Target sequencing reveals genetic diversity, population structure, core-SNP markers, and fruit shape-associated loci in pepper varieties

Heshan Du[1,2†], Jingjing Yang[1,2†], Bin Chen[1,2], Xiaofen Zhang[1,2], Jian Zhang[1,2], Kun Yang[3], Sansheng Geng[1,2*] and Changlong Wen[1,2*]

Abstract

Background: The widely cultivated pepper (*Capsicum* spp.) is one of the most diverse vegetables; however, little research has focused on characterizing the genetic diversity and relatedness of commercial varieties grown in China. In this study, a panel of 92 perfect single-nucleotide polymorphisms (SNPs) was identified using re-sequencing data from 35 different *C. annuum* lines. Based on this panel, a Target SNP-seq genotyping method was designed, which combined multiplex amplification of 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, observed heterozygosity, expected heterozygosity, and minor allele frequency, which were 0.31, 0.28, 0.4, and 0.31, respectively. Notably, the studied pepper varieties were morphologically categorized based on fruit shape as 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 kompetitive allele-specific PCR (KASPar) markers, another robust genotyping technique for variety identification. Analysis of genetic relatedness using principal component analysis 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 F_{st} statistics. Nine loci, located on chromosomes 1, 2, 3, 4, 6, and 12, were identified to be significantly associated with the fruit shape index ($p < 0.0001$).

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 Chinese pepper varieties is significantly influenced by breeding programs focused on fruit shape.

Keywords: Pepper, SNP, Genetic structure, Target SNP-seq, Association analysis

* Correspondence: gengsansheng@nercv.org; wenchanglong@nercv.org
†Heshan Du and Jingjing Yang contributed equally to this work.
[1]Beijing Vegetable Research Center (BVRC), Beijing Academy of Agricultural and Forestry Sciences, Beijing 100097, China
Full list of author information is available at the end of the article

Du *et al. BMC Plant Biology*   (2019) 19:578

Page 2 of 16

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 distinguish 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 [14–20]. 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 [21] and Infinium [22], TaqMan [23], and KASPar platform (KBiosciences, www.kbioscience.co.uk). In recent years, high-throughput transcriptome sequencing and genotyping-by-sequencing (GBS) have been successfully used in pepper, generating highly informative genome-wide SNP data [24–30]. However, SNP marker genotyping is considered expensive as it requires a comprehensive technical platform and special equipment and reagents.

Genotyping by target sequencing (GBTS) is a targeted sequence-capture strategy that can genotype more than thousands of SSRs or SNPs using 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 [31], NimbleGen [32], SureSelect [33], Geno-Baits, and GenoPlexs [34]. 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 [35]. The results of this study demonstrated that GBTS is a customizable, flexible, high-throughput, low cost, and accurate sequencing tool.

Peppers from China constitute one-third of the world's pepper production [36]. Until now, the genetic diversity of pepper accessions in China has primarily been investigated using SSR markers, but these surveys only examined either several Chinese germplasm (up to 32) [37] or a small number of SSR markers (up to 28) [36]. However, high-throughput SNP platforms used for genotyping and the identification of pepper varieties have lagged significantly behind those for SSRs, and studies on the genetic diversity between the varieties of peppers in China has not yet been extensively analyzed. Therefore, the main 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 (*C. annuum*) in this study generated a total of 872 Gb of paired-end sequence data, at an average depth of ~ 8.4. After mapping to the Zunla-1 genome [3], 40,700,040 SNPs were detected across the genomic sequences of the 31 re-sequenced lines and four previously published cultivars (Dempsey, Zunla-1, Perennial, and Chiltepin) [3, 13]. Approximately 11.3% of the *C. annuum* genome contains variable SNP sites. A total of 21,237,194 SNPs, with minor allele frequency (MAF) > 5% and missing data < 10%, were considered high-quality SNPs for downstream analyses. Using *C. annuum*'s progenitor cultivar,

Chiltepin, as an outgroup, the phylogenetic tree showed that pepper lines could generally be classified according to fruit shapes, except for three long horn-fruited lines that grouped with the linear-fruited lines. Based on the genetic distance, the transitions in fruit shapes were from Chiltepin-like peppers followed by the 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 ten of the long horn-fruited lines, whereas the remaining peppers, including three long horn-fruited, all the linear-fruited, and all the short horn-fruited peppers, as well as two cultivar progenitors Perennial and Chiltepin were assigned to Group 2. The clustering of these pepper lines appeared to be more related to fruit type, when $K = 5$. Group 1 was divided into Subgroup 1 (mostly blocky-fruited) and Subgroup 2 (long horn-fruited), whereas Group 2 was composed of Subgroup 3 (admixture, mostly short horn-fruited), Subgroup 4 (linear-fruited), and Subgroup 5 (cultivar progenitors with small fruit).

