 faba bean germplasms. Green, pink and blue represent Group 1, Group 2 and Group 3, respectively, which is corresponded to the three genetic groups obtained from ADMIXTURE results.

Fig. S10 GO analysis of the 3248 candidate selected genes between group 1 and group 2. Blue, red and orange represent Molecular function, Cellular component and Biological process, respectively. Fig. S11 Distribution of nine quantitative traits of 558 accessions of faba bean germplasm used in the seven phenotyping experiments (HB19, HB20, HB21, QH19, QH20, QH21, and YN19) at Hebei, Qinghai and Yunnan sites between 2019 and 2021. FT = Flowering time, PH = Plant height, DPL = Dry pod length, DPW = Dry pod width, PPP = Pods number per plant, HSW = 100-seed weight, SL = Seed length, SW= Seed width, ST = Seed thickness. Fig. S12 Collinearity analysis of *FT* gene in three faba bean genome assemblies of VF8137, Hedin/2, and Tiffany. Blue and orange boxes represent UTR and CDS of *FT*gene, whereas the red star and triangle represent the two unique photoperiod related regulation elements of the *FT* gene in VF8137. Fig. S13 Phenotypic differences among different genetic groups (a, b, c, d, e, f, g) and correlations (h) of PH in seven experiments. a, b, c, d, e, f, g, Pink, green, blue and purple represent Group 1, Group 2, Group 3 and Admixture, respectively. h, The darker the color intensity, the higher the correlation; whereas the weaker the color intensity, the lower the correlation.*, $P< 0.05$; **, $P<0.01$; ***,$P<0.001$; ****, $P<0.0001$. Fig. S14 GWAS and genetic analyses of candidate genes associated with PH. a and b, GWAS results of PH in HB19 (a) and HB20 (b). c, d, and e, Genotype distribution in three genetic groups and phenotypic differences of distinct haplotypes in *Vfab1G475000* (c), *Vfab1G475700* (d), and *Vfab1G475900* (e). Fig. S15 Phenotypic differences among different genetic groups (a, b, c, d, e, f, g) and correlations (h) of 100SW in seven experiments. a, b, c, d, e, f, g, Pink, green, blue and purple represent Group 1, Group 2, Group 3 and Admixture, respectively. h, The darker the color intensity, the higher the correlation; whereas the weaker the color intensity, the lower the correlation.*, $P< 0.05$; **, $P<0.01$; ***,$P<0.001$; ****, $P<0.0001$. Fig. S16 GWAS and genetic analyses of the candidate gene *Vfab1G141200* associated with seed size and seed weight. a, GWAS results of seed size and seed weight at different sites across years. b, LD block analysis for the SNPs significantly associated with 100SW. c, Genotype of the SNP significantly associated with 100SW. d, Phenotypic differences in distinct haplotypes of 100SW. e and f, Haplotype composition in different genetic groups of 558 accessions of faba bean germplasm. Fig. S17 GWAS and genetic analyses of the candidate gene *Vfab1G792100* associated with seed size and seed weight. a, GWAS results of seed size and seed weight at different sites across years. b, LD block analysis for the SNPs significantly associated with 100SW. c, Genotype of the SNP significantly associated with 100SW. d, Phenotypic differences in distinct haplotypes of 100SW. e and f, Haplotype composition in different genetic groups of 558 accessions of faba bean germplasm. Fig. S18 GWAS and genetic analyses of variations and genes associated with pod size. a, GWAS results of dry pod length (DPL) and dry pod width (DPW) across years. b, LD block analysis for the SNPs significantly associated with DPL. c, Genotype of the SNP significantly associated with DPL. d, Phenotypic differences in DPL among distinct haplotypes. e and f, Haplotype composition in different genetic groups of 558 accessions of faba bean germplasm. Fig. S19 GWAS analyses and genetic dissection of candidate genes associated with HC. a, SNPs significantly associated with HC according to GWAS. b, QQ plot of HC in GWAS analysis. c, LD block analysis for SNPs significantly associated with HC. d and e, Haplotype composition of *Vfab1G537500* (*PPO*) in different genetic groups of 558 accessions of faba bean germplasm. f and g, Phenotypic differences in HC among distinct haplotypes of *Vfab1G537500* (*PPO*). Fig. S20 Frequency distribution of 16 agronomic traits in 167 F_2 population of faba beans. The black and red arrows indicate the parents (VF8137 and H3712) of the F_2 population, respectively. HC = Hilum color, WT = Wing type, WL = Wing length, VL = Vexil length, FBN = First bud node, FPN = First pod node, BN = Branch number, EBN = Effective branch number, PPP = Pods per plant, PH = Plant height, SNPP = Seeds numer per plant, SWPP = Seeds weight per plant, 100SW = 100-Seed weight, SL = Seed length, SW = Seed width, ST = Seed thickness. Fig. S21 Genome distribution (a) and estimated recombination fractions (b) for 3369 bin markers in faba beans. A bin size of 1 Mb is applied for genetic marker density. Fig. S22 High consistency between the assembly of VF8137 and the genetic map. a, Chr1; b, Chr2; c, Chr3; d, Chr4; e, Chr5; f, Chr6. Fig. S23 LOD distribution across the genome for 12 agronomic traits in QTL analyses of faba bean. The red solid and broken lines indicate thresholds of 0.01 and 0.05, respectively. HC = Hilum color, WL = Wing length, VL = Vexil length, FBN = First bud node, BN = Branch number, PPP = Pods per plant, PH = Plant height, SWPP = Seeds weight per plant, 100SW = 100-Seed weight, SL = Seed length, SW = Seed width, ST = Seed thickness. Fig. S24 Delta index distribution across the genome for WL in the BSA analysis (a) and LOD score of WL across the genome in QTL mapping (b) of faba beans. a,The red and grey lines indicate confidence intervals of 99% and 95%, respectively. b, The red solid and broken lines indicate thresholds of 0.01 and 0.05, respectively. Fig. S25 Differences in three-dimensional structure and secondary structure prediction of FAF3 protein in H3712 (a and c) and VF8137 (b and d). The black arrows in a & b depict disparities in three-dimensional configuration, whereas the red and blue asterisks in c & d illustrate variations in secondary structure.

