Genome sequencing and population resequencing provide insights into the genetic basis of domestication and diversity of vegetable soybean

Na Liu1,*, Yongchao Niu2, Guwen Zhang1, Zhijuan Feng1, Yuanpeng Bo1, Jinmin Lian2, Bin Wang1 and Yaming Gong1,*

1State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products, Institute of Vegetables, Zhejiang Academy of Agricultural Sciences, Hangzhou, 310021, China
2Biozeron Shenzhen, Inc., Shenzhen, 518081, China
*Corresponding authors. E-mail: ln200811@163.com, gongym07@126.com

Abstract

Vegetable soybean is one of the most important vegetables in China, and the demand for this vegetable has markedly increased worldwide over the past two decades. Here, we present a high-quality de novo genome assembly of the vegetable soybean cultivar Zhenong 6 (ZN6), which is one of the most popular cultivars in China. The 20 pseudochromosomes cover 94.57% of the total 1.01 Gb assembly size, with contig N50 of 3.84 Mb and scaffold N50 of 48.41 Mb. A total of 55517 protein-coding genes were annotated. Approximately 54.85% of the assembled genome was annotated as repetitive sequences, with the most abundant long terminal repeat transposable elements. Comparative genomic and phylogenetic analyses with grain soybean Williams 82, six other Fabaceae species and Arabidopsis thaliana genomes highlight the difference of ZN6 with other species. Furthermore, we resequenced 60 vegetable soybean accessions. Alongside 103 previously resequenced wild soybean and 155 previously resequenced grain soybean accessions, we performed analyses of population structure and selective sweep of vegetable, grain, and wild soybean. They were clearly divided into three clades. We found 1112 and 1047 genes under selection in the vegetable soybean and grain soybean populations compared with the wild soybean population, respectively. Among them, we identified 134 selected genes shared between vegetable soybean and grain soybean populations. Additionally, we report four sucrose synthase genes, one sucrose-phosphate synthase gene, and four sugar transport genes as candidate genes related to important traits such as seed sweetness and seed size in vegetable soybean. This study provides essential genomic resources to promote evolutionary and functional genomics studies and genomically informed breeding for vegetable soybean.

Introduction

Vegetable soybean (Glycine max L.), called “Mao dou” in China, is harvested as a green and fully filled pod when the seeds are approximately 80% mature [1–3]. It is one of the most important vegetables in China, Japan, and other Asian countries [4]. China is the country with the highest production of vegetable soybean worldwide. Approximately 300 000 hm² of vegetable soybeans are produced in China each year [5]. In recent years, the demand for vegetable soybean as fresh and frozen vegetables has increased worldwide. They are rich in protein, soluble sugars, starch, dietary fiber, minerals, vitamins, and other phytochemicals, such as isoflavonoids, with anticancer and other health-promoting activities [4, 6]. With many nutraceutical benefits to humans, the demand for vegetable soybean may continue to increase in the future [7].

As a specialty vegetable, commercial vegetable soybean differs from grain soybean in seed composition and phenotypic appearance. Grain soybean is one of the most important crops providing plant oil and protein for food and animal feed [8]. Vegetable soybean is being selected for high-quality traits, such as larger seeds and pods, sweeter, better taste, and better color than grain soybean [9]. Vegetable soybeans are typically shorter than grain soybeans. As such, improving the overall quality, including appearance quality, nutritional quality and taste quality, as well as the yield of vegetable soybean, are the most important and ongoing objectives for vegetable soybean breeding. Understanding the associated genetic bases controlling important agricultural traits is vital to address the ongoing demand for improved vegetable soybean quality and yield [10, 11].

The first sequenced soybean genome was cv. Williams 82, a cultivar developed in America in the 1980s [12]. The availability of this high-quality genome opened a new chapter in soybean functional genomics research [13–15]. Previous studies have indicated that cultivated soybean was domesticated from its wild relative [Glycine
soja (Sieb. and Zucc.)] approximately 5000 years ago in temperate regions of China [16]. Intergenomic comparisons demonstrate that cultivated soybeans and wild soybeans exhibit extensive genetic diversity and among cultivated soybeans from various geographic areas [16–20]. Identification of genes contributing to the domestication process is vital to the continued improvement of crop species. During domestication, a number of important agronomical traits known as “domestication syndrome” were convergently selected [21, 22]. These traits typically include loss of seed shattering, reduced seed dormancy and dispersal, increased apical dominance and fruit or seed size, and changes in photoperiod sensitivity, flowering and maturation uniformity [21, 22]. To date, a number of domestication genes in soybean have been identified [23]. For instance, a NAC (NAM, ATAF1/2 and CUC2) transcription factor, SHAT1–5 (SHATTERING1–5), is involved in the loss of pod shattering [24] and CmHs1–1, encoding a calcineurin-like protein, controls hard seededness in domesticated soybean [25]. The classical stay-green G gene controlling seed dormancy [26] and the PRR (pseudoresponse regulator) gene Tof12 controlling flowering and maturity [27] were selected during soybean domestication. By using SSR markers, the genetic diversity of vegetable soybean accessions from China, Japan, and the USA was investigated, in which the germplasms from Mainland China were found to be more diverse than the others [1, 28]. However, the genetic relatedness between vegetable and grain soybeans, in particular the genes underlying their diversifications that were naturally or artificially selected, remains largely elusive. Currently, taking advantage of de novo genome assembly and whole-genome resequencing technologies shed light on vegetable soybean domestication, diversification and improvement.