Given that pepper genomes are highly repetitive, strict criteria were used to identify the perfect SNPs (See Methods). In total, 521 perfect SNPs were identified, and

92, which were distributed across the genome (Fig. 2; Additional file 9: Table S2), were selected as multiplex PCR targets. Based on the previous annotation [3], 83 and 9 perfect SNPs fall within intergenic and genic regions, respectively. The nearest flanking annotated genes for each perfect SNP are shown in Additional file 9: Table S2.

Genotyping analysis of pepper varieties using the target SNP-seq

In total, 271 pepper varieties, including 90 blocky-, 113 long horn-, 25 short horn-, and 43 linear-fruited varieties, were genotyped using the Target SNP-seq (Additional file 8: Table S1). A total of 55.9 million reads were generated from the 271 varieties, with an average target read depth of 2064, and approximately 82% of the samples were sequenced at a depth greater than $1000 \times$ (Additional file 2: Figure S2A). Among the 271 varieties, 238 varieties (87.8%) aligned to the Zunla-1 genome [3] at a rate of more than 90% (Additional file 2: Figure S2B). Of these aligned reads, 221 varieties (81.5%) exhibited an align rate to the target SNP region of over 80% (Additional file 2: Figure 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 93.68%
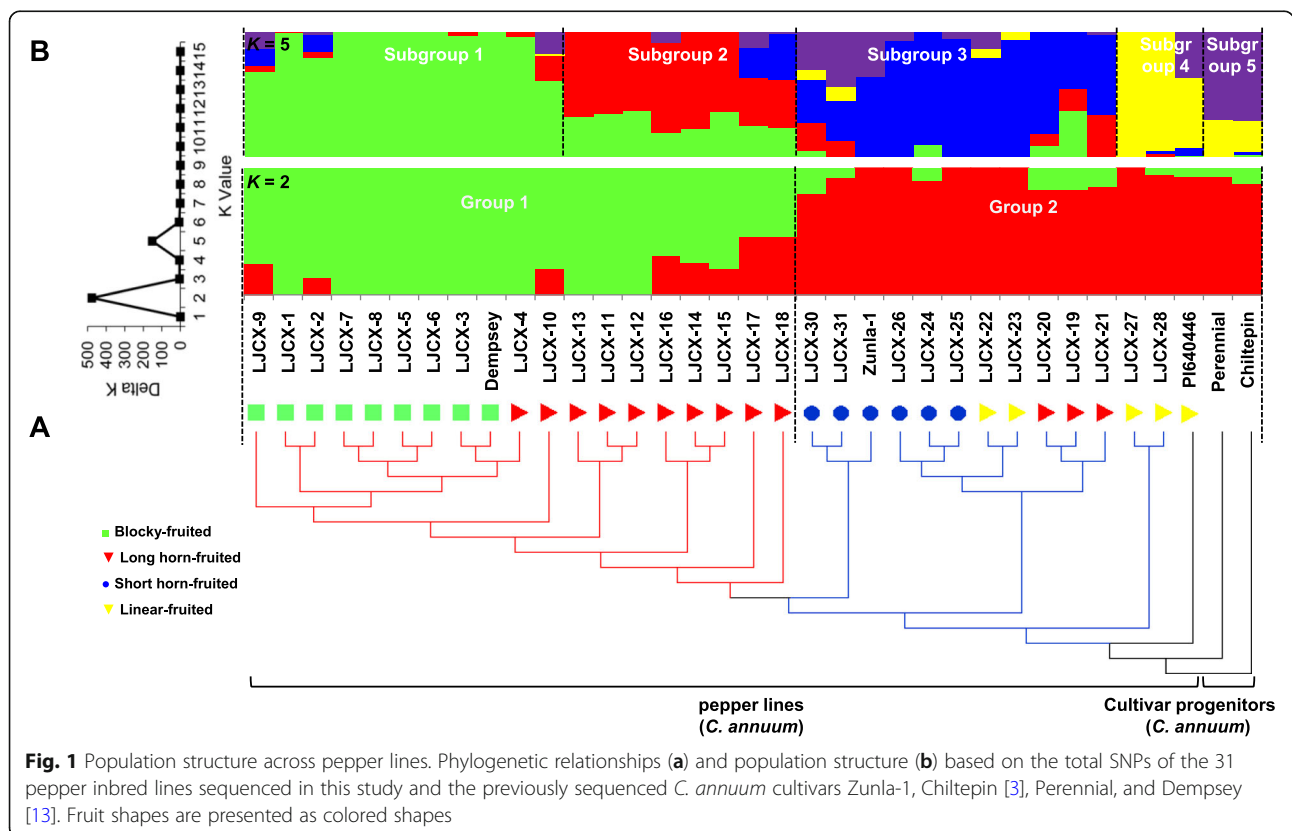

