Genotyping pepper varieties using Target SNP-seq reveals that population structure clusters according to fruit shape

Heshan Du
Beijing Vegetable Research Center
Jingjing Yang
Beijing Vegetable Research Center
Bin Chen
Beijing Vegetable Research Center
Xiaofen Zhang
Beijing Vegetable Research Center
Jian Zhang
Beijing Vegetable Research Center
Kun Yang
Chinese Academy of Agricultural Sciences Institute of Vegetables and Flowers
Sansheng Geng
Beijing Vegetable Research Center
Changlong Wen (wenchanglong@nercv.org)

Research article

Keywords: Pepper, SNP, genetic structure, Target SNP-seq, association analysis

Posted Date: July 1st, 2019

DOI: https://doi.org/10.21203/rs.2.10821/v1

Abstract

Background The widely cultivated pepper (Capsicum spp.) is one of the most diverse vegetables; however, little research has characterized the genetic diversity and relatedness of commercial varieties grown in China. In this study, a panel of single-nucleotide polymorphisms (SNPs) was created that consisted of 97 perfect SNPs, which were identified using re-sequencing data from 35 diverse C. annuum lines. Based on this panel, a Target SNP-seq was designed that combined the multiplex amplification of the perfect SNPs with Illumina sequencing to detect polymorphisms across 271 commercial pepper varieties. Results The perfect SNPs panel had a high discriminating capacity due to the average value of polymorphism information content (PIC), observed heterozygosity (Ho), expected heterozygosity (He), and minor allele frequency (MAF), which were 0.31, 0.28, 0.4, and 0.31, respectively. Notably, the studied pepper varieties were morphologically categorized based on fruit shape; blocky, long horn, short horn, and linear-fruited. The long horn-fruited population exhibited the most genetic diversity followed by the short horn, linear, and blocky-fruited populations. A set of 35 core SNPs were then used as KASPar markers, another robust genotyping technique for variety identification. Analysis of genetic relatedness using principal component analysis (PCA) and phylogenetic tree construction indicated that the four fruit shape populations clustered separately with limited overlaps. Based on STRUCTURE clustering, it was possible to divide the varieties into five subpopulations, which correlated with fruit shape. Further, the subpopulations were statistically different according to a randomization test and Fst statistics. Notably, two SNP loci, CaSNP118 and CaSNP053, which are located on chromosome 11 and 6 were significantly associated with fruit shape ($p < 1.0 \times 10^{-4}$) Conclusions Target SNP-seq developed in this study appears as an efficient power tool to detect the genetic diversity, population relatedness and molecular breeding in pepper. Moreover, this study demonstrates that the genetic structure of the pepper varieties is significantly influenced by breeding programs focused on fruit shape.

Background

Pepper are members of the genus *Capsicum*, which originated in South America and represents one of the most economically important vegetable crops worldwide [1-3]. To date, 38 species of *Capsicum* have been reported (USDA-ARS, 2011). Of these, *C. annuum*, *C. frutescens*, *C. chinense*, *C. baccatum*, and *C. pubescens* are thought to have been domesticated [4]. Globally, the most predominant species is *C. annuum*, which has numerous commercial varieties varying greatly in size, shape, pungency, and color.

As the seed trade has developed and globalized, the commercial quality of seeds, which is based on authenticity and purity, has become increasingly important [5]. Traditionally, cultivar characterization was completed by field investigation of morphological traits; however, this process is time-consuming and labor-intensive and is thus not suitable for modern inspection demands [6]. A more high-throughput approach to distinguishing varieties is the used of molecular markers [5]. Indeed, genetic markers have been used for DNA fingerprinting, diversity analysis, variety identification, and marker-assisted breeding of multiple commercial crops [7, 8]. Moreover, several PCR-based tools have been used to detect genetic

diversity in peppers, including random amplified polymorphic (RAPD), restriction fragment length polymorphism (RFLP), and amplified fragment length polymorphism (AFLP) [9-12].

Recently, the genomes of two *C. annuum* cultivars, Zunla-1 and CM334, were sequenced [3, 13], which provided an important platform for the detection and development of genome-wide simple sequence repeats (SSR) and insertion or deletion (InDel) markers [14-16]. Although a large number of SSR and InDel markers have become available, these technologies are not suitable for large scale germplasm characterization. Thus, there is an unmet need for an efficient, rapid, and high-throughput system capable of characterizing thousands of germplasm.

One approach for meeting such high standards is the use of single nucleotide polymorphisms (SNPs), which are good markers for genotyping because of their whole genome coverage and primarily biallelic nature. Accordingly, multiple high-throughput SNP genotyping platforms have been developed, including the GoldenGate [17] and Infinium [18], TaqMan [19], and KASPar platform (KBiosciences, www.kbioscience.co.uk). However, SNP marker genotyping is considered expensive as it requires a comprehensive technical platform, equipment, and reagents.

Genotyping by target sequencing (GBTS) is a targeted sequence-capture strategy that is able to genotype more than thousands of SSRs or SNPs through the use of high throughput-sequencing technology. The two main types of GBTS are multiplex PCR and probe-in-solution-based target sequencing; the technology has been commercialized as AmpliSeq [20], NimbleGen [21], SureSelect [22], GenoBaits, and GenoPlexs [23]. To date, this technology has been widely used for medical applications but has rarely been used for agriculture species. However, a Target SSR-seq technique, which is a multiplex PCR-based approach, was successfully applied to the study of genetic diversity and structure in 382 cucumber varieties [24]. The results of this study demonstrated that GBTS is a customizable, flexible, high-throughput, low cost, and accurate sequencing tool.

Until now, the genetic diversity of domesticated *Capsicum* species has primarily been investigated using SSR markers [25-27], and genetic maps have been constructed with SSRs and InDels based on inter- or intraspecific populations [15, 16, 28]. However, discovery efforts for genome-wide SNPs have lagged significantly behind those for SSRs and InDels, and studies on genetic diversity of pepper varieties are limited. The objectives of the present work were: 1) to develop a Target SNP-seq technique suitable for genotyping pepper varieties; 2) to characterize composite core-SNP markers for use with the KASPar platform to maximize variety identification; 3) to examine the level of genetic diversity, structure, and differentiation within 271 pepper varieties. This study demonstrated that a novel Target SNP-seq can be used as a rapid and efficient tool for genotyping peppers, and the genetic structure of these cultivated varieties have been strongly impacted by breeding programs that select for fruit shapes.

Results

Genome-wide perfect SNPs used for Target SNP-seq

Re-sequencing of the 31 pepper lines in this study generated a total of 872 Gb of paired-end sequence data, at an average depth of ~ 8.4. Following mapping to the Zunla-1 genome, 3,613,192 high-quality SNPs were detected across the genomic sequences of the 31 re-sequenced lines and four previously published cultivars (Dempsey, Zunla-1, Perennial, and Chiltepin). Using the cultivar progenitor Chiltepin as an out-group, the phylogenetic tree showed that pepper lines could be generally be classified according to fruit shapes, with the exception of three long horn-fruited lines that grouped with the linear-fruited lines. Based on the genetic distance, the transition in fruit shapes were from Chiltepin-like peppers followed by linear-fruited, short horn-fruited, long horn-fruited, finally to blocky-fruited peppers, which were the furthest from the Chiltepin-like peppers (Fig. 1A). Furthermore, the 35 lines can be divided into two major groups based on the optimal number of *K*=2 by STRUCTURE (Fig. 1B); Group 1 consisted of the nine bell-fruited lines and 10 of the long horn-fruited lines, whereas the remaining peppers, including three long horn, all the linear, and all the short horn-fruited, as well as PI640446, Perennial, and Chiltepin were assigned to Group 2.

Given that pepper genomes are highly repetitive, strict criteria were used to identify the perfect SNPs. In total, 521 perfect SNPs were identified, and 92 that were evenly distributed across the genome (Fig. 2; Table S2) were selected as multiplex PCR targets.

Genotyping analysis of pepper varieties using Target SNP-seq

In total, 288 pepper varieties were genotyped using the Target SNP-seq. Samples that were missing more than five of the 92 loci were removed from the analyses. The final panel contained 271 varieties, including 90 blocky, 113 long horn, 25 short horn, and 43 linear-fruited varieties (Table S1). A total of 55.9 million reads were generated from the 271 varieties, and the average Target read depth was 2064, approximately 82% of the samples were sequenced at a depth greater than 1000 × (Fig. S2A). Among the 271 varieties, 238 varieties (87.8%) aligned to the Zunla-1 genome at a rate of more than 90% (Fig. S2B). Of these aligned reads, 221 varieties (81.5%) exhibited an align rate to the target SNP region of over 80% (Fig. S2C). Furthermore, the Target SNP-seq uniformity index was analyzed, which was used to calculate the proportion of the coverage above 10% of the mean depth value for each variety. The average uniformity index in this study was 89.5% (Fig. S2D), indicating a high level of accuracy.

Perfect SNPs in 271 pepper varieties

The genetic parameters, MAF, *Ho*, *He*, and PIC revealed by each perfect SNP are given in Table S3. MAF is a measure of the discriminating ability of the markers; as such, the closer the MAF is to 0.5 for biallelic markers, the better discriminatory properties. In this study, 28.26% of perfect SNPs showed an MAF between 0.4 and 0.5, whereas only four SNP had MAF below 0.1 (Fig. 3A). The *Ho* value of each SNP ranged from 0.01 (CaSNP079) to 0.59 (CaSNP009) with an average of 0.28, and 11 SNPs exhibited higher *Ho* (>0.4) (Fig. 3B; Table S3). Furthermore, the *He* values ranged from 0.01 (CaSNP079) to 0.5 (CaSNP043 and CaSNP094) (Fig. 3C; Table S3), whereas PIC values varied among perfect SNPs from 0.01 (CaSNP079) to 0.38 (CaSNP043, CaSNP094 and CaSNP117) with a mean of 0.31 (Fig. 3D; Table S3). 71.74 % of the perfect SNPs had PIC values greater than 0.30, whereas only four SNPs showed PIC values below 0.2. These values indicate that the perfect SNPs panel has a high discriminating capacity for varieties, and CaSNP043, CaSNP94, CaSNP117, and CaSNP009 were the best at discriminating between varieties. Overall, the results indicate that the Target SNP-seq can be used as a rapid tool for genotyping peppers.