Recently, 26 soybean accessions were selected for genome assembly, and a graph-based pangenome was constructed using 26 de novo assembled genomes and three reported genomes [29]. These accessions covered 3 wild soybeans, 9 landraces, and 14 cultivars [29]. However, none of these accessions was vegetable soybean. Given the divergent selection between vegetable and grain soybean, we predict considerable differences between the vegetable soybean genome and the grain soybean genome. To test this hypothesis, we sequenced the vegetable soybean cultivar Zhenong 6 (ZN6). ZN6 is one of the most popular vegetable soybean cultivars in China. This cultivar was bred by Zhejiang Academy of Agricultural Sciences and exhibits high quality and high yield capacity.

In this study, we completed a high-quality assembly of the vegetable soybean genome using PacBio, Illumina sequencing and Dovetail Hi-C technologies. Comparative genomic and phylogenetic analyses of ZN6, Williams 82 and other representative Fabaceae species were performed. Furthermore, we resequenced 60 vegetable soybean accessions. Together with previously reported data from 103 wild soybean and 155 grain soybean accessions, we constructed population genomic analysis of these three populations to identify selective regions of the genome and candidate genes related to vegetable soybean traits. This study provides a foundation for further genomic research on vegetable soybean and will facilitate elite cultivar improvement.

Results

De novo genome sequencing and assembly

To assemble a high-quality reference genome for vegetable soybean, the cultivar “Zhenong 6” (ZN6) was used for whole genome sequencing, which is one of the most representative and popular vegetable soybean cultivars in China. As shown in Fig. 1, ZN6 and Williams 82 differed significantly in plant architecture, plant height, the number and length of internodes, and pod and seed size. The pod and seed size and weight of ZN6 were markedly higher than those of Williams 82 (Fig. 1b–e). In total, we generated ∼138× PacBio long reads (138.89 Gb), ∼50× Dovetail Hi-C reads (50.04 Gb) and ∼84× Illumina paired-end reads (84.69 Gb) (Supplementary Table 1-2).

The PacBio long reads were first assembled de novo, generating a contig assembly with an N50 of 3.84 Mb. Next, the assembly was integrated with Dovetail Hi-C reads to orient and order contigs into chromosome-scale scaffolds, and approximately 94.57% of the 1.01 Gb final vegetable soybean assembly was assigned to 20 super-scaffolds (Fig. 2, Supplementary Fig. 1), and the proportion of repeat sequences was comparable to those present in the Williams82 assembly [22]. Transposable elements (TEs) occupied 53.39% of the assembled genome. Long terminal repeats (LTRs) were the most abundant elements, representing 38.70% of the genome size, including 28.27% Gypsy-like and 10.30% Copia-like LTR elements (Supplementary Table 4).

We combined different strategies to identify protein genes. A total of 55 517 protein-coding genes were identified in the ZN6 genome. The average gene length was 3772 bp, with an average of 5.13 exons per gene (Supplementary Table 5). Further functional annotation according to BLAST analysis with public databases estimated that 52 307 (94.22%) genes
had homology with known genes in at least one of the databases (Supplementary Table 6), and 35,014 (63.07%) genes could be annotated by all five databases (Supplementary Fig. 4). In addition, 3,759 noncoding RNAs (ncRNAs) were identified in the vegetable soybean genome, which included 240 microRNAs (miRNAs), 244 ribosomal RNAs (rRNAs), 1,322 transfer RNAs (tRNAs), and 1,953 small nuclear RNAs (snRNAs) (Supplementary Table 7).

### Comparative genomic and phylogenetic analyses

We clustered the protein-coding genes into gene families for ZN6, Williams 82, *Trifolium pretense* (T. pretense), *Arachis duranensis* (A. duranensis), *Cicer arietinum* (C. arietinum), *Medicago truncatula* (M. truncatula), *Phaseolus vulgaris* (P. vulgaris), *Vigna angularis* (V. angularis) and *Arabidopsis thaliana* (A. thaliana) (Supplementary Table 8, Supplementary Fig. 5). A total of 32,957 gene families were identified, of which 10,168 were shared by all nine species. Compared with the other eight plants, there were 562 specific gene families in the ZN6 assembly (Fig. 3a); among these ZN6-specific families, 306 genes were supported by transcriptome or interpro functional annotation. These ZN6-specific genes were significantly enriched in microtubule motor activity, microtubule-based movement, RNA processing, GTP binding, and intracellular-related gene ontology (GO) categories (Supplementary Table 9).

We constructed a phylogenomic tree using the nine plant genomes ZN6, Williams 82, *T. pretense*, *A. duranensis*, *C. arietinum*, *M. truncatula*, *P. vulgaris*, *V. angularis* and *A. thaliana* using 1,818 single-copy nuclear genes (Fig. 3b). The divergence time between ZN6 and Williams 82 was suggested to be approximately 0.2 million years ago (MYA). Compared with Williams 82, ZN6 showed less gene family expansion (197 vs. 337) and more gene family contraction (187 vs. 32) (Fig. 3b). ZN6 shows the expansion of genes related to 1,3-beta-D-glucan synthase activity, phospholipid transport, and sucrose synthase activity. Functional annotation of the contracted genes in ZN6...
was enriched in recognition of pollen and iron ion binding (Supplementary Table 10).

