Molecular Plant, Volume *15*

Supplemental information

Pepper variome reveals the history and key loci associated with fruit domestication and diversification

Yacong Cao, Kang Zhang, Hailong Yu, Shumin Chen, Donghui Xu, Hong Zhao, Zhenghai Zhang, Yinqing Yang, Xiaozhen Gu, Xinyan Liu, Haiping Wang, Yaxin Jing, Yajie Mei, Xiang Wang, Véronique Lefebvre, Weili Zhang, Yuan Jin, Dongliang An, Risheng Wang, Paul Bosland, Xixiang Li, Ilan Paran, Baoxi Zhang, Giovanni Giuliano, Lihao Wang, and Feng Cheng

Supplementary Information for

Pepper variome reveals the history and key loci associated with fruit domestication and diversification

Yacong Cao[1]†, Kang Zhang[1]†, Hailong Yu[1]†, Shumin Chen[1]†, Donghui Xu[1]†, Hong Zhao[1], Zhenghai Zhang[1], Yinqing Yang[1], Xiaozhen Gu[1], Xinyan Liu [1], Haiping Wang[1], Yaxin Jing[1], Yajie Mei[1], Xiang Wang[1], Véronique Lefebvre[2], Weili Zhang[1], Yuan Jin[1], Dongliang An[1], Risheng Wang[3], Paul Bosland[4], Xixiang Li[1], Ilan Paran[5], Baoxi Zhang[1], Giovanni Giuliano[6]*, Lihao Wang[1]*, Feng Cheng[1]*

*Feng Cheng, Lihao Wang, Giovanni Giuliano

Email: chengfeng@caas.cn, wanglihao@caas.cn, giovanni.giuliano@enea.it

This file includes:

Complete version of **Materials and Methods**

Supplementary Tables S3, S4, S6, and S7

Supplementary Figures S1 to S13

SI **References**

Anther file as a spreadsheet includes:

Supplementary Tables S1, S2, and S5

Materials and Methods

Pepper sample collection

A total of 347 pepper accessions were used in this study, including the core collection of the China Gene Bank (Gu et al., 2019; Wang et al., 2018), core breeding inbred lines of Institute of Vegetables and Flowers, and international typical local varieties and wild species. In total, 12 species were collected, including 309 *C. annuum* (cultivated species), two *C. annuum* var. *glabriusculum* (wild species), 10 *C. frutescens* (cultivated species), 10 *C. chinense* (cultivated species), five *C. baccatum* (cultivated species), one *C. baccatum* var. *baccatum* (wild species), two *C. pubescens* (cultivated species), one *C. cardenasii* (wild species), two *C. chacoense* (wild species), one *C. eximium* (wild species), one *C. flexuosu*m (wild species), one *C. galapagoense* (wild species), one *C. minutiflorum* (wild species), and one *C. rhomboideum* (wild species). Samples originated from the Americas, Europe, Asia, and Africa and were obtained from several genebanks.

Agricultural characteristics investigation

The experimental materials were planted five times, including a greenhouse in Langfang during spring 2016; in a greenhouse in Beijing, in a plastic greenhouse in Dehong, and in the open field of Urumqi during spring 2017; and in the open field of Urumqi during spring 2018. Three replicates were planted in each experiment. In the Langfang and Beijing experiments, two to three plants were planted, and two fruits were measured in each replicate; and in Dehong and Urumqi experiments, 10 plants were planted, and 10 fruits were measured in each replicate.

The mature fruit traits measured were as follows: (1) fruit length: the distance from the pedicel attachment to the fruit apex; (2) fruit diameter: the maximum width; (3) fruit shape index - the ratio of fruit length to fruit diameter; (4) fruit orientation - if the fruit points upward, it is recorded as erect, otherwise it is recorded as pendent; (5) the content of capsaicin and dihydrocapsaicin of dried fruits without seeds and pedicels, as measured by ultra-performance liquid chromatography (UPLC).

For the measurements of capsaicin and dihydrocapsaicin, mature fruits without seeds and pedicels were dried at 55 to 60℃ for about two days, ground and processed according to an improved agricultural industry standard (NY/T 1381-2007, Determination of Capsaicin by high-performance liquid chromatography [HPLC]). For each determination, 0.2 g (accurate to 0.0001 g) of sample powder was extracted with 25 mL of methanol-tetrahydrofuran (1:1) (HPLC-grade; Sigma, USA) solution, shaken, extracted in an ultrasonic extractor for 30 min in a 60℃ water bath, and then filtered. The filter residue and filter paper were re-extracted with 25 mL methanol-tetrahydrofuran solution in an ultrasonic extractor for 10 min, filtered, and then the above operation was repeated. The three collected filtrates were combined, concentrated on a rotary evaporator (70–75℃ water bath) to about 30 mL, and then transferred to a 50 mL volumetric flask, where methanol was added to produce the final volume. After being filtered through a 0.22-μm micron organic phase filter membrane, chromatographic analysis was performed. The capsaicin content in the test solution was diluted to 0.13 mg/L–160 mg/L, and the dihydrocapsaicin content was diluted to 0.04 mg/L–160 mg/L. The filtrate was injected into a UPLC system (Waters UPLC, USA) with an ACQUITY UPLC BEH C18 1.7 μm, 2.1*100 mm column (Waters, USA) and Photo Diode Array Detector. The mixture of methanol and distilled water (70:30) represented the mobile phase. The flow rate was 0.20 mL/min. The injection volume was 2 μL, and the column temperature was 30℃. The detection wavelength was at 280 nm, and the retention times were 3.276 min for capsaicin and 4.395 min for dihydrocapsaicin. Capsaicin and dihydrocapsaicin external standards (Sigma, USA) were prepared as 200 mg/L stocks in absolute methanol. The stock solution was diluted with methanol to a series of standard working solutions of 100 mg/L, 50 mg/L, 20 mg/L, 10 mg/L, 1 mg/L, and 0.2 mg/L, and measured under the above liquid chromatography conditions. With

the concentration of capsaicin and dihydrocapsaicin as the ordinate, and the corresponding peak area integral as the abscissa, the standard curves and linear regression equations were calculated. After the prepared test solution was measured, it was calibrated at multiple points and quantified by the peak area integral values. Finally, the capsaicin and dihydrocapsaicin contents were converted into mg per kg dry weight.

***DNA extraction, library construction,* and sequencing**

The DNA library was prepared according to the manufacturer's instructions (NEBNext® UltraTM DNA Library Prep Kit for Illumina®). The adaptor-ligated DNA was cleaned with the MinElute Reaction Cleanup Kit instead of Vortex AMPure XP Beads. After PCR amplification, the PCR products were assessed on 2% agarose electrophoresis. Next, quantitative analysis of the DNA library was conducted with a Qubit 2.0 (Invitrogen, USA) and an Agilent 2100 Bioanalyzer (Agilent, USA). All the libraries with 300 bp insert size were sequenced on the Illumina Solexa platform, and 150 bp paired-end reads were generated.

PacBio HiFi Sequencing

Genomic DNA was extracted from the leaves of Zunla-1 using the Blood & Cell Culture DNA Kit .We used 30μl high quality genomic DNA to construct a SMRT bell target size library using the PacBio SMRT bell template prep kit. The constructed library was sequenced on the PacBio Sequel II platform using the circular consensus sequencing (CCS) mode at Berry Genomics. A total of 32.54 Gb of raw HiFi sequencing reads with the N50 length of 20.25 kb were obtained.

Genome gap closing

The assembly gaps in the Zunla-1 genome was closed using the generated HiFi reads. Specifically, the HiFi data was equally divided into four parts. TGS-GapCloser (Xu et al., 2020) (parameter: "--scaff --minimap_arg '-x asm20' --racon -tgstype pb") was used to fill the gaps with four rounds based on the four parts of data.

Protein-coding gene prediction

Prediction of the protein-coding genes in the gap-closed Zunla-1 genome was performed according to the previously described method (Yang et al., 2022). Briefly, the results from homology-based searches, mRNA-seq assisted prediction, and *ab initio* prediction were integrated using EVidenceModeler (version 1.1.1)(Haas et al., 2008). In the homology-based prediction process, protein sequences of eleven species (pepper, Arabidopsis, cacao, coffee, eggplant, grape, petunia, potato, snapdragon, tobacco, and tomato) were collected and were aligned to the updated Zunla-1 genome to identify the homologous genes using Exonerate (version 2.4.7)(Guy St C and Ewan, 2005). The RNA-seq data obtained from previous study (Qin et al., 2014) was used for mRNA-seq assisted prediction. The *ab initio* gene prediction was performed using GeneMark (version 4.61_lic) (Lomsadze et al., 2005) and Augustus (version 3.3.3) (Hoff and Stanke, 2018).

Variants calling and annotation

Trimmomatic v0.33 trimmed the Illumina fastq raw reads and removed adapters based on the manufacturer's adapter sequences to obtain clean reads. Clean reads were mapped to the Zunla-1 2.0 genome (Qin *et al.*, 2014) using BWA 0.7.17 (Li and Durbin, 2009) with the following parameters: mem -M -A 1 -B 4 -O 6 -E 1. Samtools was used to sort the sam files and convert them into bam files. The function MarkDuplicates integrated in GATK (version 4.1.7) was used to deduplicate the bam files.