Perfect SNPs across the fruit shapes

The average values of the genetic parameters across the four fruit shape populations were also compared for genetic diversity, and the results showed that the blocky-fruited population had the lowest average values for *He* (0.18), *Ho* (0.16), and PIC (0.15) (Table 1), indicating the lowest genetic diversity within this population. In contrast, the long horn-fruited population exhibited the most genetic diversity as defined by the highest average values of *He* (0.39), *Ho* (0.36), and PIC (0.31).

Identification of a core SNP Set

The perfect SNPs panel distinguished 97.7% of the 271 pepper varieties (Fig. 3), the remaining displayed the same multilocus genotypes that were also difficult to distinguish from field phenotypes. Given that some varieties may exist with multiple names, the varieties with the same genotypes may be redundant and were discarded to build non-redundant genotype varieties. Thus, a minimum of 27 of the perfect SNPs could distinguish between all the non-redundant varieties (Fig. 3).

To develop a core SNP set for the KASPar platform, each perfect SNP marker was tested on a set of 23 pepper varieties with two allele-specific forward primers and one common reverse primer. The results show that 35 SNP primers produced consistent and repeatable results with Target SNP-seq. Finally, 35 SNPs with a high discrimination power of up to 97% across all varieties and 100% in non-redundant varieties were proposed as a core SNPs set for use with the KASPar platform (Fig. 3; Table S4).

Genetic structure in pepper varieties

PCA was performed using the 92 perfect SNPs to investigate population clusters across the 271 varieties (Fig. 4A). Accordingly, the PCA plot indicates that the four fruit shape populations generally clustered separately. The distribution of blocky-fruited varieties was very concentrated, whereas that of the long horn-fruited varieties was relatively dispersed. Linear-fruited varieties showed more affinity to the short horn than either long horn or blocky-fruited varieties. Linear and blocky-fruited populations were the most diverse, and these clusters did not overlap, suggesting considerable genetic divergence over their breeding programs. Notably, a selection of both long horn- and short-fruited varieties showed affinities to the linear-fruited population.

The population structure of the 271 varieties was further inferred using the cluster program, STRUCTURE, through gradually increasing the number of clusters (*K*). The Evanno's correction [29] showed the peak of delta *K* at *K*=2, which suggests the presence of two main populations, denoted as Pop1 and Pop2. Pop1 comprised 160 varieties (59.0%), containing all blocky-fruited varieties, 60.2% of the long horn, and only two linear-fruited varieties (Fig. 4B; Table S1). The remaining 111 varieties (41.0%) were assigned to Pop2, which included all the short horn-, and linear-fruited varieties, as well as 39.8% of long horn-fruited varieties (Fig. 4B; Table S1). When *K* = 3, the Pop1 was subdivided into two clusters, blocky or long horn-fruited types. At *K* = 4, a mixture of 56% of short horn-fruited, 15 long horn and two liner-fruited varieties were assigned to a new cluster from Pop2, and these short horn-fruited, as well as a new long horn-fruited group, were assigned to independent clusters, respectively, when *K* = 5. Of note, linear-fruited types were never assigned to an independent cluster as *K* was increased. Considering the classification of populations appeared highly correlated with fruit types when *K* = 5, the two main populations were further subdivided into five subpopulations (Subpop1~Subpop5; Fig. 4B; Table S1). Subpop1, 2, 3, and 4 displayed a clearly cut structure with no or very few admixtures. Subpop1 comprised 98 varieties, 90 of which belong to blocky-fruited and the remaining eight long horn-fruited varieties. Long horn-fruited varieties were members of both Subpop2 and Subpop3, which is not surprising as long horn-fruited varieties were distributed across both Pop1 and Pop2. Subpop2 comprised 44 long horn-fruited varieties. Subpop3 comprised 24 varieties, 22 of which were long horn-fruited and the remaining two linear-fruited varieties. Subpop4 comprised 14 short horn-fruited varieties. Consistent with the results from PCA analyses, admixtures are mostly located in Subpop5, which contained 41 long linear-fruited varieties, as well as a minority of short horn- and long horn-fruited varieties.

The unrooted phylogenetic tree (Fig. 4C) is consistent with the aforementioned PCA and model-based population structure and indicated a clear distinction in the four fruit shapes, despite having admixtures. Images of the representative varieties, which were selected based on the lowest average genetic distance to other varieties within corresponding subpopulations, are presented in Fig. 4C. The representative images for two long horn-fruited varieties from Subpop2 and Subpop3 clearly indicate distinct morphologies.

In summary, three independent analysis methods strongly support the division of pepper varieties into five well-differentiated genetic populations, which correlated with distinct fruit shapes, indicating that the

genetic structure of these cultivated varieties may have been strongly impacted by the fruit shape selection of the breeding programs.

Genetic variation assessment of pepper populations

Comparison of the results between Pop1 and Pop2 using AMOVA revealed that 33.04% of the total genetic variation was partitioned among Pops, 8.47% within Pops, and the remaining 58.49% within varieties (Table 2). AMOVA analysis of the five Subpops further indicated that the maximum variation (63.83%) occurred within varieties, the minimum variation (3.54%) was accounted for within Subpops, and 32.63% of the variation occurred between Subpops (Table 2), suggesting relatively moderate differentiation among Subpops.

To test for significant variation between Pops and among Subpops, a randomization test and pairwise F_{st}estimation were performed. From the output, we can see four histograms representing the distribution of the randomized strata (Fig. 5). The observed results in the output show significant differentiation of the structure of Pops and Subpops considering all levels of the Pops and Subpops strata (Fig. 5). These results also support the separation of the varieties into two Pops and five Subpops. Furthermore, pairwise estimates of F_{st} demonstrated that population differentiation between Pop1 and Pop2 is high (F_{st} = 0.35). The pairwise F_{st} between the five Subpops ranged from 0.13 between Subpop2 and Subpop3 (both consist largely of long horn-fruited) to 0.48 between Subpop1(mostly blocky-fruited) and Subpop4 (short horn-fruited) (Table 3). Notably, a high genetic differentiation (F_{st} = 0.43) was also shown between Subpop1 and Subpop5 (mostly consisting of linear-fruited). Subpop4 also had very low genetic differentiation from Subpop5 (F_{st} = 0.14).

Identification of SNPs associated with fruit shape

A total of 21 SNP loci did not indicate any diversity (PIC = 0) within certain fruit populations, of which 16, 1, 3, and 5 loci were for blocky, long horn, short horn, and linear-fruited population, respectively (Table S3). These fruit shape-specific loci may have been under selection during breeding or were selected due to linkage with genes that are determinative of fruit traits. Using the MLM *K*+Q model and Bonferroni correction for *p*-value, two SNP loci, CaSNP118 and CaSNP053, were identified as significantly associated with fruit shape across 271 pepper varieties ($p < 1.0 \times 10^{-4}$). To match the associations with previously identified quantitative trait loci (QTL), the physical position of the two SNP loci in the both the reference genome of Zunla-1 and CM334 are given in Table 4. The two associated SNP loci were located on chromosomes 11 and 6, respectively; and contributed 13.6% and 9.2% of the phenotypic variation, respectively. In total, 252 and 1217 annotated genes were identified in the associated region of CaSNP118 and CaSNP053, respectively (Table S5).

Discussion

High-throughput genotyping by Target SNP-seq

High-throughput genotyping technology has become essential for effective crop breeding programs. Target SSR-seq, which combined the multiplexed amplification of perfect SSRs with high-throughput sequencing, was recently developed and applied to the identification of cucumber varieties, leading to the characterization of a set of core SSRs [24]. This sequencing technology can acquire thousands of data points in under 72 hours, costs less than $7/sample, and is associated with genotyping accuracy up to 100% due to the high coverage. In this study, re-sequencing tools were used to identify 92 perfect SNPs from the genomes of 35 *C. annuum.* The identified SNPs were then used for Target SNP-seq to assess genetic diversity across 271 pepper varieties that are popular in China.The results show that the perfect SNPs panel has a high discriminating capacity for varieties as 71.74 % of the perfect SNPs had PIC values > 0.30 (Table S3). Further, a minimum of 27 of the perfect SNPs could distinguish between all the non-redundant varieties (Fig. 3). Notably, the mean PIC value was found to be 0.31, which is low in comparison to the values derived from studies using SSR markers [27, 30]. These discrepancies may be explained by considering the nature of the different types of markers; SSRs are multiallelic and more polymorphic than SNP markers, which are biallelic [31].

A set of 35 core SNPs that had the same discrimination power as the 92 perfect SNPs was successfully converted into KASPar markers, representing another robust genotyping choice for pepper varieties (Fig. 3; Table S4). Unlike SSR markers, SNP markers do not require reference cultivars be included in each experiment and will also overcome the confusion between labs with regard to SSR alleles.