Fig. 1 Population structure across pepper lines. Phylogenetic relationships (**a**) and population structure (**b**) based on the total SNPs of the 31 pepper inbred lines sequenced in this study and the previously sequenced *C. annuum* cultivars Zunla-1, Chiltepin [3], Perennial, and Dempsey [13]. Fruit shapes are presented as colored shapes
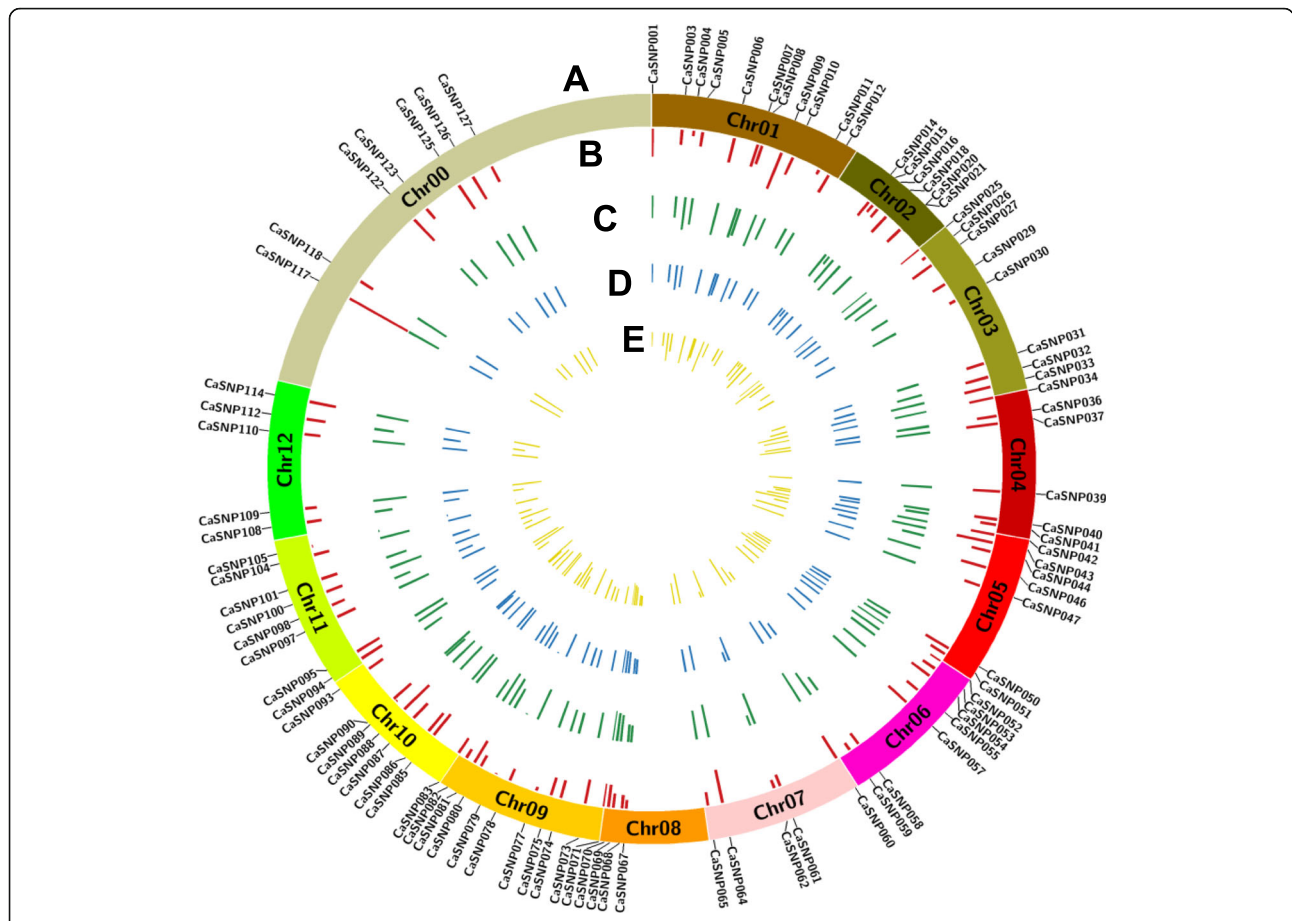
Fig. 2 Characteristics of the perfect SNPs used to genotype pepper varieties by Target SNP-seq. **a** Distribution of the 92 perfect SNPs in the ideogram of the genome of *C. annuum* Zunla-1 [3]. **b** Observed heterozygosity (*Ho*) per SNP locus, colored in red. **c** Expected heterozygosity (*He*) per SNP locus is presented in green. **d** Polymorphism information content (PIC) per SNP locus is presented 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

(Additional file 2: Figure S2D), which indicated a high uniformity of sequence depth among the 92 SNPs.