Additionally, 8982 one-to-one orthologous gene sets in the eight Fabaceae plants for positive selection gene (PSG) detection were identified using the bidirectional best hit (BBH) method. Fifty-four PSGs in ZN6 were detected [P < 0.05, likelihood ratio test (LRT)], which were significantly associated with “nucleic acid phosphodiester bond hydrolysis” (2 PSGs) and “nucleotide-sugar transmembrane transport” (2 PSGs) (Supplementary Table 11).

SNP and Indel calling of soybean populations
To understand how vegetable soybean is related to other types of soybean, a total of 318 soybean accessions were collected in this study. Sixty representative vegetable soybean accessions were sequenced in this study (17.77×) (Supplementary Table 12), and 155 grain soybean and 103 wild soybean (G. soja) accessions were from a previous study with a sequencing depth of 13.56× (Supplementary Table 13) [29]. All the sequencing reads were aligned to the ZH13 genome, and 3.99 million genome-wide SNPs were identified. These high-quality SNPs were mostly located in intergenic regions (71.73%) and intronic regions (11.19%) and were rarely located in coding sequences (3.45%) (Supplementary Table 14). In addition, 355.55 K indels were identified, of which 4856 indels were located in exonic regions.

Population structure analysis
For these 318 individuals, we characterized their population structure using neighbor-joining phylogenetic reconstruction, STRUCTURE, and principal component analysis (PCA) (Fig. 4a-c). STRUCTURE analysis was performed with different group numbers (k = 2 to 5) (Fig. 4c). All three analyses clearly divided all wild soybean (wild), grain soybeans (grain) and vegetable soybeans (vegetable) into three clades.

The linkage disequilibrium (LD; indicated by r²) decay levels of vegetable soybean and grain soybean are slower than those of wild soybean (Fig. 4d), indicating that they have experienced strong selective processes. The θπ results showed that the wild soybean population in this study has higher levels of polymorphism (θπ = 2.16-e03) than vegetable (θπ = 8.43-e06) and grain soybean populations (θπ = 1.24-e03). The FST value between wild and vegetable soybeans (FST = 0.40) was slightly greater than that between wild and grain soybeans (FST = 0.32). The population differentiation between vegetable and grain soybeans was the lowest in the three comparisons (FST = 0.23). The TreeMix model was used to detect significant introgression between vegetable and grain groups (Supplementary Fig. 6). We modeled migration events from 0 to 3, but no significant gene flow was observed. The population structure and gene flow results indicated that vegetable soybean and grain soybean may have undergone independent domestication processes.

Selective sweep
Vegetable soybeans are physiologically and morphologically distinct from grain soybean. For instance, their seeds are sweeter than grain soybean. In addition, vegetable soybean has been in a long-term breeding improvement of many traits, including consumer and processor-oriented traits, such as pod and seed size, pod shapes, color, flavor and so on [9, 30, 31]. Population genomic analyses were performed to detect sweep signals with the hypothesis that these regions may contribute to characteristic different morphologies and seed components between vegetable soybean and grain soybean during domestication from wild soybean. We analyzed selective sweeps by θπ ratio and XP-CLR using 3.99 million chromosome SNPs generated by this study. Compared with wild soybean populations, there were 1112 and 1047 candidate selected genes in vegetable soybean and grain soybean populations, respectively (Fig. 5 and Supplementary Table 15–17). Among them, only 134 selected genes were shared between vegetable and grain soybean populations (Supplementary Table 18).

The selected genes were implicated in traits related to the characteristics of vegetable soybean, such as seed sugar content and oil content. Specifically, we found that several sucrose synthase 5 (SSS) (SoyZH13_18G042801, SoyZH13_18G042802, SoyZH13_18G042803), sucrose synthase 7 (SoyZH13_20G077905) and sucrose-phosphate synthase 3 (SPS3) (SoyZH13_08G291400) were in the selected regions in vegetable soybean vs. wild soybean but not detected in grain soybean vs. wild soybean.

Additionally, four candidates for sugar transport were also detected in the selected regions of vegetable soybean between wild soybean but not in the grain soybean comparison. These genes included sugar transport...
Figure 3. Comparative analysis of the vegetable soybean genome. a Intersections of gene families between nine plant species (ZN6, Williams 82, T. pratense, Vitis vinifera, A. duranensis, C. arietinum, Medicago truncatula, P. vulgaris, Vigna angularis and A. thaliana). The figure was plotted by UpSetR (Conway J R, et al.), in which the rows represent the gene families, and the columns represent their intersections. b Gene family expansion and contraction in the 9 genomes. Numbers on the nodes of the phylogenetic tree represent divergence times. The numbers of expanded (red) and contracted (blue) gene families in each lineage after speciation are shown on the corresponding branch.

protein 5 (SoyZH13_04G095700), sugar transport protein 14 (SoyZH13_09G175600), sugar transporter SWEET6b (SoyZH13_19G009200) and sugar transporter SWEET16 (SoyZH13_19G218000).