Population structure among inbred *C. annuum* lines

Since their initial domestication in Mexico, peppers have been under strong selection for fruit shape and size [32]. Consumption habits and pepper type preference vary globally. In the US alone, more than 20 market types are recognized and consumed [33]. In China, a majority of the pepper varieties commercially cultivated belong to the species *C. annuum*, and the market types are classified by fruit shapes, such as the popular blocky, long horn, short horn, and linear-fruits [34, 35]. To date, most experiments have evaluated the genetic relationships among several *Capsicum* species [36, 37] or the genetic diversity of *C. chinense* germplasm from relatively restricted regions [38]. The phylogenetic analysis based on molecular markers and pan-genome confirmed that *C. chinense* and *C. frutescens* are more closely related to each other than to *C. annuum* [37, 39]. Only a few studies have attempted to characterize the population relatedness in cultivated *C. annuum* [31, 40, 41]. Notably, the relationships among the 35 re-sequenced *C. annuum* lines described in this study align with previous reports grouping *C. annuum* according to fruit traits [40]. Further, clustering of the blocky-fruited varieties in the most derived positions relative to small hot Chiltepin-like types can also be observed in previous studies [40, 41].

Genetic structure among *C. annuum* varieties

Although previous work has demonstrated that a small population of *C. annuum* lines clustered according to the fruit shapes, the relationships among the commercially important *C. annuum* varieties from different companies have remained unclear. In the present study, the relationships among four fruit shape populations were assessed across a broad range of pepper varieties cultivated in China. Comparison of the genetic parameters showed the lowest *Ho* was observed within the blocky-fruited population, while the highest was detected in the horn-fruited population (Table 1). These findings agree with the earlier studies that found a reduction in diversity was associated with non-pungent blocky-fruited lines relative to pungent lines [41-43]. The narrow genetic diversity associated with the blocky-fruited varieties may be a consequence of inbreeding with a limited gene pool.

Additionally, the PCA and phylogenetic tree demonstrated that the four fruit shape populations clustered separately with a little or no overlap. This aligns with the fruit shape classification system and demonstrates that the genetic structure of pepper varieties in China has been significantly influenced by breeding programs that select for fruit shape. Similarly, STRUCTURE analysis grouped the varieties into two main populations, Pop1 and Pop2, which were further divided into five subpopulations, Subpop1 to Subpop5 (Fig. 4B). Moreover, the subpopulations correlated with fruit shape. Notably, Subpop1, Subpop4, and Subpop5 corresponded to the blocky, short horn, and linear-fruited varieties, respectively. However, the majority of the long horn-fruited varieties were divided into two subpopulations, Subpop2 and Subpop3, which were statistically unique (Fig. 5)

Identification of associated loci and candidate genes for fruit shape

Fruit shape is a complex trait controlled by multiple genes. QTL analyses have been previously used to study fruit shape trait by predicting the regions of the genome that affect the trait and estimating the effect of each region. The first fruit-shape QTL in peppers, named *fs10.1*, was detected on linkage group 3 [44]. Recently, Chunthawodtiporn et al. [45] detected one fruit shape QTL, which was located at 85.3 cM on chromosome 2 and accounted for 14.9% of the phenotypic variation. In the present study, two novel SNP loci, CaSNP118 and CaSNP053, were significantly associated with fruit shape and were located on the end of chromosome 11 and 6, respectively. Combined, these two SNPs account for a total of 22.8% of the phenotypic variation associated with fruit shape (Table 4).

Within the *C. annuum* genome, CaSNP118 is located between SNP loci CaSNP098 and CaSNP095, which covers approximately 44 Mb and has 252 annotated genes; CaSNP053 is located between SNP loci CaSNP054 and CaSNP052, which covers approximately 30 Mb and has 1217 annotated genes (Table S5). Therefore, many predicted genes could be considered as candidate fruit-shape controlling gens based on their predicted roles in protein function. For example, CA11g12200 is calmodulin-domain

protein kinase gene, CA06g15400 and CA06g21580 are both Ovate family proteins, and CA06g20780 is a homeodomain-like superfamily protein. Each of these genes is a candidate for fruit shape determination in tomato. Ovate family proteins control fruit shape transformation from round to pear-shaped fruit, WUSCHEL encodes a homeodomain transcription factor that controls meristem size and locule number, and SUN encods a member of the IQD family of calmodulin-binding proteins that leads to fruit elongation [46, 47].

Future directions of Target SNP-seq in pepper

Of note, SSR or InDel loci that are suitable for specific primer design could be added to the perfect SNP panel used in our Target SNP-seq. For example, we demonstrated that disease resistance markers, such as resistance genes for *Tobamovirus* [48], *Phytophthora capsici* [49], bacterial spot [50], and *potato virus Y* [51], can be successfully added to the perfect SNP panel for Target SNP-seq (data not shown). The commercial application of this technique has the potential to increase the efficiency of marker-associated selection programs, as well as aiding in variety identification.

Conclusion

The Target SNP-seq developed in this study is a high-throughput and reliable tool for the investigation of genetic diversity, variety identification, and characterization of population structure in peppers. The use of PCA, phylogenetic tree generation, and STRUCTURE demonstrates that the genetic structure of commercially available pepper varieties in China has been significantly influenced by breeding programs that select for fruit shape. Finally, association analysis of a limited number of SNPs allowed for the identification of novel genomic regions and candidate genes that control fruit shape.

Methods

Plant materials, fruit shape categorization, and DNA extraction

A collection of 288 commercial pepper varieties (Table S1), which were acquired from 60 different seed companies in China, were used for genetic identification. These pepper varieties are the most popular throughout China.

Plants were grown under greenhouse conditions at the Vegetable Varieties Exhibition center in the Tongzhou district of Beijing. Fruit shape was categorized into one of four fruit shapes; blocky-fruited, long horn-fruited, short horn-fruited, and linear-fruited (Table S1). The classification examples for the fruit shapes are presented in Fig. S1.

First true leaves from 30 independent plants, which were identified based on the National Varieties Identification Standard, were collected and mixed to extract DNA using a CTAB-based method [52]. The DNA quality was assessed using 1.2% (w/v) agarose gel electrophoresis, and the concentration was determined using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, DE, USA).

Re-sequencing and SNP identification

In total, 31 diverse pepper lines (*C. annuum*), including 30 inbred lines from our ongoing breeding programs and small hot PI640446 provided by the U.S. National Plant Germplasm System (NPGS), were selected for re-sequencing on the Illumina X Ten platform at Shanghai Majorbio Biopharm Technology Co. Ltd. (Shanghai, China). The inbred lines had diverse genetic backgrounds and horticultural traits, including eight blocky-fruited lines, 13 long horn-fruited lines, five short horn-fruited lines and four linear-fruited lines (Fig. 1).

The raw reads of the 31 re-sequenced lines and four previously sequenced cultivars; Zunla-1 (*C. annuum*) and its wild progenitor Chiltepin (*C. annuum var. glabriusculum*) [3], *C. annuum cv.* Perennial and *C. annuum cv.* Dempsey [13], were filtered into clean data using Trimmomatic [53]. The clean reads were then mapped to the reference genome of Zunla-1 [3] using the Burrows-Wheeler Alignment Tool (BWA), and SNPs were called using the Genome Analysis Toolkit (GATK, v2.4-7g5e89f01) [54]. SNPs with minor allele frequency (MAF) > 5% and missing data < 10% were imported into MEGA to build the rooted phylogenetic tree using the cultivar progenitor, Chiltepin, as an out-group with the neighbor-joining method [55]. Population structure analysis was completed using STRUCTURE v2.3. The number of populations (*K*) was determined following the standard procedure [56] with a burn-in period of 100,000 iterations and Markov Chain Monte Carlo of 100,000. Twenty independent runs were performed for *K* varying from 1 to 15. The optimum *K* was defined according to the Evanno's correction method [29].

To acquire a dataset of genome-wide SNPs for subsequent Target SNP-seq analysis, perfect SNPs were identified by the following criteria: (i) minor allele frequency (MAF) > 0.4 to filter out uninformative SNPs; (ii) miss rate < 0.2; (iii) heterozygosity < 0.2; (iv) no sequence variation in the 100 bp flanking sequence of the SNP locus; and (v) the number of alleles per locus was two for the SNP.

Target SNP-seq

The Target SNP-seq procedure was completed as previously described using the SNPs identified above [24]. In brief, library construction for Target SNP-seq consisted of the following two rounds of PCR: the first round amplified and captured the target SNPs in DNA samples using the multiplexed panel of perfect SNP primers; the second round added a unique barcode to the capture product for each DNA sample. Thus, the samples are distinguished based on the different barcodes. The multiplexed PCR was conducted in a 30 μl reaction mixture, containing 50 ng genomic DNA template, 8 μl of the multiplexed SNP-capture panel primers (10 μM), 10 μl of 3 M enzymes (Molbreeding Biotechnology Company,

Shijiazhuang, China). The PCR mixtures were heated at 95℃ for 5 min followed by 17 cycles at 95℃ for 30 s, 60℃ for 4 min, 72℃ for 4 min with a final 4 min extension at 72℃. The PCR products were purified using a magnetic bead suspension and 80% alcohol. Similarly, the second PCR amplification was performed in a 30 μl reaction volume containing 11 μl of purified PCR product from the previous round, 10 μl of 3M *Taq* polymerase, 18 μl nuclease-free water, and 1μl of primers with the following sequences: forward 5'- AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGA -CGCTCTTCCG-3' and reverse 5'-CAAGCAGAAGACGGCATACGAGAT -*XXXXXXXX*GTGACTGGAGTTCCTTGGCACCCGAGA-3' (barcodes are indicated by underlined sequences). The PCR procedure was 95℃ for 3 min; 7 cycles of 95℃ for 15 s, 58℃ for 15 s, and 72℃ for 30 s with a final 4 min extension at 72℃. The PCR products were then purified with 100 μl of 80% alcohol and 23 μl Tris-HCl buffer (10 mM, pH8.0-8.5). After that, the Target SNP-seq library was sequenced using the Illumina X Ten platform at Shanghai Majorbio Biopharm (Shanghai, China).