Then, variants were called out using the HaplotypeCaller function of GATK with default parameters and generated the raw VCF files. Variant loci that were ungenotyped in more than 40% of the 311 *C. annuum* accessions, as well as those loci with a heterozygosity ratio of more than 10% and a minor allele frequency < 5% in the *C. annuum* population, were further filtered by an in-house Perl script. The genotype information of these remaining reliable variant loci in the 311 accessions of *C. annuum* and 37 accessions from other *Capsicum* species were then obtained for further data analysis. By doing this, we reduced the bulk and biased variants resulting from the large genetic difference between the genomes of *C. annuum* and the other *Capsicum* species. Furthermore, the ANNOVAR (Wang et al., 2010b) was used to annotate variant function. Based on annotated data, SNPs and InDels were classified according to their position in or outside genes as being in: (1) inter-genic regions. (2) CDS (coding sequences) regions, (3) intronic regions, and (4) UTR (untranslated regions) (both 5'and 3'). Variants in CDS were subdivided into two groups: one group causing changes to the coding amino acids, including non-synonymous SNPs and non-threefold frame shift InDels, and another group with synonymous SNPs and InDels without frame-shifts. Intronic variants were divided into two groups: splice site mutations (within 2 bp of the splice site) and others.

The CNVs (copy number variations) and PAVs (presence/absence variations) were called out using the CNVnator (Abyzov et al., 2011) and Control-FREEC (Boeva et al., 2012). To reduce the false positives, only the accessions with resequencing depth ≥ 8× (123 in total) were selected for the analysis. There were 23 accessions that eventually failed to meet the criterion ($4 \leq SD \leq 5$) in CNVnator when a wide range of bin sizes were used. These accessions were therefore excluded from further analyses, while for the retained 100 accessions, the bin sizes were set as 800−1200 bp, which were varied in different accessions, to optimize the SD value. For Control-FREEC, the window and the step sizes were set as 1200 bp and 600 bp, respectively. These two tools were run separately, and only the intersections of detected intervals were retained. The results were further refined by merging the variants with reciprocal overlap ≥ 80%. Finally, variants with frequency < 3 or > 80 in the 100 accessions were filtered out.

Frequency distribution of variants

To estimate the density of the variants in the pepper genome, the number of various types of SNPs and InDels were counted in a one Mb sliding window across the 12 chromosomes of pepper, and the results were drawn in a genomic circos plot using the Circos tool (Krzywinski et al., 2009).

For the frequency of average variants in the gene body and the flanking regions of the gene, firstly the 5 kb regions flanking each gene were determined. When the flanking region overlapped with the gene body or flanking region of another gene, the process was stopped at the middle site in the two neighboring genes. A one kb sliding window with a 100 bp step screened the flanking regions and gene body of all pepper genes, and the number of SNPs and InDels located in the window were calculated. After that, each of these windows were further averaged across all the genes, and the upper and lower quartiles were calculated.

KASP genotype verification

Kompetitive Allele Specific PCR (KASP) verified the SNPs. The SNPs were randomly selected for the KASP assay, and primers were designed using BatchPrimer3 online (You et al., 2008) (https://probes.pw.usda.gov/batchprimer3/). The allele-specific forward primers allele 1 and allele 2 carried the FAM- and HEX-labeled tails with the targeted SNP at the 3' end. A common reverse primer was designed to amplify the 80–200 bp target region. The test operation mainly followed the #KBS-1016-017 (Laboratory of the Government Chemist, UK) instructions (www.lgcgroup.com/genomics) and was as follows: 1) DNA samples were arrayed into 384-well PCR plate. No-template controls are

included on each plate; 2) to prepare the genotyping mix, 2.0 μL DNA (10–50 ng/μL), 2.0 μL 2× KASP Master mix, 0.055 μL KASP Assay mix were combined. The total reaction volume was 4μL; 3) the genotyping mix was dispensed onto the reaction plate, to which 2.0μL genotyping mix was added to each DNA sample in the reaction plate using a pipette; 4) the plate was sealed and centrifuged; 5) the thermal cycle was run in the Roche LightCycler 480 System®. The cycle included an initial denaturation step of 15 min at 94°C, followed by 10 touchdown cycles of 20 s at 94°C and 60 s at 61°C, then decreasing by −0.6°C per cycle, followed by 26 cycles of 20 s at 94°C and 60 s at 55°C, and nine additional cycles of 20 s at 94°C and 60 s at 57°C. 6) the plate was then read, and the data were analyzed. After completion of the thermal cycle, the reaction plate was read in a FRET-capable platter reader at 37°C. The LightCycler 480 Software® was used to analyze the data. The result was assessed based on the difference between the cycle threshold (Ct) values of the two alleles. The greater the difference, the greater the possibility of a pure gene (determine its allele according to the type of fluorescent substance connected by the special primers), and the smaller the difference, the more likely it was heterozygous.

Phylogenetic tree construction

In order to select neutral and reliable SNP variants to investigate the phylogenetic relationships of the resequenced pepper accessions, these genic SNPs that were annotated as synonymous variations by ANNOVAR were extracted (Wang *et al.*, 2010b). It was determined that 165,864 such SNPs were in the 347 resequenced pepper accessions. The distance matrixes with the genotype data for all 347 pepper accessions was generated. Within the 311 *C. annuum* accessions the pairwise allele sharing distance (ASD) was calculated for each pair of two accessions, and the data in the distance matrix were used to construct a neighbor-joining (NJ) phylogenetic tree using PHYLIP3.69 (Retief, 2000).

Population structure analysis

The software Admixture (version 1.3.0) (Pritchard et al., 2000) performed genetic structure analyses using 160,000 randomly selected SNPs from the 311 *C. annuum* accessions. Admixture was run with k = 2–15 clusters with the default parameters. Finally, k = 7 (seven major genetic components, **Fig. 1d**) that showed the smallest CV error was selected to represent the population genetic features of *C. annuum*.

Linkage disequilibrium

Correlation coefficients (r^2) were calculated for each pair of SNP loci with a distance of 1–500 kb between them and averaged per bin of 50 bp distance, producing a histogram of r^2 per such distance bin. From these histograms, the distance at which r^2 decayed to half its maximum value was determined as a linkage disequilibrium (LD) length estimate of a particular sample group.

Determination of ancestral and derived alleles

In order to estimate the frequency of genotype status—ancestral or non-ancestral (derived)—of the SNP variants and compare them in different groups of pepper, the ancestral/derived alleles for each SNP locus were determined. To obtain a reliable genotype status, the genotypes of the two wild *C. annuum* var. *glabriusculum* samples were used as the controls. Only the homozygous SNP locus shared by the two samples were retained. We then recorded the genotype of the two wild samples as the ancestral allele of the SNP locus, while the other allele in the pepper population at this SNP locus was recorded as the derived one. With this rule, the ancestral/derived alleles of 14,194,356 SNPs out of the 18,372,022 total SNPs were successfully obtained. The information of derived alleles of these SNPs was used in the data analysis.

Genome diversity and selection sweeps

Two measures were used to estimate the genomic diversity of the *C. annuum* population: π and ROD (reduction of diversity). The π measures the genomic diversity (Tajima, 1983) through computing the average difference per locus on each pair of accessions. While ROD is a measure based on π, which estimates the diversity reduction of a sample group with respect to the control sample group (Xu et al., 2011). In this study, ROD detected the genomic regions under group-specific selection as compared to the parental group.

Akey's F_{ST} calculated the pairwise genomic differentiation between two groups of samples (Akey et al., 2002), which evaluates the strength of genomic divergence between two compared groups. Values ranged from 0 (no differentiation) to 1 (complete or fixed differentiation).

XP-EHH detected signals of recent positive selection (group-specific selection signatures) (Sabeti et al., 2007) in the blocky fruit group of *C. annuum*, using the combined software of Beagle (version 5.1) (Browning et al., 2018) and SelScan (version v1.3.0) (Szpiech and Hernandez, 2014), with default parameters. Large XP-EHH values indicate unusually long haplotypes in the derived group under selection as compared to the control group. Tajima's D was applied to estimate genomic regions under negative selection/purifying selection, and outliers were detected with a threshold of less than -2 (Tajima, 1989).

To obtain the strongest selection signals from the large genome of pepper (3.6 Gb), which is comparatively larger than most crop genomes sequenced, the calculated values of these population genomic measures were further averaged using a 10 Mb window with a 1 Mb increment sliding across the whole genome of pepper, except for Tajima's D, which was calculated directly in each sliding window. The top 5% of ROD and XP-EHH were determined as outlier signals of genomic regions under selection. If two outlier windows overlapped with each other, then they were merged into one region.

Selection strength on gene units

The gene body combined with its 5 kb upstream and 5 kb downstream regions were defined as a gene unit, and 35,336 such gene units were generated using the Zunla-1 genome (version 2.0) as the reference (Qin *et al.*, 2014). Measures of π and ROD (Xu *et al.*, 2011) were then calculated using SNP variants located at each of these gene units, which were further used to estimate the gene level selection strength in the pepper population. The top 5% of ROD values, together with the top 30% of smaller π values in the control group, were determined as outliers of gene units that were under selection.

Major haplotype sharing analysis

The major haplotypes in the blocky fruit groups were first determined using a 10 Mb window with 1 Mb increment sliding across the 12 chromosomes of the pepper reference genome. In total, there were 548 such windows. In each window, the major allele frequency of each SNP in the blocky fruit pepper was calculated, and the SNP loci with a value ≥0.8 were counted and collected. If more than 80% of the SNPs in the window were counted, then these major alleles comprise the major haplotype of the blocky fruit groups in the analysed window. These major haplotypes were kept and compared to all of these resequenced *C. annuum* accessions. The average score of allele sharing between the major haplotype and each of the 311 accessions in the analyzed window was calculated and considered as a major haplotype sharing score (MHS). The MHS then estimated the relationship on the origin of the blocky fruit pepper-specific haplotype in the pepper population.