Haplotype differentiation patterns of SS5, SPS3 and SWEET6b loci showed that the vegetable soybean accessions were separated from the grain soybean and wild soybean accessions (Fig. 6 and Supplementary Table 19),
suggesting their potential contributions to increasing seed sweetness taste. The results will be useful for the functional discovery of candidate genes controlling the sugar content of vegetable soybean.

We further investigated the haplotype differentiation patterns in the vegetable, grain, and wild soybeans of some known domesticated genes, including SHAT1–5, GmHs1–1, Tof12, and G genes that control pod shattering [24], seed hard-seededness [25], flowering and maturity [27], and seed dormancy [26] in domesticated soybean, respectively. The results showed that the haplotypes of SHAT1–5, GmHs1–1 and Tof12 between vegetable soybean and grain soybean were the same. However, haplotype differentiation patterns of the G gene, which is the classical stay-green gene responsible for the green seed coat and a domesticated gene responsible for seed dormancy in soybean [26, 32], showed that the vegetable soybean accessions were clearly separated from the grain soybean accessions (Supplementary Fig. 7). The results suggested that the G gene might have played an important role in the diversification of vegetable soybean and grain soybean during domestication and improvement processes.

**Discussion**

Soybean includes grain-use and vegetable-use types important for humans and animals [2, 33]. Grain soybean is one of the most economically crucial plant protein and oil crops worldwide and provides approximately 25% of the world’s protein and 56% of the oilseed for food and animal feed [8]. Vegetable soybean is an important legume vegetable, especially in China, Japan and other Asian countries [3]. Due to its superior appearance, nutrition, taste, and ease of cooking, the demand for vegetable soybean continues to grow worldwide [34–36]. However, research on vegetable soybean is far behind that on grain soybean [4, 7, 37].

The breeding targets of vegetable soybean are distinct from grain soybean. Hence, a high-quality genome assembly is crucial for vegetable soybean in studying evolution, functional genomics, and breeding [38]. As a number of reference genomes have been sequenced, genetic diversity among various populations and varieties has been revealed [39–44]. A number of structural variations (SVs) have been proven to be responsible for important agronomic traits [29, 45–48]. In the present study, by combining PacBio, Illumina, and Hi-C sequencing technology, we obtained a high-quality genome of the vegetable soybean cultivar “ZN6.” By comparing the genomes between ZN6 and Williams 82, a total of 2380503 SNPs and 531046 indels were detected in syntenic blocks between the ZN6 and Williams 82 genomes, respectively (Supplementary Table 20). In addition, these two genomes showed 17921 SVs, including translocations, inversions and presence variations. The largest translocation was located from 11.81 to 15.59 Mb and 15.6 to 20.3 Mb on chromosome 5 of ZN6, which was shown anchored to Williams 82 from 11.75 to 20.82 Mb on chromosome 5. This analysis helped us to obtain candidate genes related to agronomically important traits. For instance, Glyma.05G083000.1, encoding 3-ketoacyl-CoA synthase 19 (KCS19), was found to be located in...
Figure 6. Haplotype distributions for SS5, SPS3, and SWEET6b. Across the grain soybean, vegetable soybean, and wild soybean accessions, haplotype distributions were shown for SNPs within these candidate genes related to seed sugar content. Each row represents the SNPs of the candidate genes. Each column represents the samples of the three populations. Sky blue represents the homozygous reference alleles, light green represents the heterozygote, red represents the homozygous variant, and gray represents the missing data.

this region. KCS19 is in the same gene family as KCS11, which is implicated in fatty acid elongation [49]. These genes involved in lipid biosynthesis could also impact fatty acid accumulation [50]. The lipid content was found to be significantly lower in vegetable soybeans than in grain soybeans [3]. Therefore, KCS19 might be a candidate gene for controlling lipid accumulation in different types of soybean. Comparative genomics analysis identified many variations, which may provide valuable resources for further study on genome structure, evolution and candidate gene identification that are responsible for important traits [38,51,52].