SNP genotype analysis of Target SNP-seq

The raw data from the Target SNP-seq was de-multiplexed to determine the exact genotypes for each variety based on the sample-specific barcodes using the Illumina bcl2fastq pipeline (Illumina, San Diogo, CA, USA). Clean data were filtered out using Trimmomatic, and the reads of each variety were mapped to the pepper reference genome of Zunla-1 [3] using BWA with default parameters. SNP genotypes were called using GATK. Based on the high-throughput sequencing results, the SNP alleles with the maximum numbers of reads and the second maximum numbers of reads were treated as the major and minor allele for each target SNP locus When the read frequency of the major allele was more than 0.7, the locus was described as homozygous. If the read frequencies of the major and minor allele were both more than 0.35, the locus was described as heterozygous.

Determination of genetic parameters for each perfect SNP

Genetic parameter statistics of the perfect SNPs, including the observed heterozygosity (H_o), expected heterozygosity (H_e), and polymorphism information content (PIC) [57] were calculated using a Perl script with the following equation:

[Due to technical limitations the formula could not be inserted here. It can be found in the supplemental files titled "Formula 1 - Heterozygosity"]

where l is the allele locus, and P_i and $P_j$represent the population frequency of the i^{th} and j^{th} allele. The chromosomal distribution of the perfect SNPs was mapped using Circos software (http://circos.ca/) with the SNP region magnified to 2 Mb.

Genetic structure analysis

Genetic relationships among varieties were investigated using three different methods: principal component analysis (PCA), STRUCTURE, and a phylogenetic tree. PCA was carried out using the FactoMineR package in R [58]. The Bayesian-based model procedure implemented by STRUCTURE v2.3 [56, 59] was also used to determine population structure. The number of populations (*K*) was determined as described above, and the unrooted phylogenetic tree was constructed using the Ape and Poppr packages in R based on the neighbor-joining method with the tree viewed using MEGA v5.1 [60, 61].

Population diversity analysis

The different fruit shape populations, as well as the subpopulations inferred from STRUCTURE, *Ho*, *He*, PIC, and MAF analyses, were calculated using the methods mentioned above. To measure genetic differences between pairs of subpopulations, the analysis of molecular variance (AMOVA) and the pairwise F_{st} was performed using the Poppr and Hierfstat R packages, respectively [61].

Core SNPs set for variety discrimination

To develop a set of core SNPs that discriminates between varieties using the KASPar platform, two allele-specific forward primers and one common reverse primer were designed for each perfect SNP marker. The 23 commercial varieties were then used to assess the potential utility of the SNPs markers through the KASPar platform; fluorescence was detected as previously described [62]. Detailed instructions are available at www.kbioscience.co.uk.

A Perl script was used to determine the core-SNPs set from the successfully verified SNP markers according to the following [24]: 1) the highest discernible SNP marker was chosen as an initial core dataset and each SNP marker was subsequently added to the initial core dataset to form a new dataset; 2) the second SNP marker was chosen from the new dataset with the highest discernibility, and added to the core dataset; 3) the steps were repeated until the maximum discernibility was reached. Finally, the SNP markers associated with the maximum variety discrimination were identified as the core-SNP set.

Association analysis

Fruit shape for each variety was scaled from 1 to 4, with 1 referring to blocky, 2 to long horn, 3 to short horn, and 4 to linear-fruited varieties (Table S1; Fig. S1). The software program TASSEL 5.2.25 was used for association analysis. A mixed linear model (MLM) that considered both fruit shape populations (Q matrix) and the kinship matrix (*K* matrix), and a general linear model (GLM) using fruit shape populations (Q matrix) as a fixed factor were used for association identification of loci conferring fruit shape.

Significance of marker-trait association was indicated when the *p*-value was less than 10^{-4}. Because it has been popularly proved that the MLM+Q+*K* model is a more effective approach than other models for detecting loci [63, 64], only data from the MLM+Q+K model is presented in this study. The phenotypic variation explained by each perfect SNP was the R^2-value obtained from the GLM model. Candidate genes between the nearest up- and down-stream SNP loci to the significantly associated loci were identified from the protein annotation published using the CM334 genome [13].

Abbreviations

PIC: Polymorphism Information Content; *Ho*: observed heterozygosity; *He*: expected heterozygosity; MAF: Minor Allele Frequency; SSR: Simple Sequence Repeats; InDel: insertion or deletion; GBTS: Genotyping By Target Sequencing; RAPD: Random Amplified Polymorphic; RFLP: Restriction Fragment Length Polymorphism; AFLP: Amplified Fragment Length Polymorphism.

Declarations

Funding

This work was partially financed by the National Key Research and Development of China (Grant No. 2017YFD0102004, 2017YFD0101901, and 2016YFD0101704), National Science Foundation of China (Grant No. 31701913), Beijing Nova Program (Grant No. Z181100006218060), Beijing Municipal Department of Organization (Grant No. 2016000021223ZK22), Ministry of Agriculture and Rural Affairs, China (Grant No. 11162130109236051), and Beijing Academy of Agricultural and Forestry Sciences (Grant No. KJCX20170402, KJCX20161503, QNJJ201810, and KJCX2017102).

Availability of data and materials

The raw sequence data reported in this paper have been deposited in the Genome Sequence Archive (Wang et al. 2017) in BIG Data Center (BIG data center members, 2019), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers CRA001576 that are publicly accessible at http://bigd.big.ac.cn/gsa.

Authors' contributions

CW and SG designed the research. HD and JY did the bioinformatics analysis. KY, BC, and XZ contributed materials and helped with data analysis. HD, BC, XZ, and JZ performed the experiments. HD and JY analyzed the data and drafted the manuscript. All authors read and approved the final version of the manuscript.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Moscone EA, Scaldaferro MA, Grabiele M, Cecchini NM, Sánchez García Y, Jarret R, Daviña JR, Ducasse DA, Barboza GE, Ehrendorfer F. The evolution of chili peppers (*Capsicum* - Solanaceae): A cytogenetic perspective. Acta Hortic. 2007;745:137-70 Available from: https://pubag.nal.usda.gov/download/17481/PDF

2. Olmstead RG, Bohs L, Migid HA, Santiago-Valentin E, Garcia VF, Collier SM. A molecular phylogeny of the Solanaceae. Taxon. 2008;57:1159-81 Available from: https://doi.org/10.1002/tax.574010

3. Qin C, Yu CS, Shen YO, Fang XD, Chen L, Min JM, Cheng JW, Zhao SC, Xu M, Luo Y *et al*. Whole-genome sequencing of cultivated and wild peppers provides insights into *Capsicum* domestication and specialization. Proc Natl Acad Sci USA. 2014;111:5135-40 Available from: https://doi.org/10.1073/pnas.1400975111

4. Andrews J. Peppers: The domesticated *Capsicums*. Austin: University of Texas Press. 1984.

5. Gao P, Ma H, Luan F, Song H. DNA fingerprinting of Chinese melon provides evidentiary support of seed quality appraisal. Plos One. 2012;7:e52431 Available from: https://doi.org/10.1371/journal.pone.0052431

6. Tian HL, Wang FG, Zhao JR, Yi HM, Wang L, Wang R, Yang Y, Song W. Development of maizeSNP3072, a high-throughput compatible SNP array, for DNA fingerprinting identification of Chinese maize varieties. Mol Breed.2015;35:136 Available from: https://doi.org/10.1007/s11032-015-0335-0

7. McCouch SR, Chen XL, Panaud O, Temnykh S, Xu YB, Cho YG, Huang N, Ishii T, Blair M. Microsatellite marker development, mapping and applications in rice genetics and breeding. Plant Mol Biol. 1997;35:89-99 Available from: https://doi.org/10.1023/a:1005711431474

8. Nagaraju J, Kathirvel M, Kumar RR, Siddiq EA, Hasnain SE. Genetic analysis of traditional and evolved Basmati and non-Basmati rice varieties by using fluorescence-based ISSR-PCR and SSR markers (vol 99, pg 5836, 2002). Proc Natl Acad Sci USA. 2002;99:13357 Available from: https://doi.org/10.1073/pnas.212463799

9. Darine T, Allagui MB, Rouaissi M, Boudabbous A. Pathogenicity and RAPD analysis of *Phytophthora nicotianae* pathogenic to pepper in Tunisia. Physiol Mol Plan Pathol. 2007;70:142-8 Available from: https://doi.org/10.1016/j.pmpp.2007.08.002

10. Lanteri S, Acquadro A, Quagliotti L, Portis E. RAPD and AFLP assessment of genetic variation in a landrace of pepper (*Capsicum annuum* L.), grown in North-West Italy. Gen Res Crop Evol. 2003;50:723-35 Available from: https://doi.org/10.1023/a:1025075118200

11. Lefebvre V, Palloix A, Rives M. Nuclear RFLP between pepper cultivars (*Capsicum annuum* L).Euphytica. 1993;71:189-99 Available from: https://10.1007/bf00040408

12. Tanksley SD, Bernatzky R, Lapitan NL, Prince JP. Conservation of gene repertoire but not gene order in pepper and tomato. Proc Natl Acad Sci USA. 1988;85:6419-23 Available from: https://doi.org/10.1073/pnas.85.17.6419