Additional file 2. Table S1. Summary of the data used for genome assembly and annotation. Table S2. Summary of the genome assembly VF8137. Table S3. Genome length of six chromosomes and anchored percentage in *Vicia faba*. Table S4. Summary of transposon elements in the VF8137 genome assembly. Table S5. Statistics of the predicted protein-coding genes of the VF8137 genome assembly. Table S6. Statistics of gene functional annotation of VF8137. Table S7. Illumina and PacBio Hifi sequence coverage depth and mapping rate for realigning to the genome assembly VF8137. Table S8. Assessment of genome completeness in faba genomes using BUSCO. Table S9. Merquury analysis of the VF8137 genome assembly of *Vicia faba*. Table S10 LTR-RT completeness assessment of the three genome assemblies of *Vicia faba* using LAI. Table S11. Information on the 37 representative sequenced plant genomes used in the comparative genomic analyses of *Vicia faba*. Table S12. KEGG enrichment of expanded and contracted gene families in the Galegoid and Millettoid clades of Leguminosae. Table S13. GO enrichment analysis with the unique genes found in only one of the three genome assemblies VF8137, Hedin/2, and Tiffany in pan-genome analysis. Table S14. Genome variations among the three genome assemblies of *Vicia faba*. Table S15. Information on 558 accessions of faba bean germplasm. Table S16. Annotation of SNPs identified in 558 accessions of faba bean germplasm. Table S17. Population differentiation across the genome (F_{ST}) between group 1 and group 2 of 558 accessions of faba bean germplasm. Table S18. Population differentiation across the genome (F_{ST}) between group 2 and group 3 of 558 accessions of faba bean germplasm. Table S19. Population

differentiation across the genome (F_{ST}) between group 1 and group 3 of 558 accessions of faba bean germplasm. Table S20. SV number present in the selective sweep region between group1 and group2. Table S21. Description of ten agronomic traits used in GWAS analysis of 558 accessions of faba bean germplasm. Table S22. Descriptive statistics of the nine quantitative agronomic traits used in seven phenotyping experiments of 558 accessions of faba bean germplasm. Table S23. GWAS results of ten agronomic traits for 558 accessions of faba bean germplasm. Table S24. Description of 16 agronomic traits used in the QTL analysis of faba beans. Table S25. Summary of the genetic linkage map of the F_2 population from VF8137 and H3712. Table S26. Correlation analysis of 16 agronomic traits in faba beans. Table S27. Results of the QTL analysis of 12 agronomic traits of the F_2 population from VF8137 and H3712. Table S28. Summary of data used for BSA analysis. Table S29. Exonic variations in VF8137 and H3712 within BSA3 (Chr3: 990,200,000-1,019,300,000). Table S30 Information on 30 DEGs located in WL3 or BSA3. Table S31 Information of two pair primers of *Vfab3G363000* (*FAF3*).

Acknowledgements
We thank Professor Song Ge for his valuable suggestions in manuscript revising. We also thank Dr. Yazhou Zhao and Dr. Hong Zhang for their help in the bee pollination experiment.

Peer review information
Eduard Akhunov and Wenjing She were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
T.Y., X.X.Z., Y.H.H., R.K.V., X.C., and Y.J.L. planned the project and designed the study. R.L., C.Q.H., D.G., M.W.L., X.X.Y. performed data analyses and drafted the manuscript. Z.M.D., W.W.H., D.X. X., Y.S.T., Y.M.D., F.M., S.K.A., and H.F.D. collected samples. Q.S., Z.H.W., H.T.Y., H.Y.Z., X.L.Z., and N.Z. performed experiments. X.Y., F.Y., A.Z., M.Y.L., C.C.T., L.B.W., X.L., and D.W. performed phenotyping. L.Y.C., V.G., C.Z.J., C.Z.D., C.X., A.C., and R.B. coordinated analysis of genome and transcriptome data, and BSA data preparation. T.Y., X.X.Z., Y.H.H., R.K.V., X.C., and Y.J.L. revised the manuscript. All authors have read and approved the final manuscript.

Funding
This work was supported by the Science and Technology Project of Yunnan Province (202202AE090003), China Agriculture Research System of MOF and MARA- Food Legumes (CARS-08), National Natural Science Foundation of China (32241042&32272134), Reginal Science Foundation of National Natural Science Foundation of China (32360466), the "JBGS" Project of Seed Industry Revitalization in Jiangsu Province (JBGS[2021]003, JBGS[2021]056), National Key R&D Program of China (2021YFD1200105), the Crop Germplasm Resources Protection (2130135), Agricultural Science and Technology Innovation Program (ASTIP) in CAAS, funding from Key Laboratory of Evaluation and Utilization for Special Crops Germplasm Resource in the Southwest Mountains, Ministry of Agriculture and Rural Affairs (Co-construction by the Ministry and Province), and Project of Sichuan Beans and Cereals Innovation Team of China Agriculture Research System (SCCXTD-2020–20), the Sichuan Provincial Science and Technology Support Breeding Project during the 14th 5-Year Plan (2021YFYZ0022). RKV thanks Grains Research & Development Corporation, Australia for supporting "Achieving improved genetic gain for yield in chickpea, faba bean, and lentil using genetic diversity" project (UMU UMU2403-009RTX) in Australia.

Data availability
All genome sequencing data and the genome assembly have been deposited at NCBI under the BioProject PRJNA1025910 [174]. The genome assembly and annotations of VF8137 are also available under figshare datasets [175]. Public genotyping data using the Faba_bean_130K TNGS SNP genotyping platform were from the BioProject PRJNA778650 at NCBI [176]. The custom scripts used in faba bean genome project have been deposited in Zenodo [177].

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
All of the authors herein affirm that there are no competing interests of any kind. Specifically, those authors affiliated with Smartgenomics Technology Institute located in Tianjin 301700, China, explicitly state that no competing interests exist in relation to either the research process or the publication of this manuscript.