The population structure and the Treemix results indicated that wild, grain, and vegetable soybean were clearly divided into three clades, and no significant gene flow was observed between vegetable soybean and grain soybean. In addition, the haplotype differentiation patterns of domestication gene G responsible for seed dormancy also showed the separation of vegetable soybean from grain soybean. Moreover, we found little overlap with genomic regions under selection in the vegetable soybean population and grain soybean population versus the wild soybean population. These results suggested that vegetable and grain soybean should be under differentially conscious and unconscious human selection during the processes of soybean domestication, diversification and improvement. This finding is consistent with the morphological and physiological distinctions between vegetable and grain soybean varieties. In the selected genes, we found that SS5 (SoyZH13_18G042801, SoyZH13_18G042802, SoyZH13_18G042803), SS7 (SoyZH13_20G077905), SPS3 (SoyZH13_08G291400), sugar transport protein S (SoyZH13_04G095700), sugar transport protein 14 (SoyZH13_09G175600), sugar transporter SWEET6b (SoyZH13_19G009200) and sugar transporter SWEET16 (SoyZH13_19G218000) were unique in the vegetable vs. wild soybean but not in grain soybean vs. wild soybean. One important difference between these two types of soybeans is that vegetable soybean seeds are much larger and sweeter than seeds of grain soybean [3]. The sucrose concentration in seeds reaches a peak at the seed rapid developmental stage in soybean [53]. Sucrose is the primary carbon source of developing seeds, and sucrose metabolism is very important for seed development [54–58]. A large number of sugar transporters and metabolic enzymes are involved in sugar accumulation and partitioning [59]. Sucrose synthase, sucrose-phosphate synthase and sugar transporter genes are the most important genes participating in sugar metabolism and transport [54,60]. An early study on wheat showed that the selection of sucrose synthase haplotypes contributed greatly to increasing kernel weight and grain yield over a century of wheat breeding worldwide [61]. Sucrose synthase is the first step to catalyze sucrose to biosynthesis of starch and is also considered as a biochemical marker of sink strength [62]. In addition, previous studies have indicated that SWEET proteins are crucial for sugar translocation to seeds and subsequently impact seed setting, development, and composition [63–66]. In a recent study, CISWEE10 and CITST2 (Tonoplast Sugar Transporter)-affected watermelon fruit sugar accumulation were under selection, which led to the derivation of modern sweet watermelon from nonsweet ancestors during evolution and domestication [67]. More sugar and biomass accumulated in modern watermelon fruit than in its wild ancestor. However, sucrose synthase genes were not under selection during watermelon domestication [67]. Recently, it was reported that simultaneous changes in seed size and sugar, protein and oil contents are driven by the domestication of GmSWEET10a and GmSWEET10b sugar transporters in...
grain soybean [68]. Thus, sucrose synthase and sugar transporter genes may be priority targets for quality and yield improvement in vegetable soybean.

In this study, we present a high-quality reference genome for vegetable soybean to complement other grain soybean assemblies and provide a foundation for functional genomics research on vegetable soybean. Moreover, population genetic analysis identified candidate genes that were differentially selected in vegetable soybean and grain soybean. These results will enhance future vegetable soybean breeding and research.

Materials and methods

Plant materials

The vegetable soybean cultivar ZN6 and grain soybean cultivar Williams 82 were grown in April in the field at Zhejiang Academy of Agricultural Sciences (Haining, Zhejiang, China) (E120.42, N30.44). The vegetable soybean cultivar ZN6 was sequenced. The vegetable soybean plants utilized for Illumina sequencing and PacBio sequencing were grown in the greenhouse at Zhejiang Academy of Agricultural Sciences (16-h light/8-h dark, 25°C day/22°C night). Young leaves of 3-week-old vegetable soybean plants were collected for PacBio and Illumina sequencing. Two-week-old seedlings were sampled for Hi-C sequencing.

Genome sequencing

The genomic DNA of vegetable soybean was extracted from young leaves using an improved CTAB method. A Dovetail Hi-C library preparation kit was used to prepare genomic DNA for Hi-C sequencing. Three genome libraries were prepared and sequenced according to the manufacturer’s instructions to construct the chromosome-scale assembly. The genome was sequenced using the PacBio Sequel platform and Illumina NovaSeq platform.

Genome assembling

The contigs of vegetable soybean (ZN6) were carried out using wtdbg2 (https://github.com/ruanjue/wtdbg2) with a minimum of 5 kb PacBio subreads, followed by consensus polishing with long reads and short reads using wtpoa-cns consenser, which was built into wtdbg2. Then, the data polishing was finished by pbmm2 (v0.12.0) and Arrow (v2.3.3). Clean Hi-C data were used to create chromosome-scale scaffolds from the draft assembly. Then, errors introduced into the assembly in the long reads were corrected by pillon (v1.22).

Genome annotation

We used Tandem Repeats Finder (TRF, version 4.0.7) to identify the tandem repetitive sequences. The interspersed repeats of the vegetable soybean genome were identified using de novo repeat identification and known repeat searches against existing databases. Then, several programs were used to identify repeat sequences in the genome. RepeatModeler (v1.0.8) and LTR_FINDER (v1.0.6) were used to predict repeat sequences in the assembly. Then, RepeatMasker (version 4.0.7) and the Repbase database (version 21) were used to identify TE repeats in the assembled genome.

Protein-coding genes were predicted using a combination of de novo, protein homology and transcriptome-based prediction approaches. Three ab initio gene prediction programs, GlimmerHMM (version 3.0.4), Augustus (version 3.2.1), and SNAP (version 2006–07–28), were used for the prediction of coding regions in the repeat-masked genome. The protein sequences downloaded from Phytozone and NCBI DataBase (G. max Williams 82, A. thaliana, T. pratense, Vitis vinifera, M. truncatula, C. arietinum, A. duranensis, P. vulgaris, V. angularis) were then aligned to the assembled genome using Genblasta (version 1.0.4). Then, we predicted the exact gene structure of the corresponding genomic regions on each Genblasta hit by GeneWise (version 2.4.1). By using stringtie (version 1.2.2), hisat2 (version 2.0.1), and TransDecoder (version 3.0.1), the RNA-seq sequences were mapped to the assembly. The mapped RNA-seq data were used to assemble the transcripts and identify the coding regions in the gene models. By using EvidenceModeler (EVM), all the gene modes were combined into a nonredundant set of gene structures. The generated gene models were finally refined with the Program to Assemble Spliced Alignments (PASA v2.3.3). We then used BLASTP (E-value 1e-05) against SwissProt and TrEMBL to analyze the functional annotations of the protein-coding genes. We used InterProScan (V5.30) to annotate the protein domains. The Gene Ontology (GO) terms for each gene were identified by applying Blast2GO to the nr protein database. Finally, we blasted the potential pathways for the genes against the KEGG databases (release 84.0), with an E-value cutoff of 1e-05.