Perfect SNPs in 271 pepper varieties

The genetic parameters, MAF, *Ho*, *He*, and PIC revealed by each perfect SNP are given in Additional file 10: 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 a MAF between 0.4 and 0.5, whereas only four SNPs had MAF below 0.1 (Additional file 3: Figure S3A). 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) (Additional file 3: Figure S3B; Additional file 10: Table S3). Furthermore, the *He* values ranged from 0.01 (CaSNP079) to 0.5 (CaSNP043 and CaSNP094) (Additional file 3: Figure S3C; Additional file 10: 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 (Additional file 3: Figure S3D; Additional file 10: 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 that 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

Table 1 Genetic diversity in 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 a descending order

highest genetic diversity as defined by the highest average values of He (0.39), Ho (0.36), and PIC (0.31).

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 the blocky-, long horn-, short horn-, and linear-fruited population, respectively (Additional file 10: Table S3). These fruit shape-specific loci may have been under selection during breeding or were selected owing to linkage with genes that determine fruit traits.

Identification of a core-SNP set

The perfect SNP 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, varieties with identical genotypes may be redundant and were discarded to build non-redundant genotype varieties. Thus, a

minimum of 27 perfect SNPs could distinguish between all 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 to 95 pepper varieties with two allele-specific forward primers and one common reverse primer. The results showed that 35 SNP primers (Additional file 11: Table S4; Additional file 4: Figure S4) 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 and Additional file 4: Figure S4; Additional file 11: Table S4).