Author details
[1]State Key Laboratory of Crop Gene Resources and Breeding, Institute of Crop Sciences, Chinese Academy of Agricultural Sciences, Haidian District, Beijing 100081, China. [2]Food Crops Research Institute, Yunnan Academy of Agricultural Sciences, Kunming, Yunnan 650205, China. [3]Smartgenomics Technology Institute, Tianjin 301700, China. [4]Institute of Industrial Crops, Jiangsu Academy of Agricultural Sciences, Nanjing, Jiangsu 210014, China. [5]State Agricultural Biotechnology Centre, Centre for Crop and Food Innovation, Food Futures Institute, Murdoch University, Murdoch, WA 6150, Australia. [6]State Key Laboratory of Plateau Ecology and Agriculture, Qinghai University, Xining, Qinghai 810016, China. [7]Qinghai Academy of Agricultural and Forestry Sciences, Qinghai University, Xining, Qinghai 810016, China. [8]Chongqing Academy of Agricultural Sciences, Chongqing 401329, China. [9]Crop Research Institute, Sichuan Academy of Agricultural Sciences, Chengdu, Sichuan 610066, China. [10]Zhangjiakou Academy of Agricultural Sciences, Zhangjiakou, Hebei 075032, China. [11]Qujing Academy of Agricultural Sciences, Qujingaq, Yunnan 655000, China. [12]Dali Academy of Agricultural Sciences, Dali, Yunnan 671005, China. [13]International Center for Agricultural Researchin the, Dry Areas (ICARDA), Beirut 1108-2010,

Lebanon. [14]Jiangsu Yanjiang Institute of Agricultural Sciences, Nantong, Jiangsu 226541, China. [15]Yuxi Academy of Agricultural Sciences, Yuxi, Yunnan 653100, China. [16]Institute of Crop Germplasm Resources, Shandong Academy of Agricultural Sciences/Shandong Provincial Key Laboratory of Crop Genetic Improvement, Ecology and Physiology, Jinan, Shandong 250100, China.

Received: 2 July 2024 Accepted: 6 March 2025
Published online: 17 March 2025

References

1. Mayer Labba IC, Frøkiær H, Sandberg AS. Nutritional and antinutritional composition of fava bean (*Vicia faba* L., var. minor) cultivars. Food Res Int. 2021;140:110038.
2. Tiwari BK, Singh N (eds). Pulse chemistry and technology. Cambridge: RSH Publishing; 2012. p. 34–133.
3. Zong XX, Yang T, Liu R. Faba bean (*Vicia faba* L.) breeding. In advances in plant breeding strategies: legumes: volume 7. Al-Khayri JM, Jain SM, Johnson DV. (eds). Cambridge: Springer International Publishing; 2019. p. 245–86.
4. Karkanis A, Ntatsi G, Lepse L, Fernández JA, Vågen IM, Rewald B, Alsiņa I, Kronberga A, Balliu A, Olle M, et al. Faba bean cultivation - revealing novel managing practices for more sustainable and competitive european cropping systems. Front Plant Sci. 2018;9:1115.
5. Mínguez MI, Rubiales D. Faba bean. In crop physiology case histories for major crops, Sadras VO, Calderini DF. (eds.). London: Academic Press; 2021. p. 452–81.
6. Adhikari KN, Khazaei H, Ghaouti L, Maalouf F, Vandenberg A, Link W, O'Sullivan DM. Conventional and molecular breeding tools for accelerating genetic gain in faba bean (*Vicia Faba* L.). Front Plant Sci. 2021;12:744259.
7. Maalouf F, Hu JG, O'Sullivan DM, Zong XX, Hamwieh A, Kumar S, Baum M. Breeding and genomics status in faba bean (*Vicia faba*). Plant Breeding. 2019;138:465–73.
8. Liu Y, Du H, Li P, Shen Y, Peng H, Liu S, Zhou GA, Zhang H, Liu Z, Shi M, et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182:162-76.e113.
9. Wu J, Wang L, Fu J, Chen J, Wei S, Zhang S, Zhang J, Tang Y, Chen M, Zhu J, et al. Resequencing of 683 common bean genotypes identifies yield component trait associations across a north-south cline. Nat Genet. 2020;52:118–25.
10. Yang T, Liu R, Luo Y, Hu S, Wang D, Wang C, Pandey MK, Ge S, Xu Q, Li N, et al. Improved pea reference genome and pan-genome highlight genomic features and evolutionary characteristics. Nat Genet. 2022;54:1553–63.
11. Khan AW, Garg V, Sun S, Gupta S, Dudchenko O, Roorkiwal M, Chitikineni A, Bayer PE, Shi C, Upadhyaya HD, et al. Cicer super-pangenome provides insights into species evolution and agronomic trait loci for crop improvement in chickpea. Nat Genet. 2024;56:1225–34.
12. Varshney RK, Roorkiwal M, Sun S, Bajaj P, Chitikineni A, Thudi M, Singh NP, Du X, Upadhyaya HD, Khan AW, et al. A chickpea genetic variation map based on the sequencing of 3,366 genomes. Nature. 2021;599:622–7.
13. Jayakodi M, Golicz AA, Kreplak J, Fechete LI, Angra D, Bednář P, Bornhofen E, Zhang H, Boussageon R, Kaur S, et al. The giant diploid faba genome unlocks variation in a global protein crop. Nature. 2023;615:652–9.
14. Caracuta V, Weinstein-Evron M, Kaufman D, Yeshurun R, Silvent J, Boaretto E. 14,000-year-old seeds indicate the levantine origin of the lost progenitor of faba bean. Sci Rep. 2016;6: 37399.
15. Khazaei H, O'Sullivan DM, Stoddard FL, Adhikari KN, Paull JG, Schulman AH, Andersen SU, Vandenberg A. Recent advances in faba bean genetic and genomic tools for crop improvement. Legume Sci. 2021;3: e75.
16. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. Plant Biotechnol J. 2016;14:1099–105.
17. Golicz AA, Bayer PE, Bhalla PL, Batley J, Edwards D. Pangenomics comes of age: from bacteria to plant and animal applications. Trends Genet. 2020;36:132–45.
18. Aguilar-Benitez D, Casimiro-Soriguer I, Ferrandiz C, Torres AM. Study and QTL mapping of reproductive and morphological traits implicated in the autofertility of faba bean. BMC Plant Biol. 2022;22:175.
19. Suso MJ, Moreno MT, Melchinger AE. Variation in outcrossing rate and genetic structure on six cultivars of *Vicia faba* L. as affected by geographic location and year. Plant Breeding. 1999;118:347–50.
20. Link W. Autofertility and rate of cross-fertilization: crucial characters for breeding synthetic varieties in faba beans (*Vicia faba* L.). Theoret Appl Genet. 1990;79:713–7.
21. Barrett SCH, Eckert C. Variation and evolution of mating systems in seed plants. In Biological approaches and evolutionary trends in plants, Kawano S. (ed.). New York: Academic Press, : 1990. p. 229–54.
22. Suso MJ, Pierre J, Moreno MTM, Esnault R, Jl Guen. Variation in outcrossing levels in faba bean cultivars: role of ecological factors. J Agr Sci. 2001;136:399–405.
23. Yu HT, Wang LP, Yang F, Lv MY, Zong XX, Yang T, Hu CQ, Yang X, Wang YB, He YH. Identification and utilization of short wing petal broad bean (*Vicia faba* L.) germplasm. J Plant Genet Resour. 2019;20:1334–9.
24. Tang H, Krishnakumar V, Bidwell S, Rosen B, Chan A, Zhou S, Gentzbittel L, Childs KL, Yandell M, Gundlach H, et al. An improved genome release (version Mt4.0) for the model legume *Medicago truncatula*. BMC Genomics. 2014;15:312.
25. Cannon SB, McKain MR, Harkess A, Nelson MN, Dash S, Deyholos MK, Peng Y, Joyce B, Stewart CN Jr, Rolf M, et al. Multiple polyploidy events in the early radiation of nodulating and nonnodulating legumes. Mol Biol Evol. 2015;32:193–210.
26. Zong XX, Liu X, Guan JP, Wang SM, Liu QJ, Paull JG, Redden R. Molecular variation among Chinese and global winter faba bean germplasm. Theor Appl Genet. 2009;118:971–8.
27. Wang CY, Liu R, Liu Y, Hou W, Wang X, Miao Y, He Y, Ma Y, Li G, Wang D, et al. Development and application of the Faba_bean_130K targeted next-generation sequencing SNP genotyping platform based on transcriptome sequencing. Theor Appl Genet. 2021;134:3195–207.