For noncoding RNA annotation, we used RNAscan-SE (version 1.3.1) to predict the tRNAs and identify rRNAs by alignment to the rice and Arabidopsis template rRNA using BlastN (version 2.2.24) with an E-value of 1e-5. Then, we used INFERNAL (version 1.1.1) to identify miRNAs and snRNAs by searching the Rfam database (release 12.0).

Evolutionary and positive selection detection

By using the OrthoMCL program [69], 1818 single-copy gene families were identified from the assembled vegetable soybean, G. max Williams 82, T. pratense, V. vinifera, M. truncatula, C. arietinum, A. duranensis, P. vulgaris, V. angularis, and A. thaliana. Using the single-copy orthologous genes, we constructed a phylogenetic tree and estimated the divergence time among these species by using PhyML v3.0 [70], the MCMCtree program (version 4.4), and the TimeTree database [71]. CAFE (v2.1) [72] was used to identify changes in gene family expansion and contraction along the phylogenetic tree.

Then, based on the constructed phylogenetic tree, we incorporated a branch-site model in the PAML package.
and detected positive selection genes (PSGs) in the vegetable soybean genome. Vegetable soybean was used as the foreground branch, while the branches created for G. max Williams 82, T. pratense, V. vinifera, M. truncatula, C. arietinum, A. duranensis, P. vulgaris, and V. angularis were used as background branches. The details to screen PSGs were described in a previous study [73]. GO enrichment analysis was determined with Fisher’s exact test and was adjusted by the Benjamini-Hochberg (BH) method with the cutoff set at \( P < 0.05 \).

**Library preparation and sequencing for population-based resequencing**

Genomic DNA was extracted from the leaves of 60 vegetable soybeans using the CTAB method. Libraries were constructed following the manufacturer’s standard protocols (Illumina). Whole genome 150-bp paired-end reads were generated using the Illumina NovaSeq 6000 platform. The sequence data for each individual reached more than 14-fold depth (Supplementary Table 12, genome size calculated according to 1.01G). Additionally, the genome sequencing data of 103 wild soybeans and 155 oil soybeans were downloaded from a previous study[^30]. The total downloaded data had a 13-fold average depth (Supplementary Table 13).

**Alignment and SNP calling**

We mapped the high-quality paired-end reads to the ZH13 reference genome using BWA (v0.7.12) [74] with the parameter “mem -t 4 -k 32 -M” and then removed the PCR or optical duplicates using SAMtools (v1.3.1) [75]. SNP calling was performed using a UnifiedGenotyper approach as implemented in the package GATK (Genome Analysis Toolkit, v3.7–0-gcfedb67). To remove possible false-positive SNPs, SNPs with \( \text{FS} > 60.0, \text{QD} < 2.0, \text{MQ} < 20.0, \text{ReadPosRankSum} < -8.0, \text{MQRankSum} < -12.5, \text{missing rate} > 0.5 \) or \( \text{MAF} < 0.05 \) were filtered.

Next, we performed gene-based SNP annotation according to the annotation of the ZH13V2 genome with the package ANNOVAR (v2013–06–21). Briefly, based on genome annotation, the SNPs were classified as occurring in exonic regions, intronic regions, splicing sites, upstream and downstream regions, or intergenic regions. The details for annotation of genetic variants were described previously [45].

**Population structure and LD analysis**

A neighbor-joining tree was constructed by MEGA7 [76]. We used ADMIXTURE (v1.3.0) to infer the population structure. In consideration of HWE violations, we also screened SNPs by reconstructing the model-based clustering analysis and testing HWE violations (\( P > 10^{-4} \)). To determine the best genetic cluster subpopulations (\( K \)), cross-validation was tested for \( K \) ranging from \( K = 2 \) to \( 5 \). The termination criterion was setting as \( 10^{-6} \). Principal component analysis (PCA) was conducted using the program GTAC (v1.92).

Linkage disequilibrium (LD) decay of wild, grain and vegetable soybeans was calculated using PopLDDecay [77]. The \( r^2 \) values in a bin of 100 bp against the physical distance of pairwise bins are shown in the LD decay plot.

To study the introgression between vegetable and gain soybean populations, Treemix (v1.13) [78] was used to model the gene flow by inferring a maximum likelihood tree based on genome-wide allele frequency data and then identifying potential gene flow from a residual covariance matrix. SNPs with missing data were filtered out, and unlinked SNPs were retained (\( r^2 < 0.1 \)) according to the script provided on the link (https://speciationgenomics.github.io/Treemix/). The number of migration events (\( ^{-\text{m}} \)) was modeled from 0 to 3.