Genetic structure in pepper varieties

The principal component analysis (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 were 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 were more closely related to the short horn-fruited varieties than to either the 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 throughout their breeding history. Notably, a selection of both long horn- and short-fruited varieties showed close relatedness to the linear-fruited population.
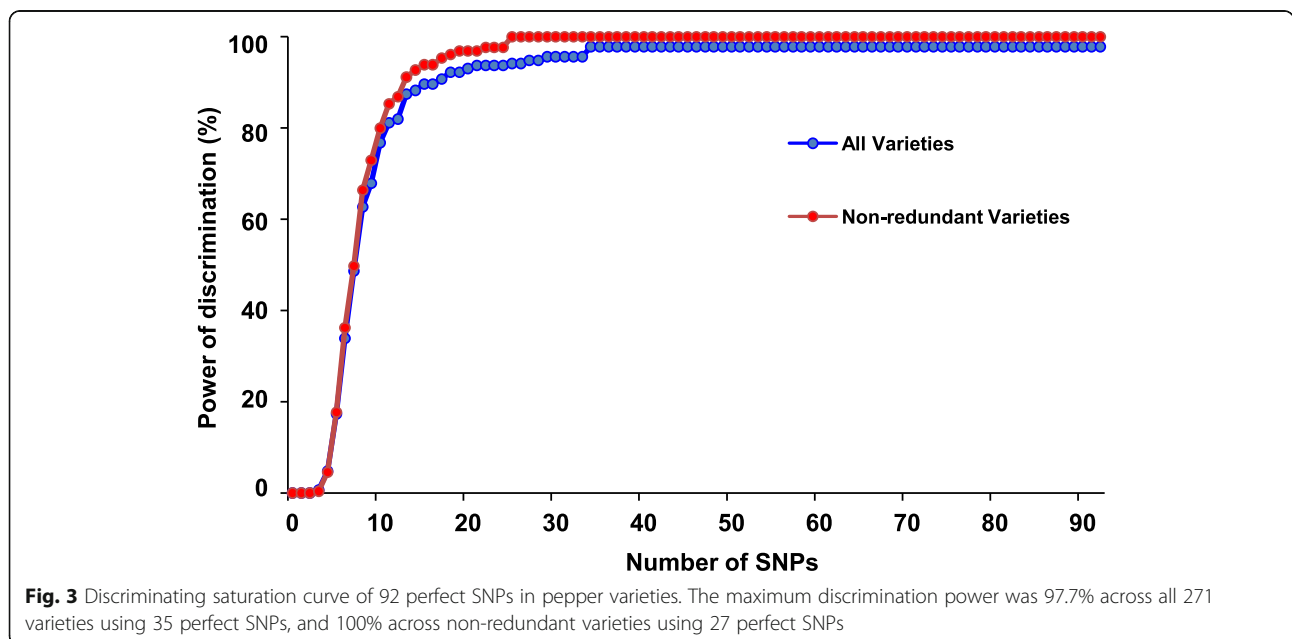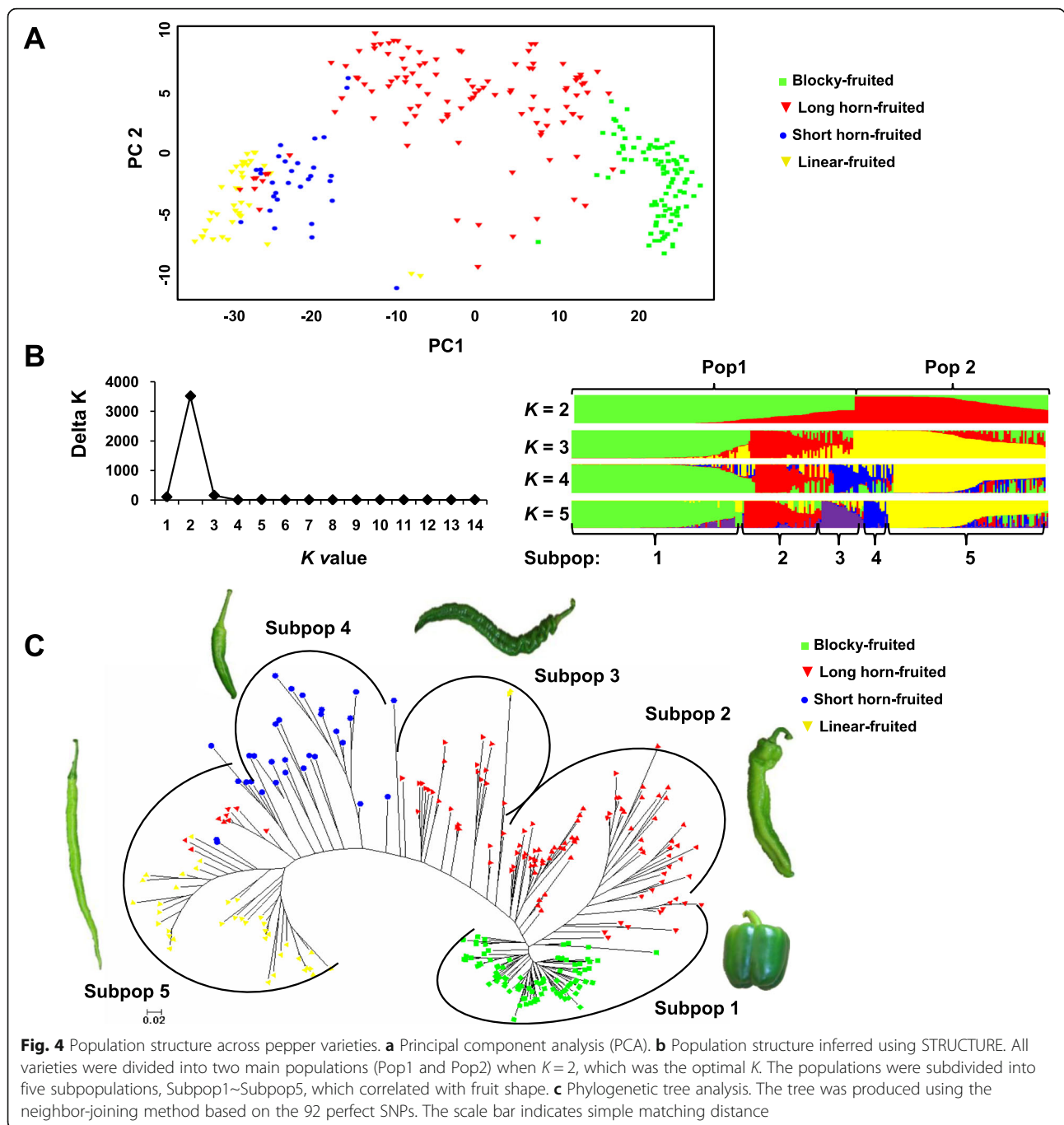

Fig. 3 Discriminating saturation curve of 92 perfect SNPs in pepper varieties. The maximum discrimination power was 97.7% across all 271 varieties using 35 perfect SNPs, and 100% across non-redundant varieties using 27 perfect SNPs

Du *et al. BMC Plant Biology*   (2019) 19:578

Page 6 of 16

Fig. 4 Population structure across pepper varieties. **a** Principal component analysis (PCA). **b** Population structure inferred using STRUCTURE. All 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 simple matching distance

The population structure of the 271 varieties was further inferred using the cluster program, STRUCTURE, testing for 2 to 5 number of clusters (*K*). Evanno's correction [38] 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 long horn-fruited varieties, and only two linear-fruited varieties (Fig. 4b; Additional file 8: 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; Additional file 8: Table S1). When *K* = 3, Pop1 was subdivided into two clusters, blocky- or long horn-fruited types. At *K* = 4, a mixture of 56% short horn-, 15 long horn-, and two linear-fruited varieties were assigned to a new cluster from Pop2, and these short horn- and a new long horn-fruited groups were assigned to independent clusters, respectively, when *K* = 5. Of note, the linear-fruited types were never assigned to an

independent cluster as K was increased. Considering that 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; Additional file 8: Table S1). Subpop1, 2, 3, and 4 showed a clear-cut structure with no or very few admixtures. Subpop1 comprised 98 varieties, 90 of which belong to blocky-fruited varieties and the remaining eight to 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 varieties and the remaining two were linear-fruited varieties. Subpop4 comprised 14 short horn-fruited varieties. Consistent with the results of PCA analyses, admixtures were 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 supported the division of pepper varieties into five well-differentiated genetic populations, which were correlated with distinct fruit shapes, indicating that the genetic structure of these cultivated varieties may have been strongly affected by fruit shape selection through breeding practices.