28. Hecht V, Laurie RE, Vander Schoor JK, Ridge S, Knowles CL, Liew LC, Sussmilch FC, Murfet IC, Macknight RC, Weller JL. The pea *GIGAS* gene is a *FLOWERING LOCUS T* homolog necessary for graft-transmissible specification of flowering but not for responsiveness to photoperiod. Plant Cell. 2011;23:147–61.
29. Samanfar B, Molnar SJ, Charette M, Schoenrock A, Dehne F, Golshani A, Belzile F, Cober ER. Mapping and identification of a potential candidate gene for a novel maturity locus, *E10*, in soybean. Theor Appl Genet. 2017;130:377–90.
30. Wickland DP, Hanzawa Y. The *FLOWERING LOCUS T/TERMINAL FLOWER 1* gene family: functional evolution and molecular mechanisms. Mol Plant. 2015;8:983–97.
31. Le L, Guo W, Du D, Zhang X, Wang W, Yu J, Wang H, Qiao H, Zhang C, Pu L. A spatiotemporal transcriptomic network dynamically modulates stalk development in maize. Plant Biotechnol J. 2022;20:2313–31.
32. Zheng Y, Zhang S, Luo Y, Li F, Tan J, Wang B, Zhao Z, Lin H, Zhang T, Liu J, et al. Rice *OsUBR7* modulates plant height by regulating histone H2B monoubiquitination and cell proliferation. Plant Commun. 2022;3: 100412.
33. Breia R, Conde A, Badim H, Fortes AM, Gerós H, Granell A. Plant SWEETs: from sugar transport to plant-pathogen interaction and more unexpected physiological roles. Plant Physiol. 2021;186:836–52.
34. Singh J, Das S, Jagadis Gupta K, Ranjan A, Foyer CH, Thakur JK. Physiological implications of SWEETs in plants and their potential applications in improving source-sink relationships for enhanced yield. Plant Biotechnol J. 2023;21:1528–41.
35. Zhu Y, Tian Y, Han S, Wang J, Liu Y, Yin J. Structure, evolution, and roles of SWEET proteins in growth and stress responses in plants. Int J Biol Macromol. 2024;263: 130441.
36. Zhang L, Ma J, Liu H, Yi Q, Wang Y, Xing J, Zhang P, Ji S, Li M, Li J, et al. SNARE proteins VAMP721 and VAMP722 mediate the post-Golgi trafficking required for auxin-mediated development in *Arabidopsis*. Plant J. 2021;108:426–40.
37. Abbo S, Pinhasi van-Oss R, Gopher A, Saranga Y, Ofner I, Peleg Z. Plant domestication versus crop evolution: a conceptual framework for cereals and grain legumes. Trends Plant Sci. 2014;19:351–60.
38. Li N, Xu R, Li Y. Molecular networks of seed size control in plants. Annual Rev Plant Biol. 2019;70:435–63.
39. Alvim Kamei CL, Boruc J, Vandepoele K, Van den Daele H, Maes S, Russinova E, Inzé D, De Veylder L. The *PRA1* gene family in *Arabidopsis*. Plant Physiol. 2008;147:1735–49.
40. Lawson T, Mayes S, Lycett GW. Plant Rabs and the role in fruit ripening. Biotechnol Genet Eng Rev. 2018;34:181–97.
41. Maillot P, Velt A, Rustenholz C, Butterlin G, Merdinoglu D, Duchêne E. Alternative splicing regulation appears to play a crucial role in grape berry development and is also potentially involved in adaptation responses to the environment. BMC Plant Biol. 2021;21:487.
42. Kawai Y, Ono E, Mizutani M. Evolution and diversity of the 2-oxoglutarate-dependent dioxygenase superfamily in plants. Plant J. 2014;78:328–43.
43. Zhao H, Kosma DK, Lü S. Functional role of Long-Chain Acyl-CoA synthetases in plant development and stress responses. Front Plant Sci. 2021;12: 640996.
44. Li MW, He YH, Liu R, Li G, Wang D, Ji YS, Yan X, Huang SX, Wang CY, Ma Y, et al. Construction of SNP genetic maps based on targeted next-generation sequencing and QTL mapping of vital agronomic traits in faba bean (*Vicia faba* L.). J Integr Agric. 2023;22:2648–59.
45. Wahl V, Brand LH, Guo YL, Schmid M. The FANTASTIC FOUR proteins influence shoot meristem size in *Arabidopsis thaliana*. BMC Plant Biol. 2010;10: 285.
46. Varshney RK, Sinha P, Singh VK, Kumar A, Zhang Q, Bennetzen JL. 5Gs for crop genetic improvement. Curr Opin Plant Biol. 2020;56:190–6.
47. Shang L, Li X, He H, Yuan Q, Song Y, Wei Z, Lin H, Hu M, Zhao F, Zhang C, et al. A super pan-genomic landscape of rice. Cell Res. 2022;32:878–96.
48. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature. 2018;557:43–9.
49. Walkowiak S, Gao L, Monat C, Haberer G, Kassa MT, Brinton J, Ramirez-Gonzalez RH, Kolodziej MC, Delorean E, Thambugala D, et al. Multiple wheat genomes reveal global variation in modern breeding. Nature. 2020;588:277–83.
50. Hufford MB, Seetharam AS, Woodhouse MR, Chougule KM, Ou S, Liu J, Ricci WA, Guo T, Olson A, Qiu Y, et al. De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. Science. 2021;373:655–62.
51. Wang B, Hou M, Shi J, Ku L, Song W, Li C, Ning Q, Li X, Li C, Zhao B, et al. De novo genome assembly and analyses of 12 founder inbred lines provide insights into maize heterosis. Nat Genet. 2023;55:312–23.
52. Yang X, Lee WP, Ye K, et al. One reference genome is not enough. Genome Biol. 2019;20:104.
53. Morneau D. Pan-genomes: moving beyond the reference. In nature milestones: genome sequencing, Koch L, Potenski C, Trenkmann M. (eds.). London: Springer Nature; 2021: S19.
54. Wang HF, Zong XX, Guan JP, Yang T, Sun XL, Ma Y, Redden R. Genetic diversity and relationship of global faba bean (*Vicia faba* L.) germplasm revealed by ISSR markers. Theor Appl Genet. 2012;124:789–97.
55. Skovbjerg CK, Angra D, Robertson-Shersby-Harvie T, Kreplak J, Keeble-Gagnère G, Kaur S, Ecke W, Windhorst A, Nielsen LK, Schiemann A, et al. Genetic analysis of global faba bean diversity, agronomic traits and selection signatures. Theor Appl Genet. 2023;136:114.
56. Guan J, Zhang J, Gong D, Zhang Z, Yu Y, Luo G, Somta P, Hu Z, Wang S, Yuan X, et al. Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding. Nat Commun. 2022;13:5707.
57. Zhou Z, Jiang Y, Wang Z, Gou Z, Lyu J, Li W, Yu Y, Shu L, Zhao Y, Ma Y, et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33:408–14.
58. Abou-Khater L, Maalouf F, Jighly A, Alsamman AM, Rubiales D, Rispail N, Hu J, Ma Y, Balech R, Hamwieh A, et al. Genomic regions associated with herbicide tolerance in a worldwide faba bean (*Vicia faba* L.) collection. Sci Rep. 2022;12:158.