**Selective sweep analysis**

Selective sweeps were detected as described in Zhou et al.[16]. First, we used the widely used cross-population composite likelihood ratio test (XP-CLR) (v1.0) to scan selective regions over the soybean genome. XP-CLR was performed with the criteria: -w1 0.0005 2001000 1 - p1 0.95. The XP-CLR scores were calculated for each 50 kb genetic window with a step size of 10 kb. Windows with XP-CLR values accounting for 5% of the genome were considered selective sweep regions. Second, selective sweeps were also expected to reduce nucleotide diversity (\( \theta_s \)) in the cultivated population experiencing the selective sweep. We used VCFTools (v0.1.13) to calculate potential selective signals between population A and population B (\( \theta_s A/\theta_s B \)) for each 50 kb sliding window with a step size of 10 kb. The windows that contained more than 10 SNPs were excluded from further analysis, and the top 5% windows were considered selective sweep regions. The results were combined, and the genes in the merged selective regions of the ZH13 genome were identified as the final candidate selective genes. The figure was drawn using RectChr (Version 1.24, https://github.com/BGI-shenzhen/RectChr).

**Acknowledgments**

This work was supported by the National Natural Science Foundation of China (31872114), Zhejiang Provincial Important Science & Technology Specific Projects (2021C02052-2), State Key Laboratory for Managing Biotic and Chemical Threats to the Quality and Safety of Agro-products (2010DS700124-ZZ2005), and Zhejiang Provincial Important Science & Technology Specific Projects of Vegetable Breeding (2021C02065). We would like to thank Prof. Mingfang Zhang and Prof. Liangsheng Zhang of Zhejiang University, Prof. Pei Xu of China Jiliang University and Dr. Isaac J. Dopp of Pennsylvania State University for their critical comments on this paper.

**Author contributions**

N. Liu, Y. Gong conceived and designed the study. Y. Niu, Jinmin Lian and N. Liu performed the genome
sequencing, assembly and bioinformatics analyses. G. Zhang, Z. Feng, Y. Bo and B. Wang contributed to data analyses. N. Liu wrote the manuscript. Y. Gong discussed and revised the manuscript.

Data Availability

The genome assembly and all the sequencing data have been deposited in NCBI (https://www.ncbi.nlm.nih.gov/) under accession number PRJNA729722.

Conflicts of interest

The authors declare that they have no conflicts of interest.

Supplementary data

Supplementary data is available at Horticulture Research online.