K_S analysis

In order to obtain the gene sequences for blocky fruit pepper, the consensus genome of blocky fruit pepper by replacing the Zunla-1 reference genome was generated (Qin *et al.*, 2014) using the major genotype of blocky fruit pepper at each variant locus. Gene sequences of blocky fruit pepper were extracted based on the gene annotation coordinate file. After that, pairwise protein sequences between blocky fruit pepper and Zunla-1 (non-blocky fruit pepper) were aligned by MUSCLE (Edgar, 2004). Protein alignments were translated into coding sequence alignments using an in-house Perl script. The K_S values were calculated with the coding sequence alignments using the method of Nei and Gojobori as integrated in the KaKs_calculator (Wang et al., 2010a).

GWAS analysis

GEMMA (Zhou and Stephens, 2012) was applied to perform GWAS. The relatedness (kinship) matrix was first calculated by GEMMA and applied to correct the pepper population stratification in the 311 accessions of *C. annuum*. The linear mixed model (LMM) analyzed the correlation of the population, and *P*-values of all SNPs to each trait were generated by GEMMA. The GWAS signals of traits analysed in this study are replicable in the five phenotypic trials, and GWAS results on the averaged values of the five trials were used to make the Manhattan plotting.

Validation of associations in another pepper population

The candidate genes were then further validated in another 241 *C. annuum* accessions. According to the mutation types, different types of primers were developed. For SNPs related to the fruit shape index, primers spanning the SNP were designed (F: GGAGGTCAGACGTGGATCATC, R: TGATGGTTTGTTTGGTTTGAGCA) and the genotype was determined by the PCR products sequencing using the forward primer as the sequencing primer. For structural variations related to fruit orientation, primers that span the variation segment were designed (F: GTGCTGCAAGAGGAAGAAAACT, R: CAGCCCTCTTTTTCCTTGTATG), and the genotype was determined by the PCR-amplified fragment size indicated by polyacrylamide gel and agarose gel electrophoresis, respectively.

RNA-Seq and data analysis

For transcriptome sequencing, 22 accessions of various fruit types and 10 accessions with different pungency levels were selected. The pericarp from accessions with various fruit types at the early stage about 5 to 7 d after pollination, and then about 15 to 20 d after pollination, the placenta of different pungency levels at the middle stage of fruit development were sampled and immediately placed into liquid nitrogen for more than 30 s, then they were stored in an ultra-low temperature refrigerator at -80°C until sequencing. The RNA-Seq data were mapped to the genome of Zunla-1 (version 2.0) (Qin *et al.*, 2014) using the software Hisat2 (version 2.1.0) (Kim et al., 2019) together with Samtools (version 1.9) (Li et al., 2009) with default parameters. The mapping results were then submitted to featureCounts (Liao et al., 2014) to calculate the TPM (tags per million reads) values to estimate the level of gene expression in each sample. To investigate the expression pattern of gene *Capana12g000954* in accessions with different fruit orientations, we sampled the flower, fruit, leaf, root, and stem organs from two accessions (erect vs pendent), which were then sequenced and analysed using the aforementioned methods.

BSA-Seq and data analysis

A wild pepper (*C. annuum var. glabriusculum*) accession Ac1979 with erect fruit was crossed with a blocky pepper accession Qiemen with pendent fruit to construct an F_2 population of ~360 plants. 30

plants exhibiting the erect and 30 exhibiting the pendent fruit phenotypes were selected from the population. DNA from the parents and the two F_2 groups was extracted, and the later were further combined into two pools for library construction. The resulting libraries were sequenced on the MGI's DNBSEQ-T7 platform from BGI-Shenzhen Company (Shenzhen, China). Clean reads were mapped to the Zunla-1 2.0 genome and the variants were called following the same method described above. Low quality (Q < 20, DP < 5) or multi-allelic variants were filtered out, and the homozygous SNPs that showed polymorphism between the two parents were obtained using an in-house Perl script. The ΔSNP-index was calculated based on a 500 kb sliding window with a 50 kb increment as described previously (Zhang et al., 2020).

Quantitative RT-PCR analysis of candidate genes of pepper fruit orientation

For qRT-PCR analysis of the gene *Capana12g000954*, total RNA was extracted from erect and pendent peppers from following tissues: leaves, flower buds, flowers and flower pedicels at 3 d post-pollination, fruis and fruit pedicels at 7 d post-pollination. Samples were immediately frozen in liquid nitrogen for at least 30 s and then stored at -80°C. The RNA isolation and cDNA synthesis were performed using an SV Total RNA Isolation System kit (Promega) and the GoScriptTM Reverse Transcription System kit (Promega), respectively, according to the manufacturer's instructions. Gene expression analysis was performed by qRT-PCR with the LightCyler®480 SYBR Green I Master on a LightCycler®480 II (Roche) real-time PCR system. The *Capsicum actin* (*GQ339766.1*) and *EF1a* (*AY496125*) were used as reference genes. The primers (F- AGTTATGCAAGCTCTGATCTGT and R-TGTTGTTTCATAGACGGGCA) were designed online (https://primer3.ut.ee/). The qRT-PCR was performed with two biological and three technical repetitions and each reaction was repeated three times. The same method was used to analyze gene silencing in the VIGS experiment.

Virus-induced gene silencing (VIGS)

VIGS was used to verify the function of the *up* gene *Capana12g000954*. We selected the specific sequence region (4 bp-303 bp) of 5' *Capana12g000954* to construct the VIGS vector. We confirmed the sequence specificity by BLAST searching of the Zunla genome databases. A 300 bp fragment of *Capana12g000954* was cloned from a pepper CDNA template using gene specific primers F-TGAGTAAGGTTACCGAATTCCAAGCTGGCCAAGAGAAGAA and R-GGAGGCCTTCTAGAGAATTCCCTCCAGGAGAAATCTGCTTA, containing an *EcoRI* site. The resulting product was inserted into the *EcoRI* site of pTRV2, to construct TRV2::*up*. A fragment of about 100 bp of *CaPDS* gene was inserted into pTRV2 to construct the control vector TRV2::*CaPDS*. The tobacco rattle virus (TRV)-based vectors TRV1 and TRV2::*rec* were tansfromed into *Agrobacterium tumefaciens* (strain GV3101) for infiltration. Bacteria were grown and resuspended in infiltration solution at an OD_{600} of about 4.0. After standing at room temperature for 3 hours, the Agrobacterium suspensions containing the TRV1 vector and one of the TRV2 vectors (TRV1+TRV2; TRV1+TRV2::*CaPDS*; TRV1+TRV2::*up*) were mixed at a ratio 1:1 and infiltrated. 2-month old plants of the accession "Changyang chili" (pendent fruit) was used for VIGS experiments. The plant growth environment was 23±2°C, the relative humidity was 60%-80%, and the light was 16h/8h (light/dark). After infiltration, the plants were placed in the dark and cultured for 24h and then reverted to 16/8h light-dark alternate culture. Five plants were used for each of the three groups (non-infiltrated, TRV2::00, and TRV2::*up* experiments. Only plants infiltrated with TRV2::*up* plants showed fruits with erect orientation, two months after infiltration. To investigate the expression of the *up* gene, total RNA was isolated from pendent fruit pedicels of wild type and TRV2::00 and from erect pedicels of TRV2::*up* plants. Gene expression analysis was performed by the qRT-PCR method described above. Four biological replicates (pedicels) and four technical replicates were performed.

Bisulfite sequencing and DNA methylation analyses

To determine the methylation levels of the 579 bp sequence in the promoter region of *Capana12g000954*. Bisulfite treatment of genomic DNA from erect pepper was conducted using EpiArt DNA Methylation Bisulfite Kit according to the manufacturer's instructions (Vazyme). Un-methylated cytosine (C) was converted into uracil (U), which was further interpreted as thymine (T). Treated DNA was used for PCR amplification using specific primers (F: AGTTAAATAAAGAAAGGTTTAGTGAAAT, R: TTTTATCCCTATCTATTAACCATAAAAC). PCR conditions were as follows: 98 °C for 30 s; 35 cycles of 98 °C for 10 s, 53 °C for 5 s, and 72 °C for 15 s and 72 °C for 1 min. The PCR products were then sequenced to determine the methylation status of C loci.

Supplementary Tables

Table S1. Summary of the sampled collection of pepper.

Presented in the spreadsheet file.

Table S2. SNP loci verification by Kompetitive Allele-Specific PCR (KASP).

Presented in the spreadsheet file.

Table S3. Positions of the SNPs and InDels identified in the pepper population*.

Variant	Genic						Intergenic	Total
	CDS		Intron		UTR			
	NS/FS	S/NFS	splice	intron	5'UTR	3'UTR		
SNP	55,711	33,346	471	274,840	464,143	386,470	17,157,041	18,372,022
INDEL	2,027	900	153	27,197	39,799	33,621	699,539	802,875

**: Abbreviations: CDS: coding sequences; UTR: untranslated region; NS: nonsynonymous and S: synonymous mutations of SNPs; FS: frameshift and NFS: non-frameshift mutations of InDel.*

Table S4. Genomic regions under selection in transitions between different C. annuum groups, measured by ROD (reduction of diversity).