59. Zhao N, Xue D, Miao Y, Wang Y, Zhou E, Zhou Y, Yao M, Gu C, Wang K, Li B, et al. Construction of a high-density genetic map for faba bean (*Vicia faba* L.) and quantitative trait loci mapping of seed-related traits. Front. Plant Sci. 2023;14:1201103.
60. Koo BH, Yoo SC, Park JW, Kwon CT, Lee BD, An G, Zhang Z, Li J, Li Z, Paek NC. Natural variation in *OsPRR37* regulates heading date and contributes to rice cultivation at a wide range of latitudes. Mol Plant. 2013;6:1877–88.
61. Kou K, Yang H, Li H, Fang C, Chen L, Yue L, Nan H, Kong L, Li X, Wang F, et al. A functionally divergent *SOC1* homolog improves soybean yield and latitudinal adaptation. Curr Biol. 2022;32:1728-42.e1726.
62. Wu T, Lu S, Cai Y, Xu X, Zhang L, Chen F, Jiang B, Zhang H, Sun S, Zhai H, et al. Molecular breeding for improvement of photothermal adaptability in soybean. Mol Breed. 2023;43:60.
63. Brambilla V, Gomez-Ariza J, Cerise M, Fornara F. The importance of being on time: regulatory networks controlling photoperiodic flowering in cereals. Front Plant Sci. 2017;8: 665.
64. Chen R, Deng Y, Ding Y, Guo J, Qiu J, Wang B, Wang C, Xie Y, Zhang Z, Chen J, et al. Rice functional genomics: decades' efforts and roads ahead. Sci China Life Sci. 2022;65:33–92.
65. Lin X, Liu B, Weller JL, Abe J, Kong F. Molecular mechanisms for the photoperiodic regulation of flowering in soybean. J Integr Plant Biol. 2021;63:981–94.
66. Maple R, Zhu P, Hepworth J, Wang JW, Dean C. Flowering time: from physiology, through genetics to mechanism. Plant Physiol. 2024;195:190–212.
67. Lin X, Dong L, Tang Y, Li H, Cheng Q, Li H, Zhang T, Ma L, Xiang H, Chen L, et al. Novel and multifaceted regulations of photoperiodic flowering by phytochrome A in soybean. Proc Nati Acad Sci. 2022;119: e2208708119.
68. Yuan J, Ott T, Hiltbrunner A. Phytochromes and flowering: legumes do it another way. Trends Plant Sci. 2023;28:379–81.
69. Kong F, Liu B, Xia Z, Sato S, Kim BM, Watanabe S, Yamada T, Tabata S, Kanazawa A, Harada K, Abe J. Two coordinately regulated homologs of *FLOWERING LOCUS T* are involved in the control of photoperiodic flowering in soybean. Plant Physiol. 2010;154:1220–31.
70. Liu L, Xuan L, Jiang Y, Yu H. Regulation by *FLOWERING LOCUS T* and *TERMINAL FLOWER 1* in flowering time and plant architecture. Small Struct. 2021;2:2000125.
71. Metz PLJ, Buiel AAM, van Norel A, Helsper JPFG. Rate and inheritance of cross-fertilization in faba bean (*Vicia faba* L.). Euphytica. 1992;66:127–33.
72. Lawes DA. The develpment of self-fertile field beans. Annual report for 1972, Welsh plant breeding station, Wales. 1973. p. 163–75.
73. Marçais G, Kingsford C. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. Bioinformatics. 2011;27:764–70.
74. Cheng H, Concepcion GT, Feng X, Zhang H, Li H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. Nat Methods. 2021;18:170–5.
75. Cheng H, Jarvis ED, Fedrigo O, Koepfli KP, Urban L, Gemmell NJ, Li H. Haplotype-resolved assembly of diploid genomes without parental data. Nat Biotechnol. 2022;40:1332–5.
76. Wingett S, Ewels P, Furlan-Magaril M, Nagano T, Schoenfelder S, Fraser P, Andrews S. HiCUP: pipeline for mapping and processing Hi-C data. F1000Res. 2015;4:1310.
77. Zhang J, Zhang X, Tang H, Zhang Q, Hua X, Ma X, Zhu F, Jones T, Zhu X, Bowers J, et al. Allele-defined genome of the autopolyploid sugarcane *Saccharum spontaneum* L. Nat Genet. 2018;50:1565–73.
78. Burton JN, Adey A, Patwardhan RP, Qiu R, Kitzman JO, Shendure J. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat Biotechnol. 2013;31:1119–25.
79. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.
80. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34:3094–100.
81. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31:3210–2.
82. Rhie A, Walenz BP, Koren S, Phillippy AM. Merqury: reference-free quality, completeness, and phasing assessment for genome assemblies. Genome Biol. 2020;21:245.
83. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinform. 2008;9:18.
84. Ou S, Chen J, Jiang N. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 2018;46: e126.
85. Ou S, Jiang N. LTR_retriever: a highly accurate and sensitive program for identification of long terminal repeat retrotransposons. Plant Physiol. 2017;176:1410–22.
86. Xu Z, Wang H. LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. 2007;35:W265–8.
87. She R, Chu JSC, Wang K, Pei J, Chen N. genBlastA: enabling BLAST to identify homologous gene sequences. Genome Res. 2009;19:143–9.
88. Yu XJ, Zheng HK, Wang J, Wang W, Su B. Detecting lineage-specific adaptive evolution of brain-expressed genes in human using rhesus macaque as outgroup. Genomics. 2006;88:745–51.
89. Birney E, Clamp M, Durbin R. Genewise and genomewise. Genome Res. 2004;14:988–95.
90. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14: R36.
91. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
92. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.

93. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, Couger MB, Eccles D, Li B, Lieber M, et al. De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat Protoc. 2013;8:1494–512.
94. Haas BJ, Delcher AL, Mount SM, Wortman JR, Smith RK Jr, Hannick LI, Maiti R, Ronning CM, Rusch DB, Town CD, et al. Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
95. Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. AUGUSTUS: ab initio prediction of alternative transcripts. Nucleic Acids Res. 2006;34:W435–9.
96. Burge C, Karlin S. Prediction of complete gene structures in human genomic DNA. J Mol Biol. 1997;268:78–94.
97. Guigó R. Assembling genes from predicted exons in linear time with dynamic programming. J Comput Biol. 1998;5:681–702.
98. Majoros WH, Pertea M, Salzberg SL. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics. 2004;20:2878–9.
99. Korf I. Gene finding in novel genomes. BMC Bioinform. 2004;5:59.
100. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, White O, Buell CR, Wortman JR. Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. 2008;9: R7.
101. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
102. Consortium U. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:D506–15.
103. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L, et al. InterPro: the integrative protein signature database. Nucleic Acids Res. 2009;37:D211–5.
104. Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, et al. Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature. 2000;408:823–6.
105. Kahlau S, Aspinall S, Gray JC, Bock R. Sequence of the tomato chloroplast DNA and evolutionary comparison of solanaceous plastid genomes. J Mol Evol. 2006;63:194–207.
106. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, et al. The genome of black cottonwood, *Populus trichocarpa* (Torr. and Gray). Science. 2006;313:1596–604.
107. Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature. 2007;449:463–7.
108. Ouyang S, Zhu W, Hamilton J, Lin H, Campbell M, Childs K, Thibaud-Nissen F, Malek RL, Lee Y, Zheng L, et al. The TIGR rice genome annotation resource: improvements and new features. Nucleic Acids Res. 2007;35:D883–7.
109. Sato S, Nakamura Y, Kaneko T, Asamizu E, Kato T, Nakao M, Sasamoto S, Watanabe A, Ono A, Kawashima K, et al. Genome structure of the legume, *Lotus japonicus*. DNA Res. 2008;15:227–39.
110. Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. The Sorghum bicolor genome and the diversification of grasses. Nature. 2009;457:551–6.
111. Schmutz J, Cannon SB, Schlueter J, Ma J, Mitros T, Nelson W, Hyten DL, Song Q, Thelen JJ, Cheng J, et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83.
112. Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, Estep M, Feng L, Vaughn JN, Grimwood J, et al. Reference genome sequence of the model plant Setaria. Nat Biotechnol. 2012;30:555–61.
113. Mayer KFX, Waugh R, Langridge P, Close TJ, Wise RP, Graner A, Matsumoto T, Sato K, Schulman A, Muehlbauer GJ, et al. A physical, genetic and functional sequence assembly of the barley genome. Nature. 2012;491:711–6.
114. Varshney RK, Chen W, Li Y, Bharti AK, Saxena RK, Schlueter JA, Donoghue MTA, Azam S, Fan G, Whaley AM, et al. Draft genome sequence of pigeonpea (*Cajanus cajan*), an orphan legume crop of resource-poor farmers. Nat Biotechnol. 2012;30:83–9.
115. Varshney RK, Song C, Saxena RK, Azam S, Yu S, Sharpe AG, Cannon S, Baek J, Rosen BD, Tar'an B, et al. Draft genome sequence of chickpea (*Cicer arietinum*) provides a resource for trait improvement. Nat Biotechnol. 2013;31:240–6.
116. Kang YJ, Kim SK, Kim MY, Lestari P, Kim KH, Ha BK, Jun TH, Hwang WJ, Lee T, Lee J, et al. Genome sequence of mungbean and insights into evolution within *Vigna* species. Nat Commun. 2014;5:5443.
117. Schmutz J, McClean PE, Mamidi S, Wu GA, Cannon SB, Grimwood J, Jenkins J, Shu S, Song Q, Chavarro C, et al. A reference genome for common bean and genome-wide analysis of dual domestications. Nat Genet. 2014;46:707–13.
118. De Vega JJ, Ayling S, Hegarty M, Kudrna D, Goicoechea JL, Ergon A, Rognli OA, Jones C, Swain M, Geurts R, et al. Red clover (*Trifolium pratense* L.) draft genome provides a platform for trait improvement. Sci Rep. 2015;5:17394.
119. Sakai H, Naito K, Ogiso-Tanaka E, Takahashi Y, Iseki K, Muto C, Satou K, Teruya K, Shiroma A, Shimoji M, et al. The power of single molecule real-time sequencing technology in the *de novo* assembly of a eukaryotic genome. Sci Rep. 2015;5: 16780.
120. Bertioli DJ, Cannon SB, Froenicke L, Huang G, Farmer AD, Cannon EKS, Liu X, Gao D, Clevenger J, Dash S, et al. The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. Nat Genet. 2016;48:438–46.
121. Gordon SP, Contreras-Moreira B, Woods DP, Marais DLD, Burgess D, Shu S, Stritt C, Roulin AC, Schackwitz W, Tyler L, et al. Extensive gene content variation in the *Brachypodium distachyon* pan-genome correlates with population structure. Nat Commun. 2017;8:2184.
122. Gupta S, Nawaz K, Parween S, Roy R, Sahu K, Pole AK, Khandal H, Srivastava R, Parida SK, Chattopadhyay D. Draft genome sequence of *Cicer reticulatum* L., the wild progenitor of chickpea provides a resource for agronomic trait improvement. DNA Res. 2017;24:1–10.
123. Hane JK, Ming Y, Kamphuis LG, Nelson MN, Garg G, Atkins CA, Bayer PE, Bravo A, Bringans S, Cannon S, et al. A comprehensive draft genome sequence for lupin (*Lupinus angustifolius*), an emerging health food: insights into plant-microbe interactions and legume evolution. Plant Biotechnol J. 2017;15:318–30.
124. Appels R, Eversole K, Feuillet C, Keller B, Rogers J, Stein N, Pozniak CJ, Choulet F, Distelfeld A, Poland J, et al. Shifting the limits in wheat research and breeding using a fully annotated reference genome. Science. 2018;361:eaar7191.