13. Kim S, Park M, Yeom SI, Kim YM, Lee JM, Lee HA, Seo E, Choi J, Cheong K, Kim KT *et al*. Genome sequence of the hot pepper provides insights into the evolution of pungency in *Capsicum* species. Nat Genet. 2014;46:270-8 Available from: https://doi.org/10.1038/ng.2877

14. Guo GJ, Zhang GL, Pan BG, Diao WP, Liu JB, Ge W, Gao CZ, Zhang Y, Jiang C, Wang SB. Development and application of InDel markers for *Capsicum* spp. based on whole-genome re-sequencing. Sci Rep. 2019;9:3691 Available from: https://doi.org/10.1038/s41598-019-40244-y

15. Li WP, Cheng JW, Wu ZM, Qin C, Tan S, Tang X, Cui JJ, Zhang L, Hu KL. An InDel-based linkage map of hot pepper (*Capsicum annuum*). Mol Breed. 2015;35:32 Available from: https://doi.org/10.1007/s11032-015-0219-3

16. Tan S, Cheng JW, Zhang L, Qin C, Nong DG, Li WP, Tang X, Wu ZM, Hu KL. Construction of an interspecific genetic map based on InDel and SSR for mapping the QTLs affecting the initiation of flower primordia in pepper (*Capsicum* spp.). Plos One. 2015;10:e0119389 Available from: https://doi.org/10.1371/journal.pone.0119389

17. Fan JB, Oliphant A, Shen R, Kermani BG, Garcia F, Gunderson KL, Hansen M, Steemers F, Butler SL, Deloukas P *et al*. Highly parallel SNP genotyping. Cold Spring Harbor Symposia on Quantitative Biology. 2003;68:69-78 Available from: https://doi.org/10.1101/sqb.2003.68.69

18. Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray™ platform. Biotechnol J. 2007;2:41-9 Available from: https://doi.org/10.1002/biot.200600213

19. Livak KJ, Flood SJA, Marmaro J, Giusti W, Deetz K. Oligonucleotides with fluorescent dyes at opposite ends provide a quenched probe system useful for detecting PCR product and nucleic acid hybridization. Genome Res. 1995;4:357-62 Available from: https://www.ncbi.nlm.nih.gov/pubmed/7580930

20. Li L, Fang ZW, Zhou JF, Chen H, Hu ZF, Gao LF, Chen LH, Ren S, Ma HY, Lu L, Zhang WX, Peng H.An accurate and efficient method for large-scale SSR genotyping and applications. Nucleic Acids Research. 2017;45 Available from: https://doi.org/10.1093/nar/gkx093

21. Krasileva KV, Vasquez-Gross HA, Howell T, Bailey P, Paraiso F, Clissold L, Simmonds J, Ramirez-Gonzalez RH, Wang XD, Borrill P, Fosker C, Ayling S, Phillips AL, Uauy C, Dubcovsky J. Uncovering hidden variation in polyploid wheat. Proc Natl Acad Sci USA. 2017;114:913-21 Available from: https://doi.org/10.1073/pnas.1619268114

22. Jiang L, Liu X, Yang J, Wang HF, Jiang JC, Liu LL, He S, Ding XD, Liu JF, Zhang Q. Targeted resequencing of GWAS loci reveals novel genetic variants for milk production traits. BMC Genomics 2014;15:1105 Available from: https://doi.org/10.1186/1471-2164-15-1105

23. Guo ZF, Wang HW, Tao JJ, Ren YH, Xu C, Wu KS, Zou C, Zhang JN, Xu YB. Development of multiple SNP marker panels affordable to breeders through genotyping by target sequencing (GBTS) in maize. Mol Breed. 2019;39:37 Available from: https://doi.org/10.1007/s11032-019-0940-4

24. Yang JJ, Zhang J, Han RX, Zhang F., Mao AJ, Luo J, Dong BB, Liu H, Tang H, Zhang JN, Wen CL. Target SSR-seq: A novel SSR genotyping technology associate with perfect SSRs in genetic analysis of cucumber varieties. Front Plant Sci. 2019;10:531 Available from: https://doi.org/10.3389/fpls.2019.00531

25. Aguilar-Meléndez A, Morrell PL, Roose ML, Kim SC. Genetic diversity and structure in semiwild and domesticated chiles (*Capsicum annuum*; Solanaceae) from Mexico. Am J Bot. 2009;96:1190-202 Available from: https://doi.org/10.3732/ajb.0800155

26. Ibiza VP, Blanca J, Canizares J, Nuez F. Taxonomy and genetic diversity of domesticated *Capsicum* species in the Andean region. Gen Res Crop Evol. 2012;59:1077-88 Available from: https://doi.org/10.1007/s10722-011-9744-z

27. Yumnam JS, Tyagi W, Pandey A, Meetei NT, Rai M. Evaluation of genetic diversity of chilli landraces from North Eastern India based on morphology, SSR markers and the *Pun1* locus. Pl Mol Reporter. 2012;30:1470-79 Available from: https://doi.org/10.1007/s11105-012-0466-y

28. Zhang XF, Sun HH, Xu Y, Chen B, Yu SC, Geng SS, Wang Q. Development of a large number of SSR and InDel markers and construction of a high-density genetic map based on a RIL population of pepper (*Capsicum annuum* L.). Mol Breed. 2016;36:92 Available from: https://doi.org/10.1007/s11105-012-0466-y

29. Evanno G, Regnaut S, Goudet J. Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Mol Ecol. 2005;14:2611-20 Available from: https://doi.org/10.1111/j.1365-294X.2005.02553.x

30. Lee JM, Nahm SH, Kim YM, Kim BD. Characterization and molecular genetic mapping of microsatellite loci in pepper. Theor Appl Genet. 2004;108:619-27 Available from: https://doi.org/10.1007/s00122-003-1467-x

31. Taranto F, D'Agostino N, Greco B, Cardi T, Tripodi P. Genome-wide SNP discovery and population structure analysis in pepper (*Capsicum annuum*) using genotyping by sequencing. BMC Genomics. 2016; 17 Available from: https://doi.org/10.1186/s12864-016-3297-7

32. Kraft KH, Brown CH, Nabhan GP, Luedeling E, Ruiz JDL, d'Eeckenbrugge GC, Hijmans RJ, Gepts P. Multiple lines of evidence for the origin of domesticated chili pepper, *Capsicum annuum*, in Mexico. Proc Natl Acad Sci USA. 2014;111:6165-70 Available from: https://doi.org/10.1073/pnas.1308933111

33.Bosland PW, Votava E. Peppers: vegetable and spice *Capsicums*. Oxford, Wallingford: Cabi. 2000

34. Geng SS, Chen B, Zhang XF, Sun JT. Hot pepper breeding development and its varieties's distribution in China. Journal of China Capsicum. 2011;1:1-5 (In Chinese) Available from: https://www.ifabiao.com/lj/201103/15393478.html

35. Geng SS, Chen B, Zhang XF, Du HS. The trend of market demand and breeding strategies of pepper varieties in China. China Vegetables. 2015;3:1-5 (In Chinese) Available from: http://www.cnveg.org/UserFiles/File/3-1.pdf

36. Moreira AFP, Ruas PM, Ruas CD, Baba VY, Giordani W, Arruda IM, Rodrigues R, Goncalves LSA. Genetic diversity, population structure and genetic parameters of fruit traits in *Capsicum chinense*. Sci Hortic. 2018;236:1-9 Available from: https://doi.org/10.1016/j.scienta.2018.03.012

37. Ou LJ, Li D, Lv JH, Chen WC, Zhang ZQ, Li XF, Yang BZ, Zhou SD, Yang S, Li WG *et al*. Pan-genome of cultivated pepper (*Capsicum*) and its use in gene presence-absence variation analyses. New Phytol. 2018;220:360-63 Available from: https://doi.org/10.1111/nph.15413

38. Moses M, Umaharan P, Dayanandan S. Microsatellite based analysis of the genetic structure and diversity of *Capsicum chinense* in the Neotropics. Gen Res Crop Evol. 2014;61:741-55 Available from: https://doi.org/10.1007/s10722-013-0069-y

39. Baral JB, Bosland PW. Unraveling the species dilemma in *Capsicum frutescens* and *C. chinense* (Solanaceae): A multiple evidence approach using morphology, molecular analysis, and sexual compatibility. J Amer Soc Hort Sci. 2004;129:826-32 Available from: https://doi.org/10.21273/JASHS.129.6.0826

40. Gonzalez-Perez S, Garces-Claver A, Mallor C, de Miera LES, Fayos O, Pomar F, Merino F, Silvar C. New insights into *Capsicum* spp relatedness and the diversification process of *Capsicum annuum* in Spain. PLoS One. 2014;9:e116276 Available from: https://doi.org/10.1371/journal.pone.0116276

41. Hill TA, Ashrafi H, Reyes-Chin-Wo S, Yao JQ, Stoffel K, Truco MJ, Kozik A, Michelmore RW, Van Deynze A. Characterization of *Capsicum annuum* genetic diversity and population structure based on parallel polymorphism discovery with a 30K unigene Pepper GeneChip. Plos One. 2013;8:e56200 Available from: https://doi.org/10.1371/journal.pone.0056200

42. Nicolai M, Cantet M, Lefebvre V, Sage-Palloix AM, Palloix A. Genotyping a large collection of pepper (*Capsicum* spp.) with SSR loci brings new evidence for the wild origin of cultivated *C. annuum* and the structuring of genetic diversity by human selection of cultivar types. Gen Res Crop Evol.2013;60:2375-90 Available from: https://doi.org/10.1007/s10722-013-0006-0