Groups (a *vs.* b)*	Chromosome	Start	Stop
II vs. I	Chr02	1	3,500,001
II vs. I	Chr04	80,500,002	100,500,001
II vs. I	Chr08	55,500,002	91,500,001
II vs. I	Chr09	127,500,002	185,500,001
II vs. I	Chr11	86,500,002	102,500,001
II vs. I	Chr11	105,500,002	106,500,001
III vs. II	Chr04	144,500,002	152,500,001
III vs. II	Chr05	166,500,002	169,500,001
III vs. II	Chr05	171,500,002	181,500,001
III vs. II	Chr05	188,500,002	205,500,001
III vs. II	Chr07	82,500,002	91,500,001
III vs. II	Chr07	129,500,002	180,500,001
III vs. II	Chr10	85,500,002	106,500,001
III vs. II	Chr12	88,500,002	104,500,001
III vs. II	Chr12	144,500,002	145,500,001
VII-IX vs. II	Chr04	127,500,002	138,500,001
VII-IX vs. II	Chr06	154,500,002	155,500,001
VII-IX vs. II	Chr06	159,500,002	164,500,001
VII-IX vs. II	Chr06	169,500,002	170,500,001
VII-IX vs. II	Chr07	82,500,002	90,500,001
VII-IX vs. II	Chr09	64,500,002	72,500,001
VII-IX vs. II	Chr09	124,500,002	128,500,001
VII-IX vs. II	Chr09	131,500,002	144,500,001
VII-IX vs. II	Chr11	143,500,002	150,500,001
VII-IX vs. II	Chr12	75,500,002	151,500,001
IV,VI vs. III	Chr07	130,500,002	158,500,001
IV,VI vs. III	Chr08	62,500,002	77,500,001
IV,VI vs. III	Chr08	93,500,002	96,500,001
IV,VI vs. III	Chr10	74,500,002	106,500,001
IV,VI vs. III	Chr11	61,500,002	105,500,001
IV,VI vs. III	Chr11	124,500,002	137,500,001

** Group a compared to its progenitor group b.*

Table S5. Significantly enriched GO (gene ontology) terms in genes under selection transitions between different C. annuum groups. Note: GO functions colored in red indicate important candidate gene sets involved in pepper fruit evolution.

Presented in the spreadsheet file.

Table S6. Gene units under selection in different pepper groups compared to their progenitor groups in the pepper population.

Gene	ROD	π_a	π_b	#SNPs	Gene symbol/annotation	Groups (a *vs.* b)*
Capana04g002188	0.92	0.03	0.44	62	AP2-A, sepal regulation	II vs. I
Capana02g000700	0.98	0.01	0.48	78	AP2-A, sepal regulation	III vs. II
Capana05g000060	1	0	0.3	15	SUN, SlIQD12, fruit shape regulation	IV, VI vs. III
Capana07g001005	1	0	0.47	33	agamous family gene, flower development related	IV, VI vs. III
Capana10g000984	0.96	0.02	0.43	73	cyclin-dependent protein kinase	IV, VI vs. III
Capana10g001014	0.96	0.02	0.44	35	cyclin-dependent protein kinase	IV, VI vs. III
Capana03g002426	0.72	0.1	0.35	52	mitotic phase inducer phosphatase-like protein	IV, VI vs. III
Capana09g001401	0.79	0.06	0.31	191	glycine-rich cell wall structural protein-like	IV, VI vs. III

** Group a compared to its progenitor group b.*

390 **Table S7**. The expression of up gene (Capana12g000954) in different organs of pepper samples with
391 erect or pendent fruits.

Type	Leaf*	Stem	Root	Flower	Fruit
Erect	0.08	0.76	0.36	50.19	2.18
Pendent	0.00	2.00	2.43	23.99	35.83

392 **: The gene expression values were estimated by TPM (Transcripts Per kilobase of exon model per*
393 *Million mapped reads).*

394

Supplementary Figures

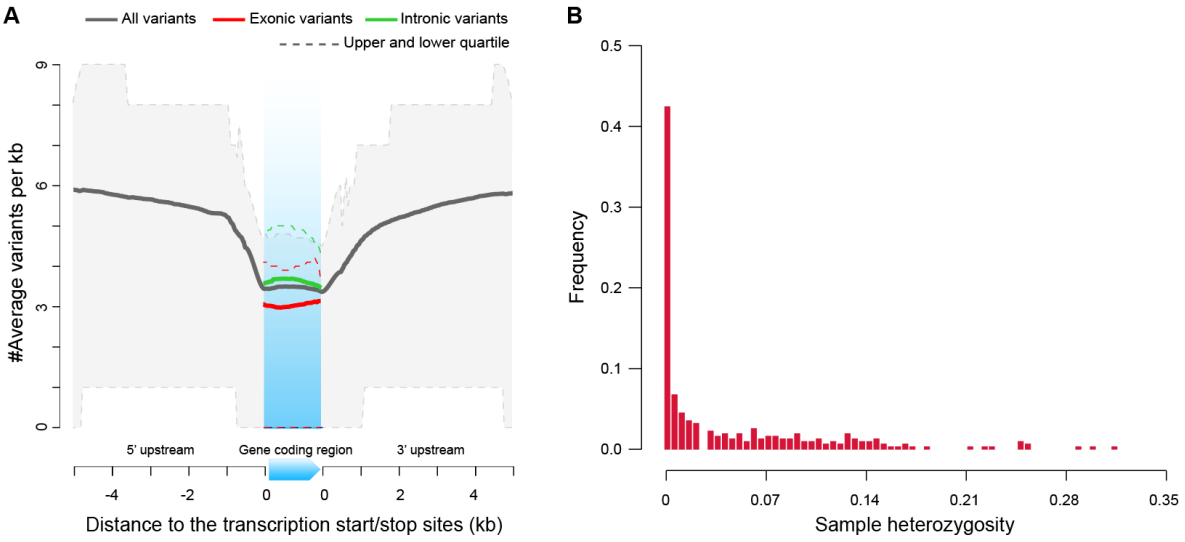
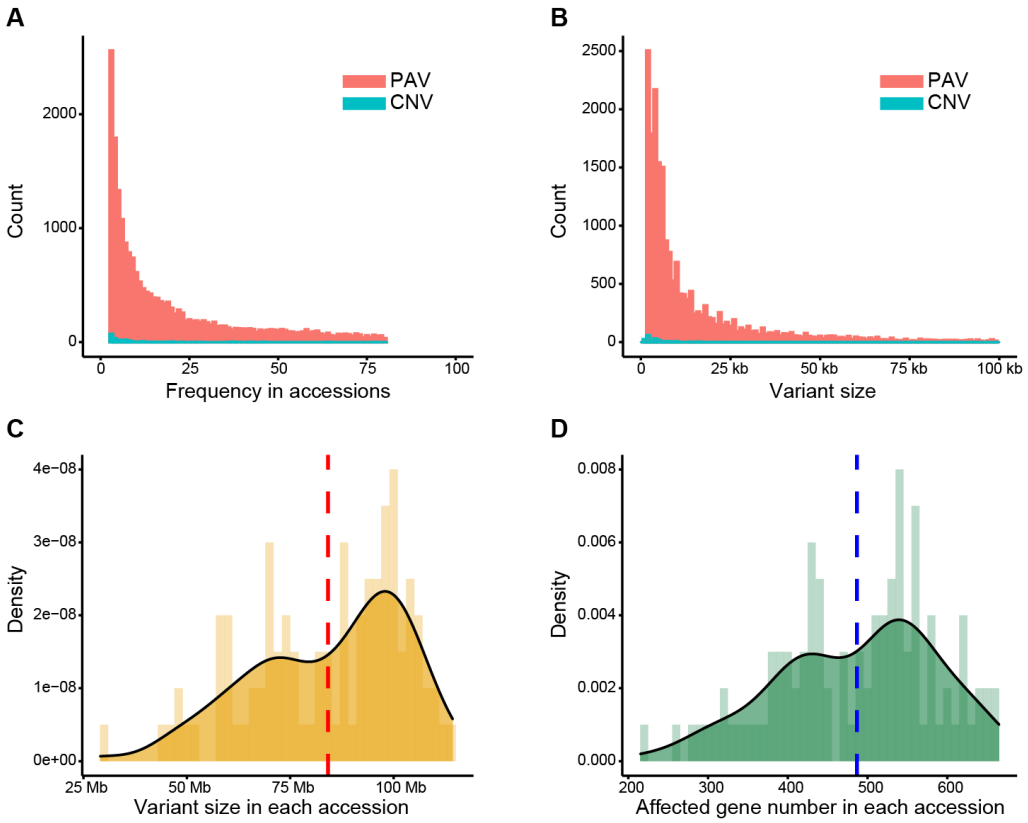


Fig. S1. (**A**) Density of the variants in the gene bodies and the 5 kb gene-flanking regions in the *C. annuum* population. (**B**) Frequency distribution of the sample heterozygosity in the resequenced pepper population.



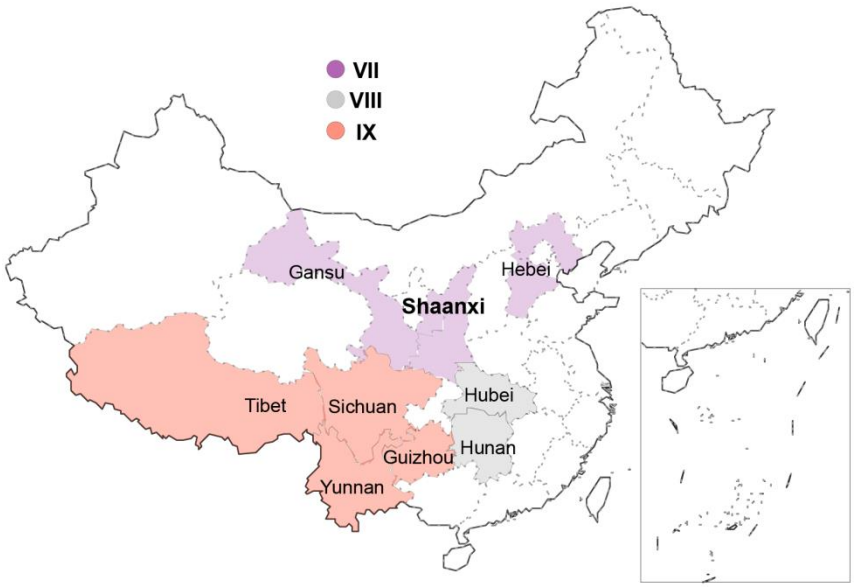
404

405

406 **Fig. S2**. Characteristics of the identified PAVs and CNVs in 100 pepper accessions. The numbers of
407 PAVs and CNVs with different frequency in 100 accessions (**A**) and with different size (**B**). Density
408 distributions of PAV/CNV size (**C**) and affected gene number (**D**) in each accession. The average size
409 and the affected gene number are indicated by the dashed lines.