Genetic variation assessment of pepper populations

Comparison of the results between Pop1 and Pop2 using analysis of molecular variance (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 variations between Pops and among Subpops, a randomization test was performed (Additional file 5: Figure S5). The output revealed six histograms representing the distribution of the randomization strata. The observed results in the output showed significant differentiation of the structure of Pops and Subpops considering all levels of Pops and Subpops strata (Additional file 5: Figure S5). These results also supported the separation of the varieties into two Pops and five Subpops. Furthermore, pairwise estimates of F_{st} showed that population differentiation between Pop1 and Pop2 was 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 varieties) to 0.48 between Subpop1 (mostly blocky-fruited varieties) and Subpop4 (short horn-fruited varieties) (Table 3). Notably, high genetic differentiation ($F_{st} = 0.43$) was observed between Subpop1 and Subpop5 (mostly consisting of linear-fruited varieties), whereas lower genetic differentiation was observed between Subpop4 and Subpop5 ($F_{st} = 0.14$).

Identification of the loci associated with fruit shape

A wide range of variation was observed for fruit shape index (FSI) in the 271 pepper varieties (Additional file 12: Table S5). The average FSI was 1.34, 4.98, 4.70, and 16.56 in blocky-, long horn-, short horn-, and linear-fruited populations, respectively. Significant differences were observed among blocky-, horn-, and linear-shaped populations ($p < 0.01$), but no differences were detected between long horn and short horn-fruited populations. FSI values of more than 9.5 are typical of linear fruits.

Having observed concordance between the population structure and fruit shapes (Fig. 2), we next performed association analyses of the FSI in 271 varieties and 165 genetic loci, including 92 SNPs and an additional 73 SSRs, which were all detected using Target sequencing (Additional file 6: Figure S6). Using the $K + Q$ mixed linear model (MLM), a total of nine loci (CaSSR013, CaSSR090, CaSSR105, CaSSR091, CaSSR039, CaSSR044, CaSSR107, CaSSR077, and CaSNP112) were identified as

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	20,288.23 / 20,288.23	45.03 / 41.26	100.00 / 100.00%

Du *et al. BMC Plant Biology*   (2019) 19:578

Page 8 of 16

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

significantly associated with FSI under a threshold *p*-value of 0.0001 (Additional file 7: Figure S7; Additional file 13: Table S6). To pair the associations with previously identified quantitative trait loci (QTL), the physical position of the nine loci in both the reference genome of Zunla-1 [3] and CM334 [13] are provided in Table 4. Loci CaSSR091 and CaSSR039 are within 820 kb on the same chromosome and were considered a unit. Therefore, the nine loci were located at eight chromosomal regions on six chromosomes, including chromosomes 1, 2, 3, 4, 6, and 12, and the phenotypic variation explained by each locus ranged from 7.9 to 12.7%. Two loci, CaSSR044 and CaSSR107, spanning approximately 39 Mb on chromosome 6, explained the highest phenotypic variation, which was 12.4 and 12.7%, respectively (Table 4 and Additional file 13: Table S6; Additional file 7: Figure S7).

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 [35]. This sequencing technology can acquire thousands of data points in under 72 h, costs less

than \$7/sample, and is associated with genotyping accuracy up to 100% due to the high coverage. The cost of Target SNP-seq developed in this study was similar as that of Target SSR-seq because the same procedure was used for target library construction in these two technologies.

In this study, re-sequencing tools were used to identify 92 perfect SNPs from the genomes of 35 *C. annuum* based on strict screening criteria. Only 9.8% of the perfect SNPs fell within genic regions, which is in agreement with the previous result that variant density is significantly lower in the genic region than in the intergenic regions [30]. The identified perfect SNPs were then used for target SNP sequencing to assess genetic diversity across 271 pepper varieties that are popular in China. The results showed that the perfect SNP panel had a high discriminating capacity for varieties, as 71.74% of the perfect SNPs had PIC values of > 0.30 (Additional file 10: Table S3). Further, a minimum of 27 perfect SNPs could distinguish between all non-redundant varieties (Fig. 3). Notably, the mean PIC value was found to be 0.31, which is lower than the values derived from studies using SSR markers [17, 39]. These discrepancies may be explained by the nature of the different types of markers; SSRs are multiallelic and more polymorphic than SNP markers, which are biallelic [40]. Another reason for the discrepancies may be due to the commercial varieties tending to be less variable compared to the landraces or the wide germplasm collection.