43. Tam SM, Lefebvre V, Palloix A, Sage-Palloix AM, Mhiri C, Grandbastien MA. LTR-retrotransposons Tnt1 and T135 markers reveal genetic diversity and evolutionary relationships of domesticated peppers. Theor Appl Genet. 2009;119:973-89 Available from: https://doi.org/10.1007/s00122-009-1102-6

44. Chaim AB, Borovsky Y, De Jong W, Paran I. Linkage of the A locus for the presence of anthocyanin and *fs10.1*, a major fruit-shape QTL in pepper. Theor Appl Genet. 2003;106:889-94 Available from: https://doi.org/10.1007/s00122-002-1132-9

45. Chunthawodtiporn J, Hill T, Stoffel K, Van Deynze A. Quantitative trait loci controlling fruit size and other horticultural traits in bell pepper (*Capsicum annuum*). Plant Genome. 2018;11:160125 Available from: https://doi.org/10.3835/plantgenome2016.12.0125

46. van der Knaap E, Chakrabarti M, Chu YH, Clevenger JP, Illa-Berenguer E, Huang ZJ, Keyhaninejad N, Mu Q, Sun L, Wang YP,Wu S. What lies beyond the eye: the molecular mechanisms regulating tomato fruit weight and shape. Front Plant Sci. 2014;5:227 Available from: https://doi.org/10.3389/fpls.2014.00227

47. Wu S, Zhang BY, Keyhaninejad N, Rodriguez GR, Kim HJ, Chakrabarti M, Illa-Berenguer E, Taitano NK, Gonzalo MJ, Diaz A, Pan YP, Leisner CP, Halterman D, Buell CR, Weng YQ, Jansky SH, van Eck H, Willemsen J, Monforte AJ, Meulia T, van der Knaap E. A common genetic mechanism underlies morphological diversity in fruits and other plant organs. Nat Commun. 2018;9:4734 Available from: https://doi.org/10.1038/s41467-018-07216-8

48. Yang HB, Liu WY, Kang WH, Kim JH, Cho HJ, Yoo JH, Kang BC. Development and validation of L allele-specific markers in *Capsicum.* Mol Breed. 2012;30:819-29 Available from: https://doi.org/10.1007/s11032-011-9666-7

49. Rehrig WZ, Ashrafi H, Hill T, Prince J, Van Deynze A. *CaDMR1* Cosegregates with QTL *Pc5.1* for resistance to *Phytophthora capsici* in pepper (*Capsicum annuum*). Plant Genome. 2014;7:1-12 Available from: https://doi.org/10.3835/plantgenome2014.03.0011

50. Romer P, Hahn S, Jordan T, Strauss T, Bonas U, Lahaye T. Plant pathogen recognition mediated by promoter activation of the pepper *Bs3* resistance gene. Science. 2007;318:645-48 Available from: https://doi.org/10.1126/science.1144958

51. Yeam I, Kang BC, Lindeman W, Frantz JD, Faber N, Jahn MM. Allele-specific CAPS markers based on point mutations in resistance alleles at the *pvr1* locus encoding eIF4E in *Capsicum*. Theor Appl Genet. 2005;112:178-86 Available from: https://doi.org/10.1007/s00122-005-0120-2

52. Fulton TM, Chunwongse J, Tanksley SD. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol Biol Reporter. 1995;13:207-9 Available from: https://doi.org/10.1007/bf02670897

53. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30:2114-20 Available from: https://doi.org/10.1093/bioinformatics/btu170

54. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. The Genome Analysis Toolkit: A mapreduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297-303 Available from: https://doi.org/10.1101/gr.107524.110

55. Price MN, Dehal PS, Arkin AP. FastTree 2-approximately maximum-likelihood trees for large alignments. Plos One. 2010;5:e9490 Available from: https://doi.org/10.1371/journal.pone.0009490

56. Pritchard JK, Stephens M, Donnelly P. Inference of population structure using multilocus genotype data. Genetics. 2000;155:945-59 Available from: https://www.ncbi.nlm.nih.gov/pubmed/10835412

57. Botstein D, White RL, Skolnick M, Davis RW. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. Am J Hum Genet.1980;32:314-31 Available from: https://www.ncbi.nlm.nih.gov/pubmed/6247908

58. Husson F, Josse J, Pages J. Principal component methods-hierarchical clustering-partitional clustering: why would we need to choose for visualizing data? Technical report-Agrocampus, Applied Mathematics Department. 2010 Available from: http://www.agrocampus-ouest.fr/math/

59. Falush D, Stephens M, Pritchard JK. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics. 2003;164:1567-87 Available from: https://www.ncbi.nlm.nih.gov/pubmed/12930761

60. Nei, M. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics. 1978;89:583-90 Available from: https://www.ncbi.nlm.nih.gov/pubmed/17248844

61. Kamvar ZN, Tabima JF, Grünwald NJ. Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. Peerj. 2014;2:e281 Available from: https://doi.org/10.7717/peerj.281

62. Su TB, Li PR, Yang JJ, Sui GL, Yu YJ, Zhang DS, Zhao XY, Wang WH, Wen CL, Yu SC, Zhang FL.Development of cost-effective single nucleotide polymorphism marker assays for genetic diversity analysis in *Brassica rapa*. Mol Breed. 2018;38:42 Available from: https://doi.org/10.1007/s11032-018-0795-0

63. Pace J, Gardner C, Romay C, Ganapathysubramanian B, Lubberstedt T. Genome-wide association analysis of seedling root development in maize (*Zea mays* L.). BMC Genomics. 2015;16:47 Available from: https://doi.org/10.1186/s12864-015-1226-9

64. Sim SC, Robbins MD, Wijeratne S, Wang H, Yang WC, Francis DM. Association analysis for bacterial spot resistance in a directionally selected complex breeding population of tomato. Phytopathology. 2015;105:1437-45 Available from: https://doi.org/10.1094/phyto-02-15-0051-r

Tables

Table 1 Genetic diversity in the fruit shape populations and across all varieties.

Fruit shapes	Varieties Size	PIC [a]	*He*	*Ho*	MAF
Blocky-fruited	90	0.15 (4) [b]	0.18(4)	0.16 (4)	0.13 (4)
Long horn-fruited	113	0.31 (1)	0.39 (1)	0.36 (1)	0.30 (1)
Short horn-fruited	25	0.29 (2)	0.37 (2)	0.34 (2)	0.29 (2)
Linear-fruited	43	0.25 (3)	0.31 (3)	0.33 (3)	0.23 (3)
Total	271	0.31	0.40	0.28	0.31

[a] For each fruit population: polymorphism information content (PIC), expected heterozygosity (*He*), observed heterozygosity (*Ho*), and minor allele frequency (MAF).

[b] The numbers in parentheses refer to the numerical ranking of diversity in descending order.

Table 2. Analysis of molecular variance (AMOVA) among Pops and Subpops

	Sum of squares	Variance components	Percentage of variation
Among Pops / Subpops	4013.64 / 5368.25	14.88 / 13.46	33.04% / 32.63%
Within Pops / Subpops	9137.06 / 7782.44	3.81 / 1.46	8.47% / 3.54%
Within varieties of Pops / Subpops	7137.53 / 7137.53	26.34 / 26.34	58.49% / 63.83%
Total	20288.23 / 20288.23	45.03 / 41.26	100.00 / 100.00%

Table 3. Pairwise *F* statistics (F_{st}) estimates among subpopulations.

Subpopulations	Subpop2	Subpop3	Subpop4	Subpop5
Subpop1	0.21	0.23	0.48	0.43
Subpop2		0.13	0.31	0.28
Subpop3			0.22	0.18
Subpop4				0.14

Table 4. SNP loci significantly associated with fruit shape as identified by association analysis.

Marker name	CM334 (v.1.55)		Zunla-1 (v2.0)		*p*-value	Phenotypic variation explained (%)
	Chromosome	Physical position	Chromosome	Physical position		
CaSNP118	11	217566921	0	185313467	1.80E-08	13.6
CaSNP053	6	219972322	6	15889098	1.80E-05	9.2

Additional File Legend

Figure S1. Examples of fruit shape classification. Fruit shapes were categorized into four types as (A) blocky-fruited: blocky shape, 5.0-12.5 cm wide at the shoulder, 7.0-18 cm long, 3-4 lobes, including Fang Jiao, Chang Fang Jiao, and Ma La Jiao, as named in China; (B) long horn-fruited: long horn shape, 3.0-8.0 cm wide at the shoulder, 10.0-35.0 cm long, without lobe, including Niu Jiao Jiao, Yang Jiao Jiao, and

Luo Si Jiao, as named in China; (C) short horn-fruited: cone-shaped, medium-hot, 1.0-3.0 cm in diameter at the base, 3.5-10.0 cm in length, and with very thin pericarp, including Gan Jiao and Chao Tian Jiao, as named in China; (D): linear-fruited: cayenne type, 1.0-3.0 cm wide by 3.5-11.0 cm long, without shoulder and lobe, including Xian Jiao, Tiao Jiao and Mei Ren Jiao, as named in China.

Figure S2. Target SNP-seq genotyping result analysis. Distribution of the average read depths (A), the reads alignment rate to the pepper reference genome (B), the target region alignment rate (C), and the uniformity index for 271 pepper varieties (D).

Figure S3. Genetic diversity analysis for the 92 perfect SNPs across 271 pepper varieties. Minor allele frequency (MAF; A), observed heterozygosity (H_o; B), expected heterozygosity (H_e; C), and polymorphism information content (PIC; D)

Table S1. Classification and information on the pepper varieties used in this study.

Table S2. Multiplexed primers panel of 92 perfect SNPs used for Target SNP-seq.

Table S3. Characteristics of the 92 perfect SNPs and the diversity detected in the 271 pepper varieties and four fruit shape populations.