410

411



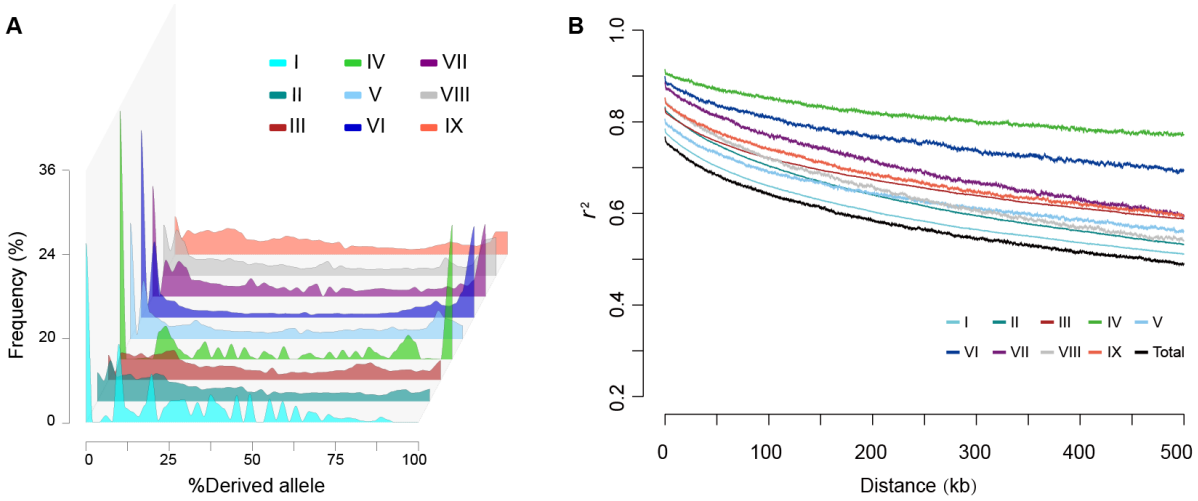
412

413

414 **Fig. S3.** Locations of China inland pepper groups VII, VIII, and IX in provinces from the north, middle
415 and southwest of China.

416

417



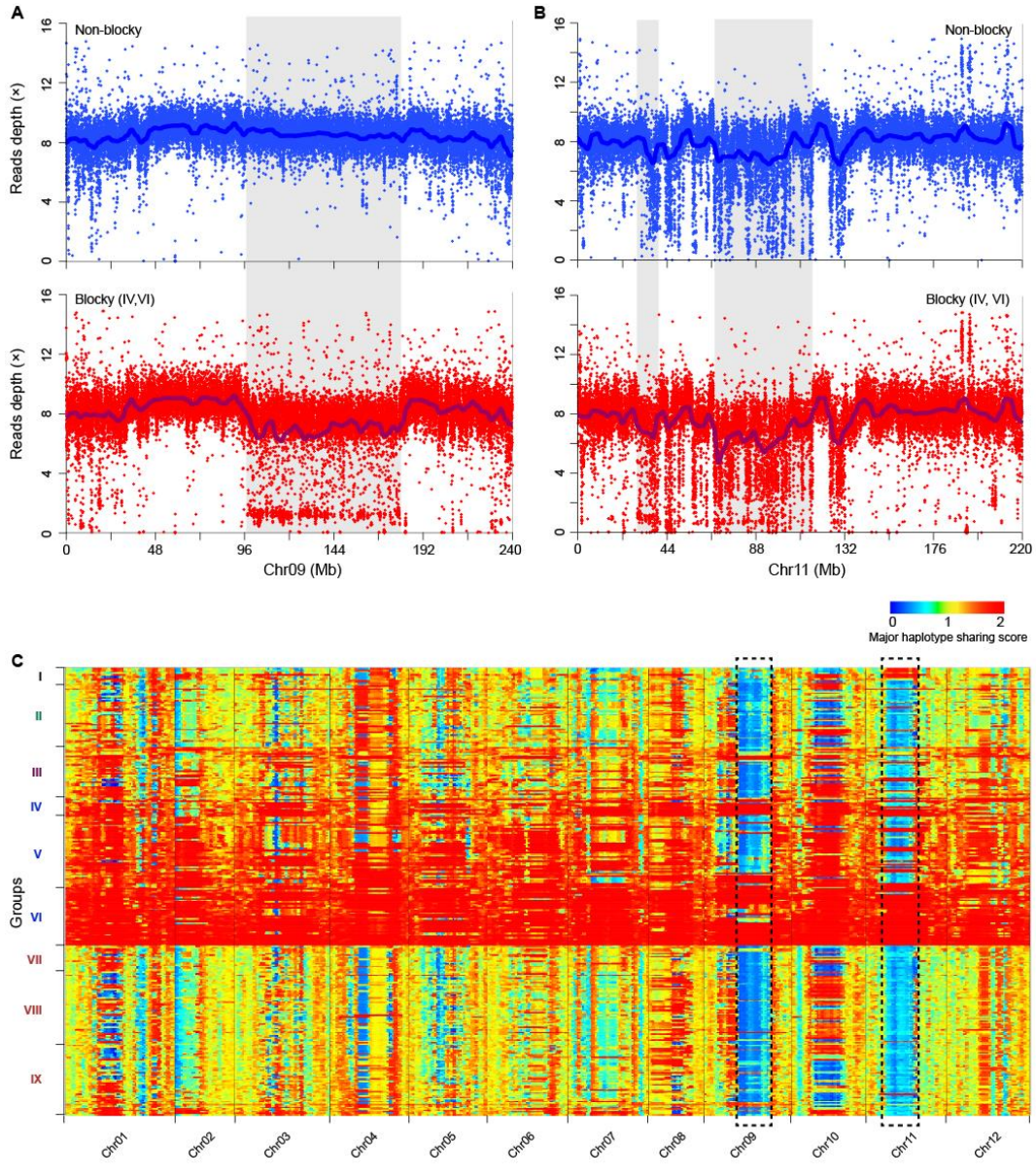
418

419

420 **Fig. S4**. (**A**) Frequency distribution of derived alleles of all SNPs in the nine pepper groups. (**B**) Decay
421 of genomic linkage disequilibrium (LD) in the total *C. annuum* population, and in the nine groups (I-
422 IX).

423

424



425

426

427 **Fig. S5.** The difference on depth of reads mapped to chromosomes 9 (**A**) and 11 (**B**) between the blocky
428 and non-blocky fruit peppers. Blue and red points indicate average reads depth in 10 kb windows, dark
429 blue lines denote average reads depth in 1 Mb window, and grey rectangles indicate the genomic regions
430 showing lower depth of reads mapped to blocky fruit peppers. (**C**) Genome-wide haplotype sharing
431 between blocky and non-blocky fruit peppers, with the major haplotype of the blocky fruit pepper as
432 the reference. The *y*-axis lists samples following the same order as shown in main-text **Fig. 1C-D**. The
433 heatmap colored from blue to red denotes low to high levels of major haplotype sharing score (MHS)
434 of each sample to the major haplotype of blocky fruit pepper in a 10 Mb window across the whole
435 genome.

436

437

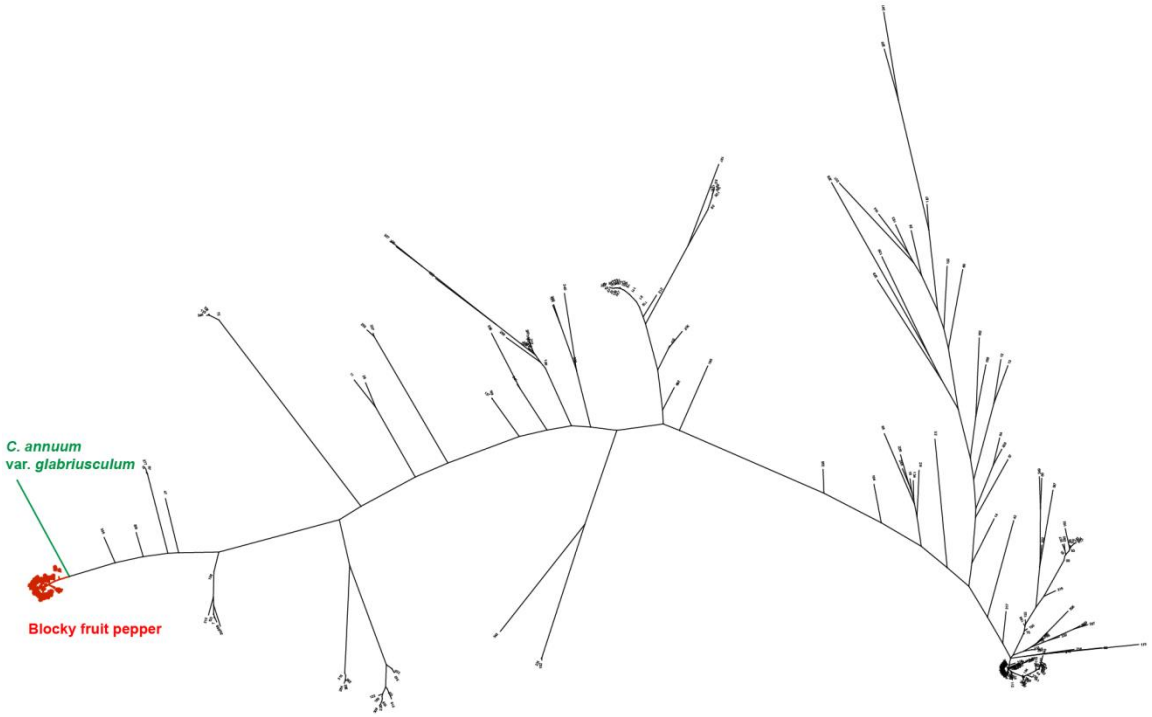


Fig. S6. Phylogenetic tree of the *C. annuum* population plus *C. annuum var. glabriusculum* whose genome has been released previously (Qin *et al.*, 2014), using genotypes of loci located at F9.