125. Griesmann M, Chang Y, Liu X, Song Y, Haberer G, Crook MB, Billault-Penneteau B, Lauressergues D, Keller J, Imanishi L, et al. Phylogenomics reveals multiple losses of nitrogen-fixing root nodule symbiosis. Science. 2018;361: eaat1743.
126. Pecrix Y, Staton SE, Sallet E, Lelandais-Briere C, Moreau S, Carrere S, Blein T, Jardinaud M-F, Latrasse D, Zouine M, et al. Whole-genome landscape of *Medicago truncatula* symbiotic genes. Nat Plants. 2018;4:1017–25.
127. Chang Y, Liu H, Liu M, Liao X, Sahu SK, Fu Y, Song B, Cheng S, Kariba R, Muthemba S, et al. The draft genomes of five agriculturally important African orphan crops. Gigascience. 2019;8: giy152.
128. Lonardi S, Muñoz-Amatriaín M, Liang Q, Shu S, Wanamaker SI, Lo S, Tanskanen J, Schulman AH, Zhu T, Luo MC, et al. The genome of cowpea (*Vigna unguiculata* [L.] Walp.). Plant J. 2019;98:767–82.
129. Xie M, Chung CYL, Li MW, Wong FL, Wang X, Liu A, Wang Z, Leung AKY, Wong TH, Tong SW, et al. A reference-grade wild soybean genome. Nat Commun. 2019;10:1216.
130. Yang N, Liu J, Gao Q, Gui S, Chen L, Yang L, Huang J, Deng T, Luo J, He L, et al. Genome assembly of a tropical maize inbred line provides insights into structural variation and crop improvement. Nat Genet. 2019;51:1052–9.
131. Chen ZJ, Sreedasyam A, Ando A, Song Q, De Santiago LM, Hulse-Kemp AM, Ding M, Ye W, Kirkbride RC, Jenkins J, et al. Genomic diversifications of five *Gossypium* allopolyploid species and their impact on cotton improvement. Nat Genet. 2020;52:525–33.
132. Hufnagel B, Marques A, Soriano A, Marques L, Divol F, Doumas P, Sallet E, Mancinotti D, Carrere S, Marande W, et al. High-quality genome sequence of white lupin provides insight into soil exploration and seed quality. Nat Commun. 2020;11:492.
133. Wang H, Sun S, Ge W, Zhao L, Hou B, Wang K, Lyu Z, Chen L, Xu S, Guo J, et al. Horizontal gene transfer of *Fhb7* from fungus underlies *Fusarium* head blight resistance in wheat. Science. 2020;368:eaba5435.
134. Kuang L, Shen Q, Chen L, Ye L, Yan T, Chen ZH, Waugh R, Li Q, Huang L, Cai S, et al. The genome and gene editing system of sea barleygrass provide a novel platform for cereal domestication and stress tolerance studies. Plant Commun. 2022;3: 100333.
135. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
136. Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89.
137. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
138. Stamatakis A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics. 2006;22:2688–90.
139. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol. 2007;24:1586–91.
140. Han MV, Thomas GW, Lugo-Martinez J, Hahn MW. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. Mol Biol Evol. 2013;30:1987–97.
141. Mao X, Cai T, Olyarchuk JG, Wei L. Automated genome annotation and pathway identification using the KEGG Orthology (KO) as a controlled vocabulary. Bioinformatics. 2005;21:3787–93.
142. Wang Y, Tang H, Debarry JD, Tan X, Li J, Wang X, Lee TH, Jin H, Marler B, Guo H, et al. MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. Nucleic Acids Res. 2012;40: e49.
143. Zhang H, Liu Y, Zong X, Teng C, Hou W, Li P, Du D. Genetic diversity of global faba bean germplasm resources based on the 130K TNGS genotyping platform. Agronomy. 2023;13: 811.
144. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. Bioinformatics. 2018;34:i884–90.
145. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9.
146. Weber J, Aldana R, Gallagher B, Edwards J. Sentieon DNA pipeline for variant detection - Software-only solution, over 20× faster than GATK 3.3 with identical results. Peer J Preprints. 2016;4: e1672.
147. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. The variant call format and VCFtools. Bioinformatics. 2011;27:2156–8.
148. Wang K, Li M, Hakonarson H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res. 2010;38:e164.
149. Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome Res. 2009;19:1655–64.
150. Vilella AJ, Severin J, Ureta-Vidal A, Heng L, Durbin R, Birney E. EnsemblCompara genetrees: complete, duplication-aware phylogenetic trees in vertebrates. Genome Res. 2009;19:327–35.
151. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet. 2011;88:76–82.
152. Zhang C, Dong SS, Xu JY, He WM, Yang TL. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35:1786–8.
153. Zong XX, Bao SY, Guan JP. Descriptors and data standard for faba bean (*Vicia faba* L.). Beijing: China Agriculture Press; 2006. p. 9–23.
154. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong SY, Freimer NB, Sabatti C, Eskin E. Variance component model to account for sample structure in genome-wide association studies. Nat Genet. 2010;42:348–54.
155. Duggal P, Gillanders EM, Holmes TN, Bailey-Wilson JE. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome wide association studies. BMC Genomics. 2008;9: 516.
156. Guan J, Zhang J, Gong D, et al. Genomic analyses of rice bean landraces reveal adaptation and yield related loci to accelerate breeding. Nature Communication. 2022;13:5707.
157. Li M, Sui Y, Wang X, Ma Z, Luo Y, Aluthwattha ST, McKey D, Pujol B, Chen J, Zhang L. High outcrossing rates in a self-compatible and highly aggregated host-generalist mistletoe. Mol Ecol. 2022;31:6489–504.
158. Broman KW, Wu H, Sen S, Churchill GA. R/qtl: QTL mapping in experimental crosses. Bioinformatics. 2003;19:889–90.
159. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES. TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics. 2007;23:2633–5.