Table S4. Primer sequences of the 35 core SNP markers developed in this study.

Table S5. Annotated genes in the region associated with fruit shape.

Figures

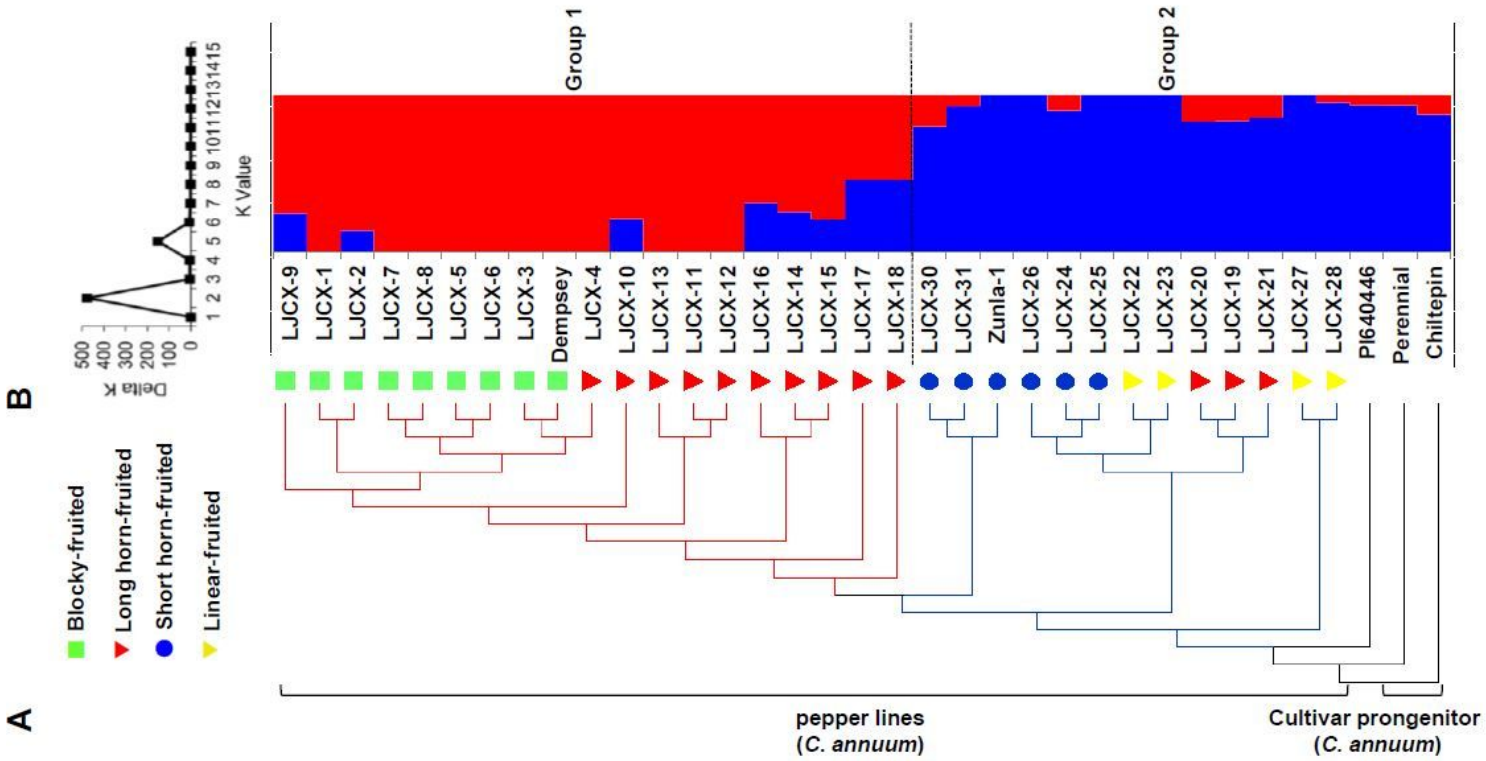

Figure 1

Population structure across the pepper lines. Phylogenetic relationships (A) and population structure (B) based on the total SNPs of the 31 pepper inbred lines sequenced in this study and the previously sequenced C. annuum cultivars Zunla-1, Chiltepin, Perennial, and Dempsey. Fruit shapes are given as colored shapes.

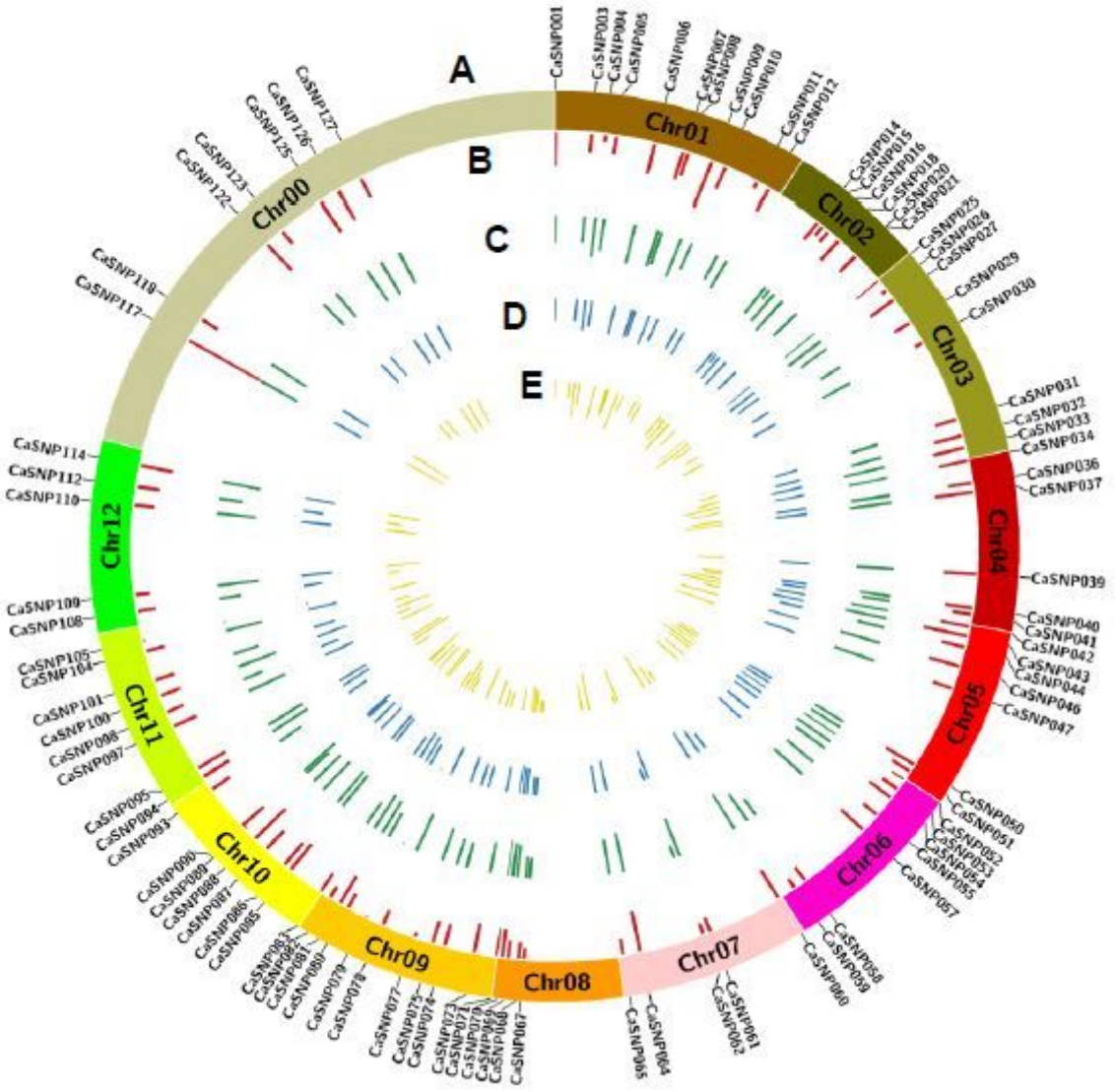

Figure 2

Characteristics of perfect SNPs used to genotype pepper varieties by Target SNP-seq. (A) The distribution of the 92 perfect SNPs in the ideogram of the genome of C. annuum Zunla-1. (B) Observed heterozygosity (Ho) per SNP locus, colored in red.(C) Expected heterozygosity (He) per SNP locus is given in green. (D) Polymorphism information content (PIC) per SNP locus is given in blue. (E) Minor allele frequency (MAF) per SNP locus is given in yellow. This figure was generated using Circos (http://circos.ca/) with the SNP region magnified to 2 Mb.