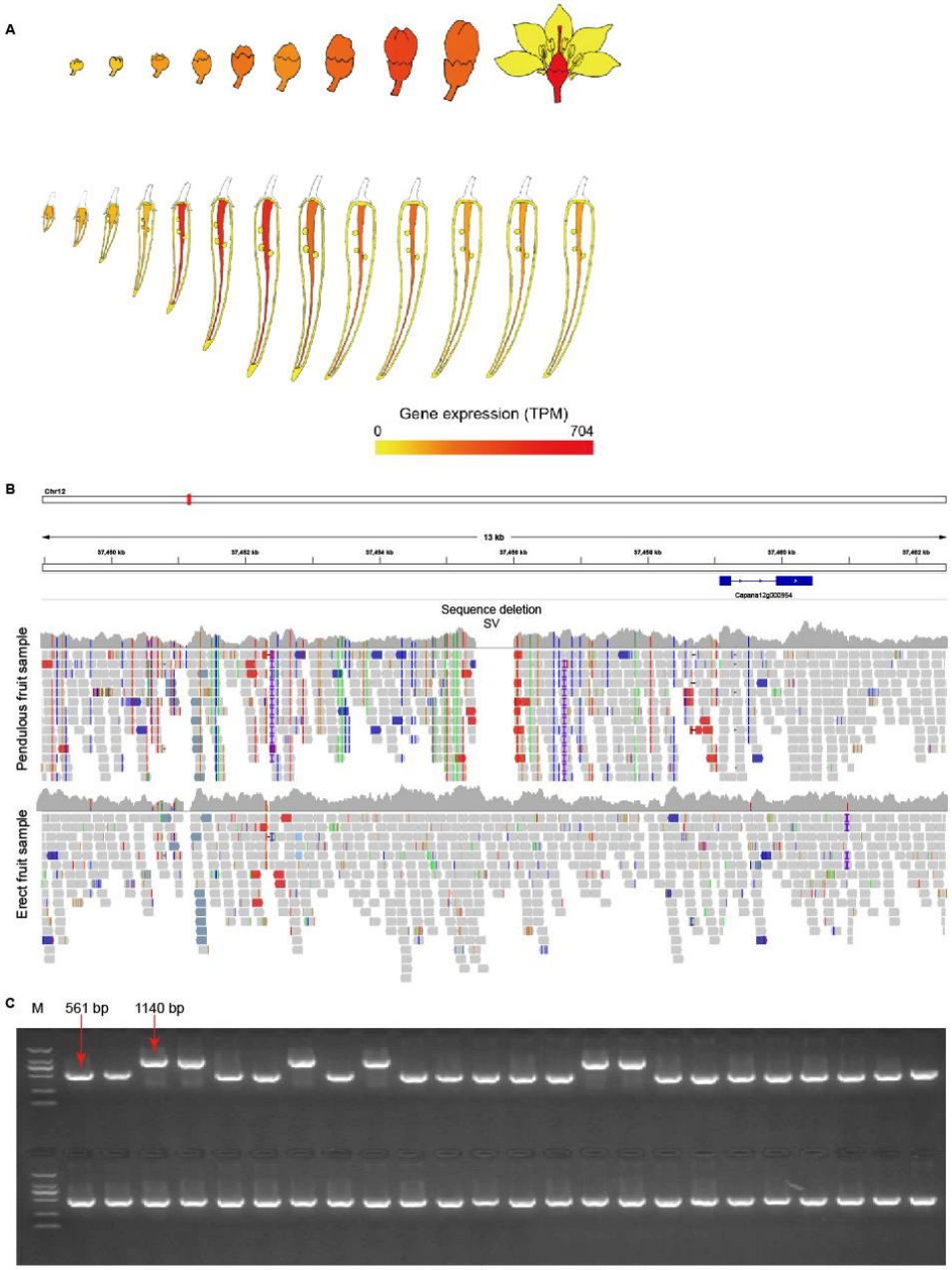


Fig. S7. (**A**) Gene expression of *Capana12g000954* (*Up*) in different developmental stages of pepper flower and fruit (Liu et al., 2017). (**B**) Visualization example of reads mapped onto the genomic region around gene *Capana12g000954* (*Up*) between samples with pendent (top) and erect (bottom) pepper fruits. The region that was not covered by any reads in the pendent fruit pepper indicates a large sequence deletion corresponding to the 579 bp deletion. (**C**) Electropherogram of the PCR products that cover the deletion upstream of gene *Capana12g000954* (*Up*) in different pepper samples. The 561 bp electrophoretic bands indicate samples with the 579 bp sequence deletion, while the 1,140 bp bands indicate samples without the sequence deletion. M: DNA marker DM2000.

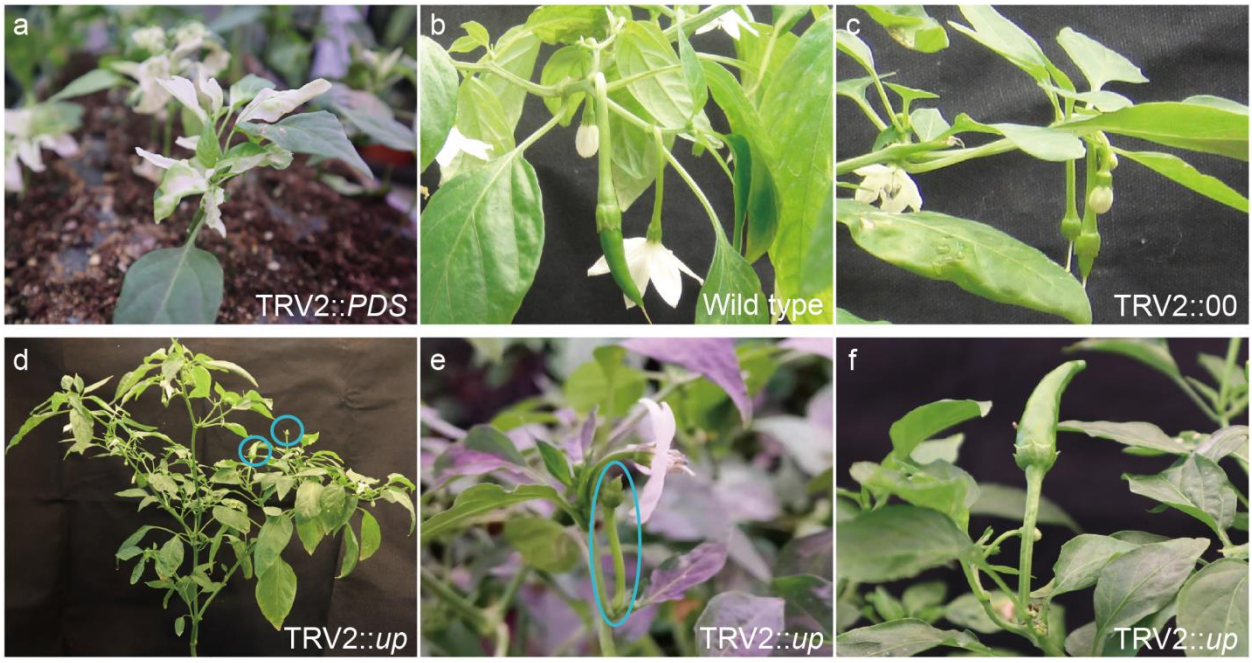


Fig. S8. The phenotypes of virus-induced gene silencing experiments on pepper accession "Changyang chili" with pendent fruit (wild type). (**a**) The TRV2::*CaPDS* pepper plant shows photobleaching; (**b**) Pendent fruit observed in the control plant (wild type); (**c**) Pendent fruit observed in plants treated with empty vector (TRV2::00); (**d**) Whole plant treated with TRV2::*up*, exhibiting with erect fruits; (**e**) Close-up of an erect flower and a small erect pepper fruit in a plant treated with TRV2::*up*; (**f**) Close-up of a large erect pepper fruit in a plant treated with TRV2::*up*.