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; Additional file 11: Table S4). Unlike SSR markers, SNP markers do not require reference cultivars to be included in each experiment and will also overcome the confusion between labs regarding SSR alleles.

Table 4 Loci significantly associated with fruit shape index as identified by association analysis

Loci[a]	CM334 (v.1.55) [13]		Zunla-1 (v2.0) [3]		*p*-value	Phenotypic variation explained (%)
	Chr.	Physical position (bp)	Chr.	Physical position (bp)		
CaSSR013	1	2,895,632	1	2,383,961	2.16E-06	9.7
CaSSR090	2	153,032,594	2	143,851,440	1.36E-06	8.2
CaSSR105	3	143,624,183	3	95,195,822	6.99E-05	9.1
CaSSR091	4	762,569	4	214,859,753	2.17E-05	8.9
CaSSR039	4	ND	4	215,677,699	1.99E-05	8.9
CaSSR044	6	234,686,966	6	1,041,327	6.95E-08	12.7
CaSSR107	6	193,777,105	6	39,731,428	1.88E-05	12.4
CaSSR077	12	226,618,758	12	9,517,953	7.54E-08	10.1
CaSNP112	12	47,349,405	12	181,587,490	8.45E-06	7.9

[a]The interval between CaSSR091 and CaSSR039 was within ~ 820 kb; thus, these two loci were considered at the same associated region

Population structure among inbred *C. annuum* lines

Since their initial domestication in Mexico, peppers have been under strong selection for fruit shape and size [56]. Consumption habits and pepper type preference vary globally. In the US alone, more than 20 market types are recognized and consumed [57]. In China, most 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 [58, 59]. To date, most experiments have evaluated the genetic relationships among several *Capsicum* species [29, 30, 36, 37, 60–62] or the genetic diversity of *C. chinense* and *C. baccatum* germplasm from relatively restricted regions [26, 63]. Phylogenetic analysis based on molecular markers, pangenome sequencing, and GBS confirmed that *C. chinense* and *C. frutescens* are more closely related to each other than to *C. annuum* [29, 61, 64]. Several studies attempted to characterize the population relatedness of cultivated *C. annuum* in restricted geographical areas [29, 36, 40–42, 65]. They revealed that the population structures of the *C. annuum* accessions were mainly associated to distinct cultivar types with respect to the plant and fruit descriptors, and thus mostly result from human selection for cultivar types in agreement with consumption modes and adaptation to the highly diversified agro-climatic conditions. 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 [29, 41, 65]. Further, clustering of blocky-fruited peppers in the furthest positions relative to small hot Chiltepin-like types can also be observed in previous studies [29, 41, 65].

Genetic structure among *C. annuum* varieties

Although the previous work has shown that population of *C. annuum* landraces in China clusters according to cultivar type [36], the relationships among commercially important *C. annuum* varieties from different companies have not been investigated with a fine set of genetic markers. 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 [41–44]. 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 (Additional file 5: Figure S5). The best fit of genetic structures of the pepper lines and varieties were both divided into two groups in this study, different to that observed in the 368 Chinese *C. annuum* accessions analyzed by Zhang et al. [36, 18], which included 28 SSR markers to structure the accessions into three STRUCTURE groups. These differences may be attributed to the different types of pepper materials and the number of markers used in the two studies. However, the clustering of fruit types in both studies appeared to be somewhat similar, although different classifications of fruit shape were used. For example, Group1 mainly included rectangular, square, and triangular fruit types [36], which were also mainly clustered in Pop1 of our study. Group 3 mainly comprised cultivars with small and long fruits characterized by a very high fruit length: width ratio [36], which is the characteristic of linear-fruited and some short horn-fruited varieties in Pop2 of this study. In summary, our study provided valuable insight into the population structure underlying the fruit shapes of pepper varieties, as well as confirmed the strong effect of fruit shape selection by breeders on the genetic structure of Chinese pepper varieties.

Identification of associated loci for fruit shape

Fruit shape is an important trait in pepper breeding programs. A number of QTLs controlling FSI have been identified in intraspecific and interspecific populations from a cross between the bell pepper and a small-fruited hot pepper [45–53]. The first FSI QTL in peppers, named *fs3.1*, was detected on linkage group 3 [48]. This QTL was subsequently detected in other linkage analyses [46, 47, 51]. Using genome-wide associations in 373 pepper accessions, Colonna et al. (2019) recently identified that the SNP 3:183386147 on chromosome 3, located in the exon of gene CA03g16080, is significantly associated with FSI [30]. In the current study, the FSI associated loci CaSSR105 and its nearest upstream marker, CaSNP029, are located within approximately 26 Mb intervals on chromosome 3 (Additional file 13: Table S6).