160. Gonda I, Ashrafi H, Lyon DA, Strickler SR, Hulse-Kemp AM, Ma Q, Sun H, Stoffel K, Powell AF, Futrell S, et al. Sequencing-based bin map construction of a tomato mapping population, facilitating high-resolution quantitative trait loci detection. Plant Genome. 2019;12:1–14.
161. Ouellette LA, Reid RW, Blanchard SG, Brouwer CR. LinkageMapView—rendering high-resolution linkage and QTL maps. Bioinformatics. 2017;34:306–7.
162. Takagi H, Abe A, Yoshida K, Kosugi S, Natsume S, Mitsuoka C, Uemura A, Utsushi H, Tamiru M, Takuno S, et al. QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations. Plant J. 2013;74:174–83.
163. Kim D, Langmead B, Salzberg SL. HISAT: a fast spliced aligner with low memory requirements. Nat Methods. 2015;12:357–60.
164. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT. StringTie and Ballgown Nat Protoc. 2016;11:1650–67.
165. Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics. 2014;30:923–30.
166. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.
167. Benjamini Y, Hochberg Y. Controlling the false discovery rate. A practical and powerful approach to multiple testing. J R Stat Soc B. 2018;57:289–300.
168. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 1997;25:4876–82.
169. Hall TA. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. Nuclc Acids Symp Ser. 1999;41:95–8.
170. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A. DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol. 2017;34:3299–302.
171. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, Heer FT, de Beer TAP, Rempfer C, Bordoli L, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res. 2018;46:W296-303.
172. Buchan DWA, Jones DT. The PSIPRED protein analysis workbench: 20 years on. Nucleic Acids Res. 2019;47:W402–7.
173. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol. 1999;292:195–202.
174. Liu R, Hu CQ, Gao D, Li MW, Yuan XX, Chen LY, Shu Q, Wang ZH, Yang X, Dai ZM, et al. Genome sequencing and assembly of *Vicia faba* VF8137. NCBI. BioProject accession: PRJNA1025910. Genome Sequence Archive. 2023. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1025910.
175. Liu R. VF8137 genome sequence and annotation result. figshare. Dataset. 2025. https://doi.org/10.6084/m9.figshare.25632861.v1.
176. Li MW, He YH, Liu R, Li G, Wang D, Ji YS, Yan X, Huang SX, Wang CY, Ma Y, et al. Construction of SNP genetic maps based on targeted next-generation sequencing and QTL mapping of vital agronomic traits in faba bean (*Vicia faba* L.) NCBI. BioProject accession: PRJNA778650. The raw sequencing data of faba bean. 2021. https://www.ncbi.nlm.nih.gov/bioproject/PRJNA778650.
177. Liu R. Custom script for *Vicia faba* VF8137. Zenodo. 2023. https://zenodo.org/record/8423762.

Publisher's Note