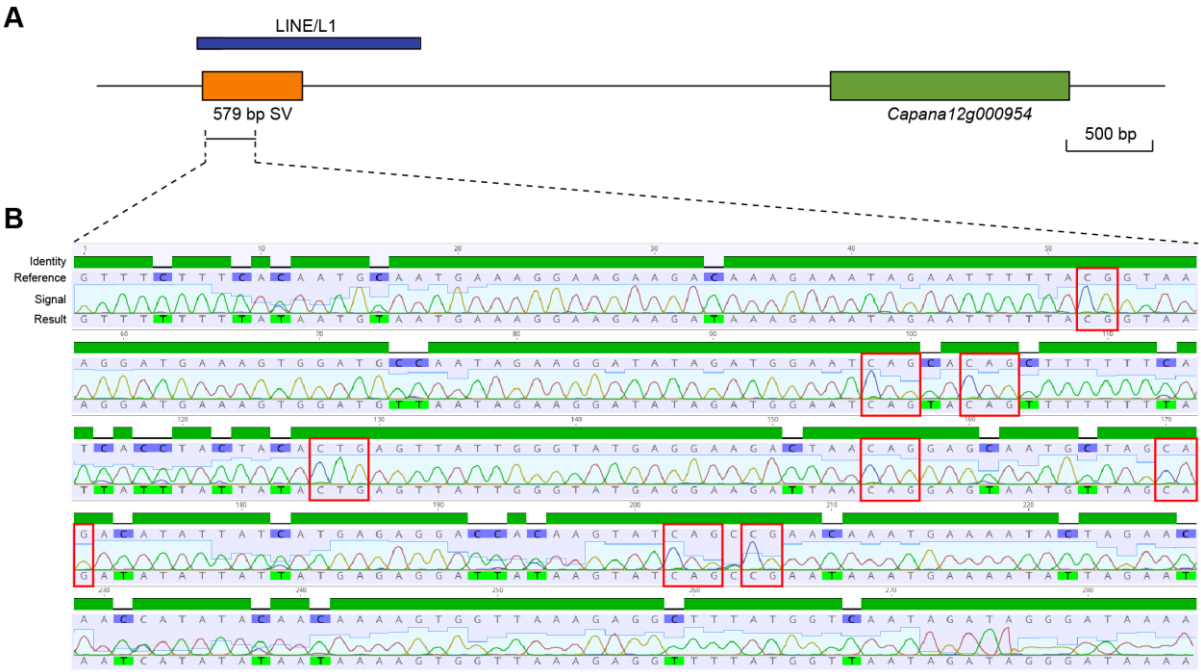


Fig. S9. DNA methylation status of the 579 bp SV in the promoter region of *Capana12g000954*. (**A**) Diagram illustrating the relative positions of the SV (orange), the overlapped TE (navy blue), and the *Capana12g000954* (green). The region bordered by dashed line was selected to do bisulfite sequencing to estimate the methylation status of the deleted sequence. (**B**) Comparison of the reference sequence and the bisulfite treated sequence with the un-methylated cytosine (C) converted into thymine (T). The Cs, especially those locating at the CpG and CHG sites (red box), in this region were heavily methylated, since they kept unchanged as Cs (purple signal) after bisulfite treatment.

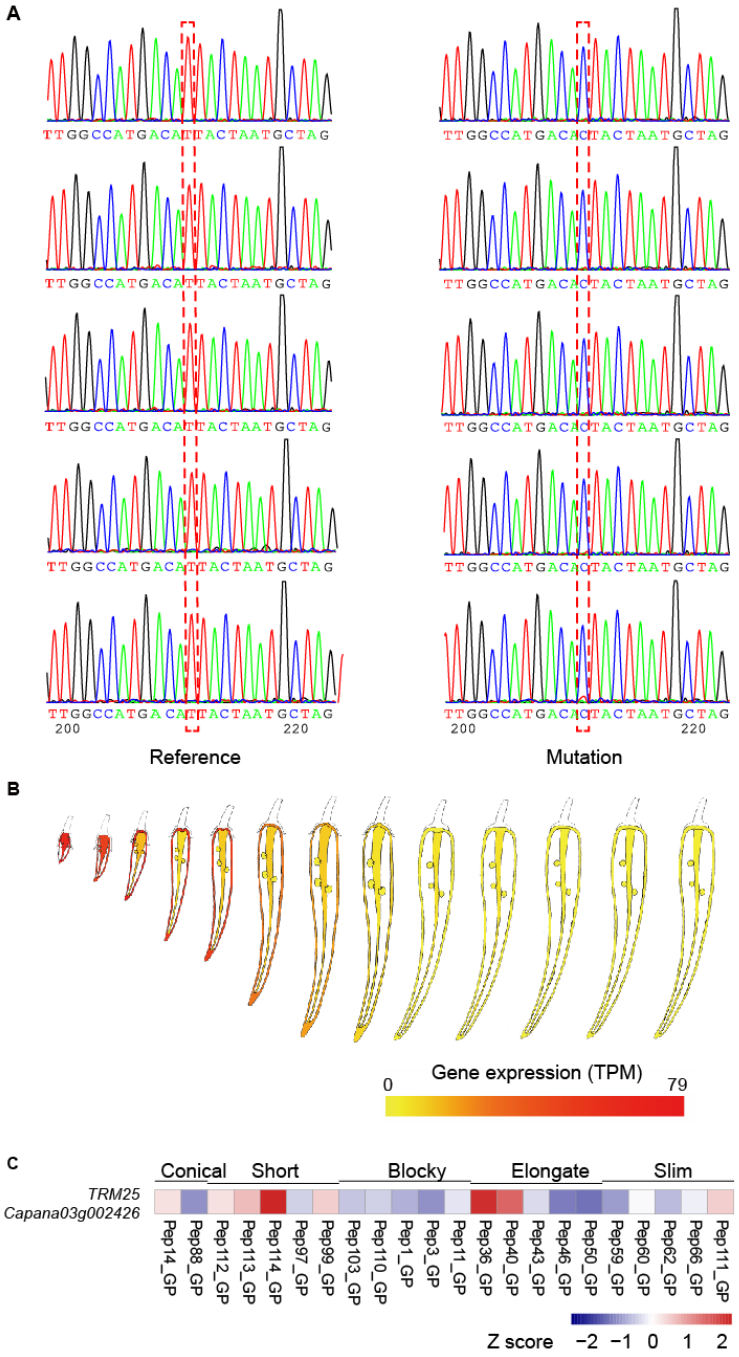


Fig. S10. (**A**) Sanger sequencing of the nonsynonymous mutation in *TRM25* in the test population. The red dashed-line rectangle denotes the position of the mutation. (**B**) Gene expression of *Capana03g002426* (*TRM25*) in different developmental stages of pepper fruit (Liu *et al.*, 2017). (**C**) Expression heatmap of *Capana03g002426* (*TRM25*) in the pericarp of 22 accessions with different fruit types evaluated through RNA-Seq.

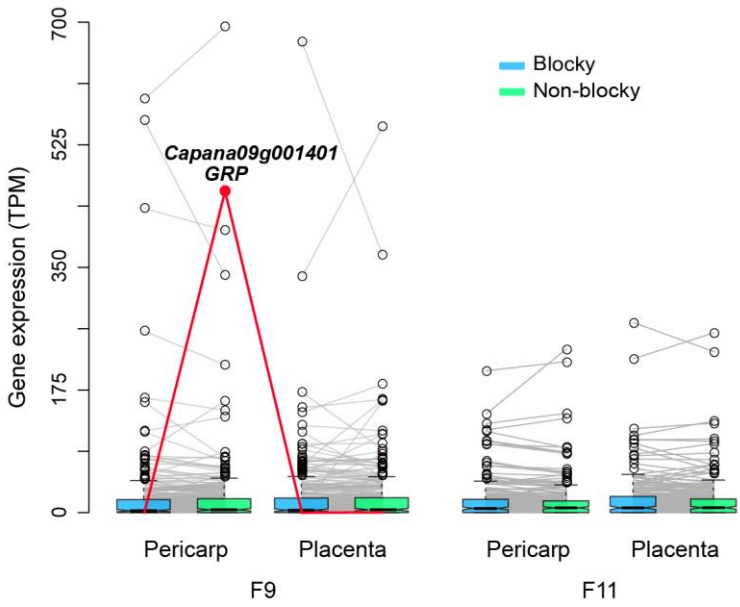


Fig. S11. Boxplot of gene expression for genes located at F9 and F11 in two pepper fruit tissues (pericarp and placenta) between blocky and non-blocky fruit peppers. The lines link the same genes in different boxplots; the red line indicates the gene *Capana09g001401* that is highly and specifically expressed in the pericarp of non-blocky fruit pepper. TPM: tags per million reads.