References

1. Zhang GW, Xu SC, Mao WH et al. Determination of the genetic diversity of vegetable soybean [Glycine max (L.) Merr.] using EST-SSR markers. J Zhejiang Univ Sci B. 2013;14:279–88. https://10.1631/jzus.B1200243.
2. Young G, Mebrahtu T, Johnson J. Acceptability of green soybeans as a vegetable entity. Plant Foods Hum Nutr. 2000;55:323–33.
3. Hao CF, He SS, Wang CS et al. A modified Rogers distance algorithm for mixed quantitative-qualitative phenotypes to establish a Core collection for Taiwanese vegetable soybeans. Front Plant Sci. 2021;11:612106. https://10.3389/fpls.2020.612106.
4. Xu S, Liu N, Mao W et al. Identification of chilling-responsive microRNAs and their targets in vegetable soybean (Glycine max L.). Sci Rep. 2016;6:26619. https://10.1038/srep26619.
5. Kumar V, Rani A, Pratap D et al. Evaluation of vegetable-type soybean for sucrose, taste-related amino acids, and Isoflavonoid contents. Int J Food Prop. 2011;14:1142–51. https://10.1080/1094291010392761.
6. Jiang GL, Rutto PK, Ren S et al. Genetic analysis of edamame seed composition and trait relationships in soybean lines. Euphytica. 2018;214:158. https://10.1007/s10681-018-2237-9.
7. Graham PH, Vance CP. Legumes: importance and constraints to greater use. Plant Physiol. 2003;131:872–7. https://10.1104/pp.017004.
8. Saldívar X, Wang YJ, Chen P et al. Effects of blanching and storage conditions on soluble sugar contents in vegetable soybean. JWST - Food Science and Technology. 2010;43:1368–72. https://10.1016/j.jwst.2010.04.017.
9. Golicz AA, Batley J, Edwards D. TOWARDS plant pangenomics. Plant Biotechnol J. 2016;14:1099–105. https://10.1111/pbi.12499.
10. Varshney RK, Sinha P, Singh VK et al. 5Gs for crop genetic improvement. Curr Opin Plant Biol. 2020;56:190–6. https://10.1016/j.pop.2019.12.004.
11. Schmutz J, Cannon SB, Schlueter J et al. Genome sequence of the palaeopolyploid soybean. Nature. 2010;463:178–83. https://10.1038/nature08670.
12. Chan C, Qi XP, Li MW et al. Recent developments of genomic research in soybean. J Genet Genomics. 2012;39:317–24. https://10.1016/j.jgg.2012.02.002.
13. Wang Z, Tian ZX. Genomics progress will facilitate molecular breeding in soybean. Sci China Life Sci. 2015;58:813–5. https://10.1007/s11427-015-4908-2.
14. Li YH, Li W, Zhang C et al. Genetic diversity in domesticated soybean (Glycine max) and its wild progenitor (Glycine soja) for simple sequence repeat and single-nucleotide polymorphism loci. New Phytol. 2010;188:242–53. https://10.1111/j.1469-8137.2010.03344.x.
15. Zhou Z, Ziang J, Wang Z et al. Resequencing 302 wild and cultivated accessions identifies genes related to domestication and improvement in soybean. Nat Biotechnol. 2015;33:408–14. https://10.1038/nbt.3096.
16. Hyten DL, Song Q, Zhu Y et al. Impacts of genetic bottlenecks on soybean genome diversity. Proc Nat Acad Sci U S. 2006;103:16666–71. https://10.1073/pnas.0604379103.
17. Lam HM, Xu X, Liu X et al. Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet. 2010;42:1053–9. https://10.1038/ng.715.
18. Li YH, Zhou G, Ma J et al. De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol. 2014;32:1045. https://10.1038/nbt.2979.
19. Liu ZX, Li H, Wen Z et al. Comparison of genetic diversity between Chinese and American soybean (Glycine max (L.)) across accessions revealed by high-density SNPs. Front Plant Sci. 2017;8:2014. https://10.3389/fpls.2017.02014.
20. Olsen KM, Wendel JF. A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol. 2013;64:47–70. https://10.1146/annurev-arplant-050312-120048.
21. Doebley JF, Gaut BS, Smith BD. The molecular genetics of crop domestication. Cell. 2006;127:1309–21. https://10.1016/j.cell.2006.12.006.
22. Sedivy EJ, Wu FQ, Hanzawa Y. Soybean domestication: the origin, genetic architecture and molecular bases. New Phytol. 2017;214:539–53. https://10.1111/nph.14418.
23. Dong Y, Yang X, Han BH et al. Pod shattering resistance associated with domestication is mediated by a NAC gene in soybean. Nat Commun. 2014;5:3552. https://10.1038/ncomms3552.
24. Sun L, Miao Z, Cai C et al. MsMa1-1, encoding a calcineurin-like protein, controls hard-seededness in soybean. Nat Genet. 2015;47:399–+. https://10.1038/ng.3339.
25. Wang M, Li W, Fang C et al. Parallel selection on a dormancy gene during domestication of crops from multiple families. Nat Genet. 2018;50:1435. https://10.1038/s41588-018-0229-2.
26. Lu SJ, Dong L, Fang C et al. Stepwise selection on homeologous PRR genes controlling flowering and maturity during soybean domestication. Nat Genet. 2020;52:428. https://10.1038/s41588-020-0604-7.
27. Dong DK, Fu XJ, Yuan F et al. Genetic diversity and population structure of vegetable soybean (Glycine max (L.) Merr.) in China as revealed by SSR markers. Kulturpflanze. 2014;61:173–83. https://10.1007/s10722-013-0024-y.
28. Liu Y, Du H, Li P et al. Pan-genome of wild and cultivated soybeans. Cell. 2020;182:162, e113–176.e13. https://10.1016/j.cell.2020.05.023.
29. Song JF, Liu CQ, Li DJ et al. Evaluation of sugar, free amino acid, and organic acid compositions of different varieties of vegetable soybean (Glycine max (L.) Merr.). Ind Crops Prod. 2013;50:743–9. https://10.1016/j.indcrop.2013.08.064.
68. Wang SD, Liu S, Wang J et al. Simultaneous changes in seed size, oil content and protein content driven by selection of SWEET homologues during soybean domestication. Natl Sci Rev. 2020;7:1776–86. https://10.1093/nsr/nwaa110.

69. Li L, Stoeckert CJ, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178–89. https://10.1101/gr.1224503.

70. Guindon S, Dufayard JF, Lefort V et al. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol. 2010;59:307–21. https://10.1093/sysbio/syq010.

71. Hedges SB, Dudley J, Kumar S. TimeTree: a public knowledge-base of divergence times among organisms. Bioinformatics. 2006;22:2971–2. https://10.1093/bioinformatics/btl505.

72. De Bie T, Cristianini N, Demuth JP et al. CAFE: a computational tool for the study of gene family evolution. Bioinformatics. 2006;22:1269–71. https://10.1093/bioinformatics/btl097.

73. Yang JH, Deng G, Lian J et al. The chromosome-scale genome of melon dissects genetic architecture of important agronomic traits. iScience. 2020;23:101422. https://10.1016/j.isci.2020.101422.

74. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. eprint arXiv. 2013;00:1–3.

75. Li H, Handsaker B, Wysoker A et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25:2078–9. https://10.1093/bioinformatics/btp352.

76. Kumar S, Stecher G, Tamura K. MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets. Mol Biol Evol. 2016;33:1870–4.

77. Zhang C, Dong SS, Xu JY et al. PopLDdecay: a fast and effective tool for linkage disequilibrium decay analysis based on variant call format files. Bioinformatics. 2019;35:1786–8. https://10.1093/bioinformatics/bty875.

78. Pickrell JK, Pritchard JK. Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet. 2012;8:e1002967. https://10.1371/journal.pgen.1002967.