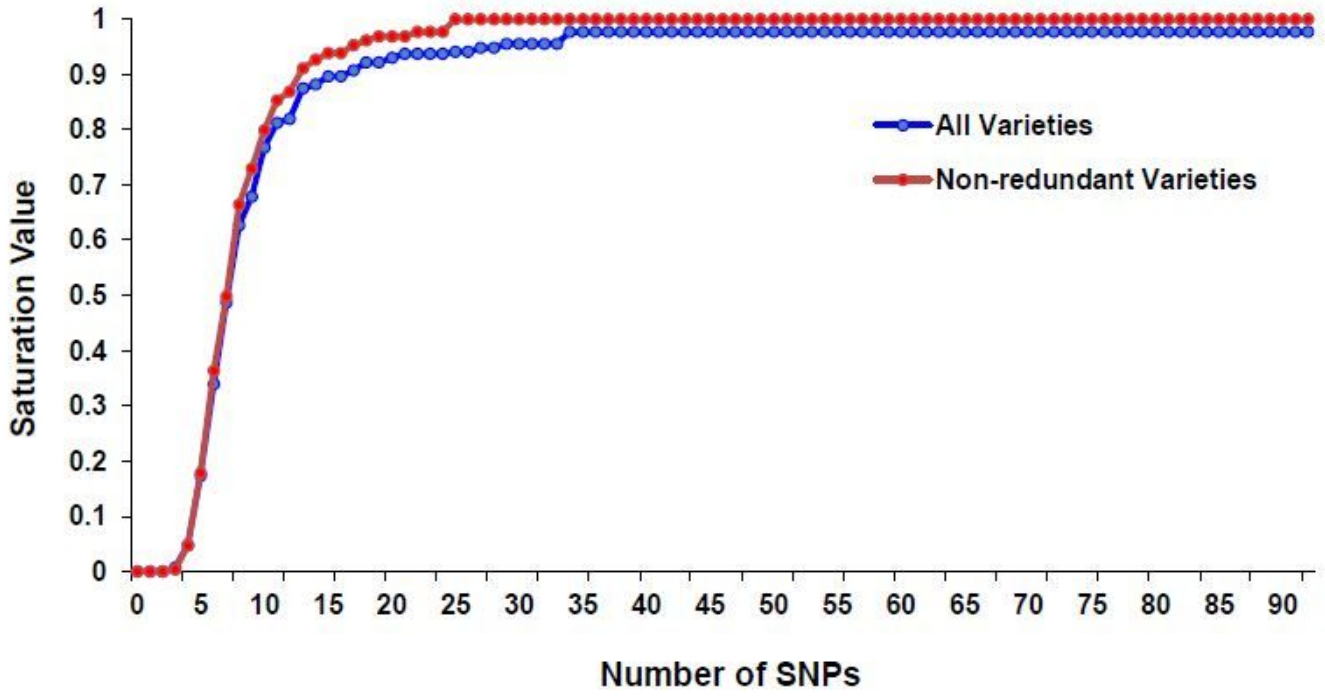

Figure 3

The discriminating saturation curve of 92 perfect SNPs in pepper varieties. The maximum discrimination capacity is 97.7% across all 271 varieties using 35 perfect SNPs, and 100% across the non-redundant varieties using 27 perfect SNPs.

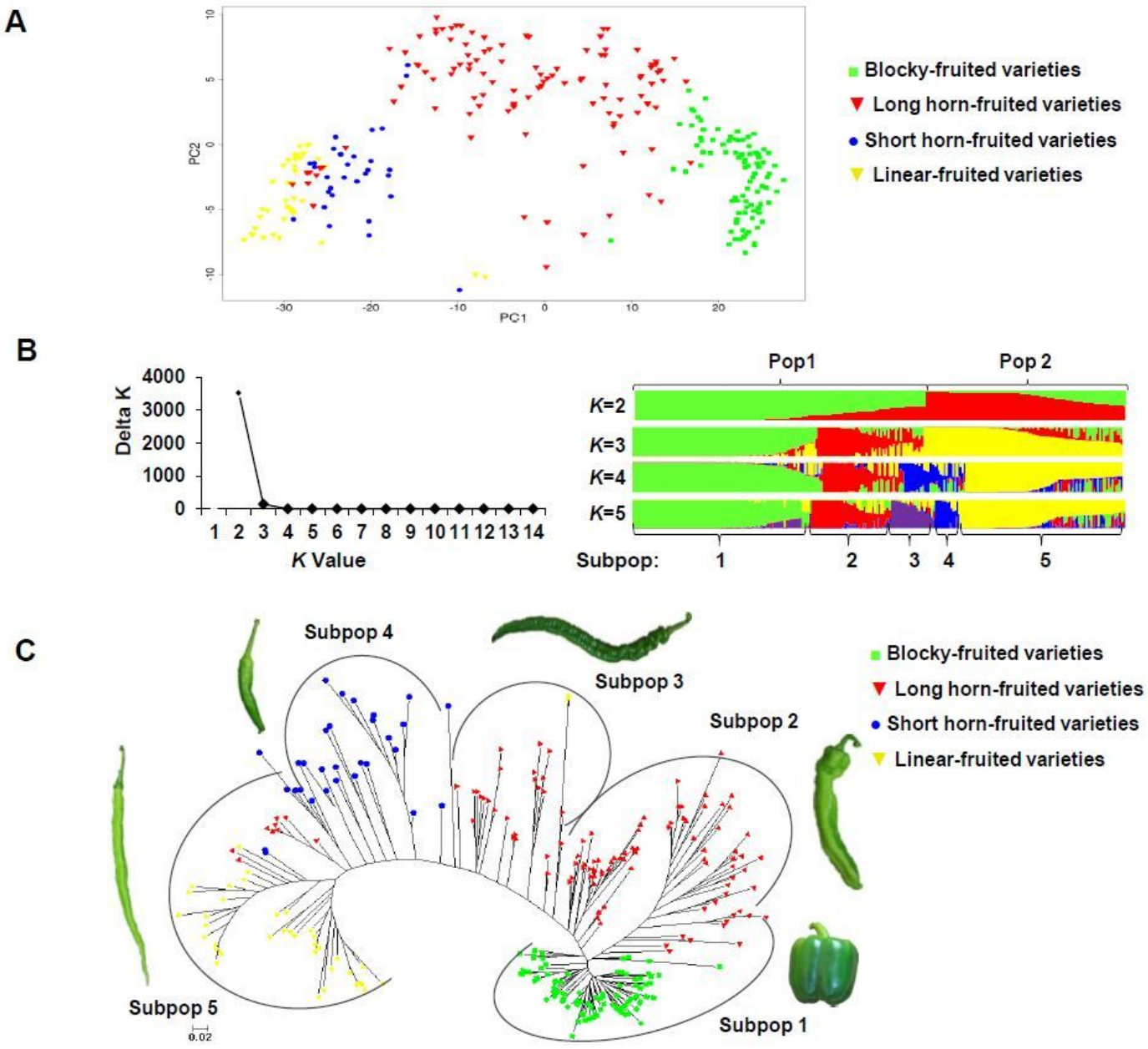

Figure 4

Population structure across the pepper varieties. (A) Principal component analysis (PCA). (B) Population structure inferred using STRUCTURE. All the varieties were divided into two main populations (Pop1 and Pop2) when K=2, which was the optimal K. The populations were subdivided into five subpopulations, Subpop1~Subpop5, which correlated with fruit shape. (C) Phylogenetic tree analysis. The tree was produced using the neighbor-joining method based on the 92 perfect SNPs. The scale bar indicates the simple matching distance.

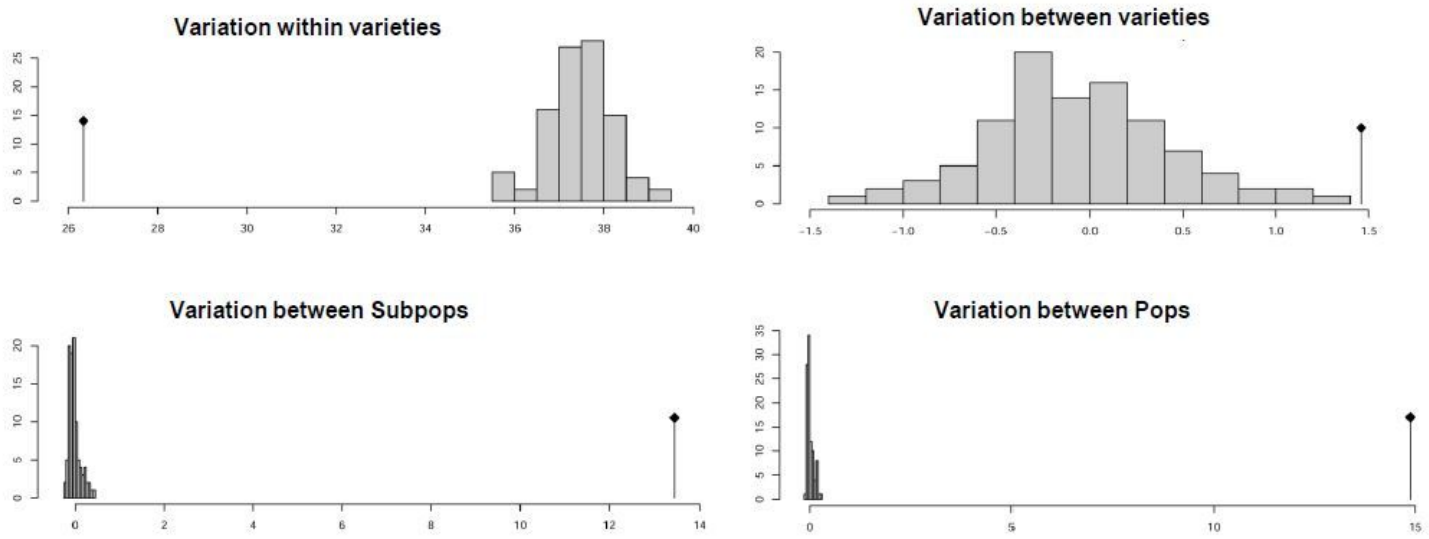

Figure 5

Significance testing of differentiation between the Pops and among the Subpops. The graphs show significant population differentiation at all levels given that the observed line (black) does not fall within the distribution expected of the permutation.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Fig.S3.pdf
- TableS2.xlsx
- Fig.S1.pdf
- TableS3.xlsx
- Fig.S2.pdf
- TableS4.xls
- Formula1Heterozygosity.JPG
- TableS5.xlsx
- TableS1.xlsx