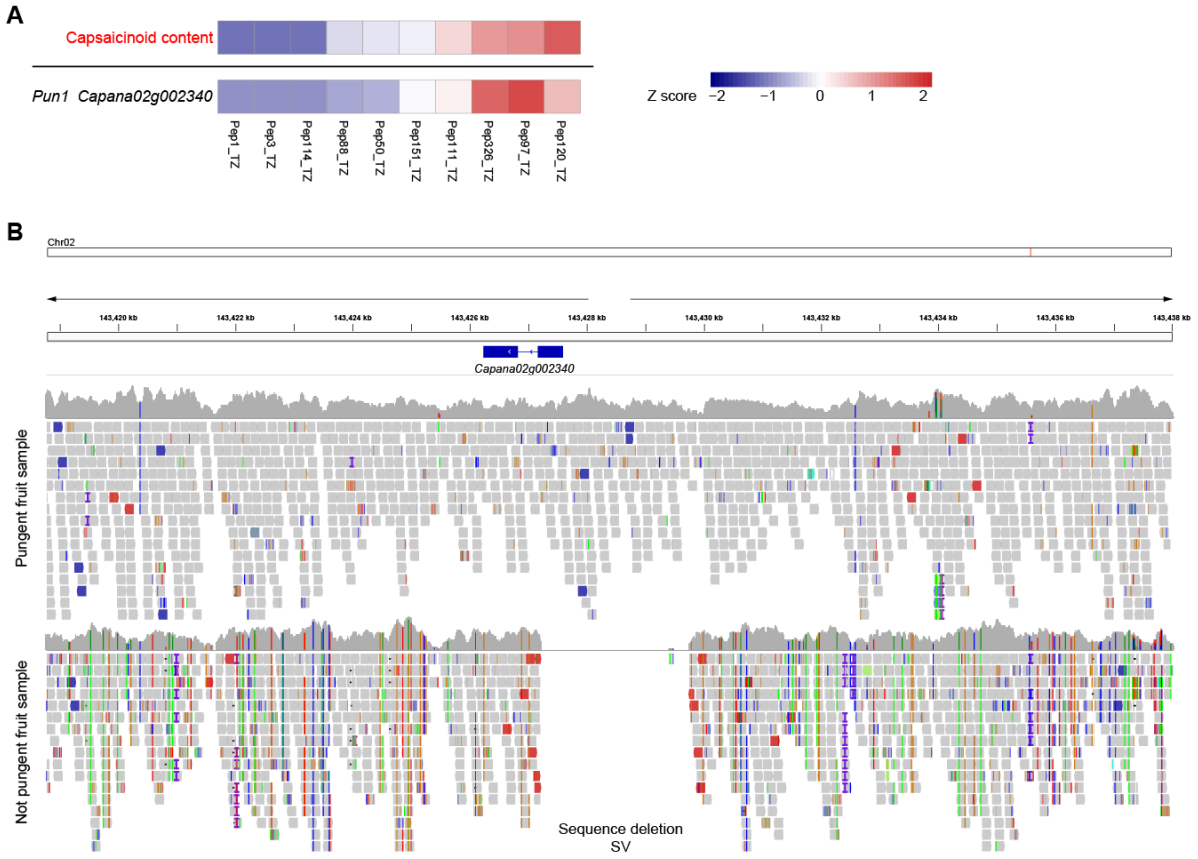


Fig. S12. (**A**) Expression heatmap of *Capana02g002340* (*Pun1*) in the placenta of 10 accessions with different capsaicinoid content evaluated through RNA-Seq. (**B**) Mapping of reads mapped onto the genomic region around gene *Capana02g002340* (*Pun1*) between peppers with pungent (top) and non-pungent (bottom) fruits. The region that was not covered by any reads in non-pungent fruit peppers, indicating a large sequence deletion.

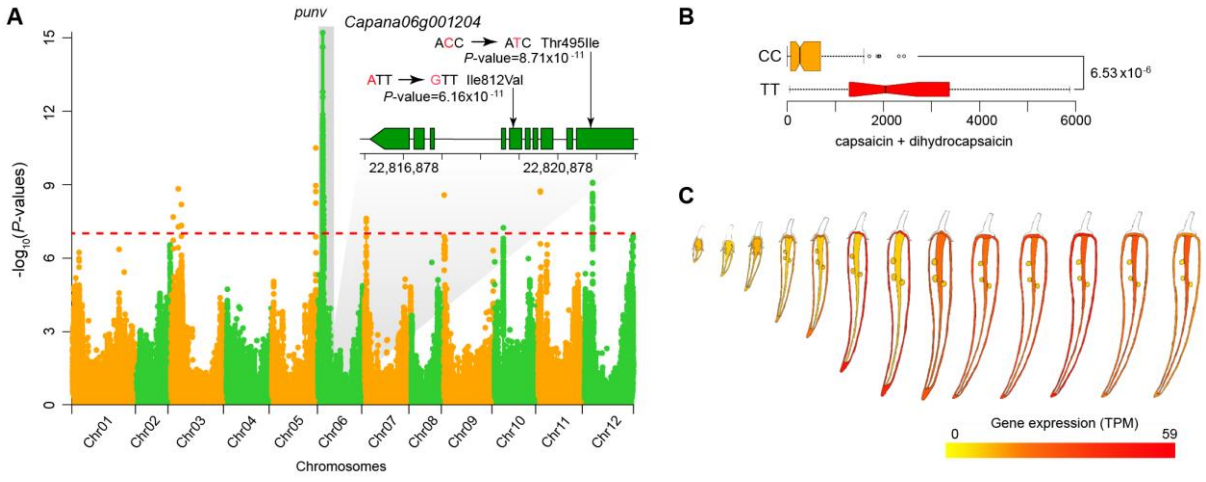


Fig. S13. The mapping of genomic loci associated with the pungency variation in the narrow-fruit pepper. (**A**) GWAS identified a locus *punv* that showed association signal of capsaicinoid content, and there were two (Thr495Ile and Ile812Val) mutations in the candidate gene *Capana06g001204*; (**B**) differences in capsaicinoid content in genotypes carrying the nonsynonymous mutations, in GWAS and test populations. (**C**) *Capana06g001204* expression during development of the pepper fruit.

SI References

Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome ResearchRes **12**:1805-1814. 10.1101/gr.631202.

Browning, B.L., Zhou, Y., and Browning, S.R. (2018). A One-Penny Imputed Genome from Next-Generation Reference Panels. The American Journal of Human Genetics **103**:338-348. 10.1016/j.ajhg.2018.07.015.

Edgar, R.C. (2004). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics **5**:113. 10.1186/1471-2105-5-113.

Gu, X.-z., Cao, Y.-c., Zhang, Z.-h., Zhang, B.-x., Zhao, H., Zhang, X.-m., Wang, H.-p., Li, X.-x., and Wang, L.-h. (2019). Genetic diversity and population structure analysis of Capsicum germplasm accessions. Journal of Integrative Agriculture **18**:1312-1320. https://doi.org/10.1016/S2095-3119(18)62132-X.

Guy St C, S., and Ewan, B. (2005). Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics **6**:31-31. 10.1186/1471-2105-6-31.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biology **9**:R7. 10.1186/gb-2008-9-1-r7.

Hoff, K.J., and Stanke, M. (2018). Predicting Genes in Single Genomes with AUGUSTUS. Current Protocols in Bioinformatics:e57. 10/gfpr44.
Kim, D., Paggi, J.M., Park, C., Bennett, C., and Salzberg, S.L. (2019). Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. Nat Biotechnol **37**:907-915. 10.1038/s41587-019-0201-4.
Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J., and Marra, M.A. (2009). Circos: an information aesthetic for comparative genomics. Genome Research **19**:1639-1645. 10.1101/gr.092759.109.
Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics **25**:1754-1760.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics **25**:2078-2079. 10.1093/bioinformatics/btp352.
Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics **30**:923-930. 10.1093/bioinformatics/btt656.
Liu, F., Yu, H., Deng, Y., Zheng, J., Liu, M., Ou, L., Yang, B., Dai, X., Ma, Y., Feng, S., et al. (2017). PepperHub, an Informatics Hub for the Chili Pepper Research Community. Molecular Plant **10**:1129-1132. 10.1016/j.molp.2017.03.005.
Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y.O., and Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res **33**:6494-6506. 10.1093/nar/gki937.
Pritchard, J.K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. Genetics **155**:945-959.
Qin, C., Yu, C., Shen, Y., Fang, X., Chen, L., Min, J., Cheng, J., Zhao, S., Xu, M., and Luo, Y. (2014). Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proceedings of the National Academy of Sciences of the United States of America **111**:5135-5140.
Retief, J.D. (2000). Phylogenetic analysis using PHYLIP. Methods in Molecular Biology **132**:243-258.
Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature **449**:913-918. 10.1038/nature06250.
Szpiech, Z.A., and Hernandez, R.D. (2014). selscan: an efficient multithreaded program to perform EHH-based scans for positive selection. Molecular Biology and Evolution **31**:2824-2827. 10.1093/molbev/msu211.
Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics **105**:437-460.
Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123**:585-595.
Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010a). KaKs_Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. Genomics Proteomics Bioinformatics **8**:77-80. 10.1016/S1672-0229(10)60008-3.
Wang, H.P., Li, X.X., and Song, J.P. (2018). Vegetable Genetic Resources in China.

Horticultural Plant Journal **4**:83-88. 10.1016/j.hpj.2018.03.003.
Wang, K., Li, M., and Hakonarson, H. (2010b). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res **38**:e164-e164.
Xu, M., Guo, L., Gu, S., Wang, O., Zhang, R., Peters, B.A., Fan, G., Liu, X., Xu, X., Deng, L., et al. (2020). TGS-GapCloser: A fast and accurate gap closer for large genomes with low coverage of error-prone long reads. GigaScience **9**:giaa094. 10.1093/gigascience/giaa094.
Xu, X., Liu, X., Ge, S., Jensen, J.D., Hu, F., Li, X., Dong, Y., Gutenkunst, R.N., Fang, L., Huang, L., et al. (2011). Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. Nature Biotechnology **30**:105-111. 10.1038/nbt.2050.
Yang, Y., Zhang, K., Xiao, Y., Zhang, L., Huang, Y., Li, X., Chen, S., Peng, Y., Yang, S., Liu, Y., et al. (2022). Genome Assembly and Population Resequencing Reveal the Geographical Divergence of Shanmei (Rubus corchorifolius). Genomics, Proteomics & Bioinformatics 10.1016/j.gpb.2022.05.003.
You, F.M., Huo, N., Gu, Y.Q., Luo, M.C., Ma, Y., Hane, D., Lazo, G.R., Dvorak, J., and Anderson, O.D. (2008). BatchPrimer3: a high throughput web application for PCR and sequencing primer design. BMC Bioinformatics **9**:253. 10.1186/1471-2105-9-253.
Zhang, X., Zhang, K., Wu, J., Guo, N., Liang, J., Wang, X., and Cheng, F. (2020). QTL-Seq and Sequence Assembly Rapidly Mapped the Gene *BrMYBL2.1* for the Purple Trait in *Brassica rapa*. Scientific Reports **10**:2328. 10.1038/s41598-020-58916-5.
Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nature Genetics **44**:821-824. 10.1038/ng.2310.