Du *et al. BMC Plant Biology*   (2019) 19:578

Page 10 of 16

This FSI association region covers the reported FSI associated loci SNP 3:183386147 in the gene *Longifolia 1-like* (CA03g16080) [30]. Zygier et al. (2005) mapped a fruit shape QTL (*fs2.1*) on chromosome 2 [50]. This QTL was also detected in other studies and was found to be close to the *Ovate* gene (CA02g22830) at approximately 158 Mb of the CM334 genome [52, 53]. In the current study, FSI associated loci CaSSR090 and its nearest downstream loci CaSSR024, with an interval spanning of approximately 12.6 Mb on chromosome 2, covered the reported *Ovate* gene (CA02g22830). The *Ovate* gene was initially discovered in the tomato, where it controlled the fruit shape transformation from round to pear-shaped fruit [54, 55]. A single FSI QTL (*fs4.2*) was detected at the end of chromosome 4, which explained the 26.1% phenotypic variation [50]. We found that two FSI associated loci, CaSSR091and CaSSR039, were located between CaSNP041 and the end of chromosome 4, covering approximately 12.9 Mb (Additional file 13: Table S6). The presence of FSI association loci on chromosomes 1, 6, and 12 was also detected in this study (Additional file 6: Figure S6 and Additional file 7: Figure S7; Table 4 and Additional file 13: Table S6). After screening for protein function in these association loci, we found that two *Ovate* genes, CA06g21580 and CA12g07370, could be considered candidate genes, as they had significant effects on fruit shape.

Future directions of target SNP-seq in pepper
Of note, foreground selecting markers that are suitable for specific primer design could also be added to the perfect SNP panel used in our Target SNP-seq. For example, based on the functional site of the *Tobamovirus* resistance gene L^3 and L^4 [66], the *Phytophthora capsici* resistance genes *CaDMR1* and Phyto5NBS1 [67, 68], bacterial spot resistance gene *Bs3* [69], and *potato virus Y* resistance gene *pvr1* [70], specific primers at the flanking region of the functional site have successfully been developed (Additional file 14: Table S7) and added to the perfect SNP panel. These functional loci of resistance genes, combined with the perfect SNP and SSR markers, could be detected simultaneously across hundreds of pepper accessions through Target SNP-seq. 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.

Conclusions
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 revealed that the genetic structure of commercially available pepper varieties in China had been significantly influenced by fruit shape selection through breeding. Finally, association analysis of a limited number of markers allowed for the identification of previously reported and novel genomic regions that control fruit shape.

Methods
Plant materials, fruit shape categorization, and DNA extraction
A total of 271 pepper varieties, which were kindly supplied by 60 different breeding companies in China, were analyzed in this study. Information on these hybrid seeds, including variety name and source, is available in Additional file 8: Table S1. Fruit trait investigation and genetic identification were carried out by the pepper genetic breeding group and high-throughput molecular breeding platform at the Beijing Vegetable Research Center (BVRC).

Varieties were planted under greenhouse conditions at the Vegetable Varieties Exhibition Center in the Tongzhou District of Beijing. The greenhouse temperature ranged from 25 to 30 °C (08:00−20:00) and 20−25 °C (20:00−08:00), with natural light. Each variety consisted of at least four plants. The fruits were categorized into one of four fruit shapes: blocky-, long horn-, short horn-, and linear-fruited types (Additional file 8: Table S1).Examples of fruit shape classification are presented in Additional file 1: Figure S1. Four to ten ripe fruits from each plant were subjected to measurements of maximum height and width using a Vernier caliper (Hangzhou Tool and Measuring Tool Company, Hangzhou, China).

DNA was extracted from four young plantlet randomly selected from individuals of each variety using a CTAB-based method [71]. The DNA integrity was assessed using 1.5% (w/v) agarose gel electrophoresis, and the concentration was determined using a Nanodrop 2000 Spectrophotometer (Thermo Fisher Scientific, DE, USA).

Re-sequencing and perfect SNP identification
In total, 31 diverse pepper lines (*C. annuum*), including 30 inbred lines from our ongoing breeding programs in BVRC and PI640446 provided by the U.S. National Plant Germplasm System, were selected for re-sequencing on the Illumina X Ten platform at Shanghai Majorbio Biopharm Technology Co. Ltd. (Shanghai, China). The 31 pepper lines had diverse genetic backgrounds and horticultural traits, including eight blocky-fruited lines, 13 long horn-fruited lines, five short horn-fruited lines and five 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 [72]. The clean reads were then mapped to the reference genome of Zunla-1 chromosome version 2.0 [3] using the Burrows-Wheeler Alignment Tool (BWA) with default parameters, and SNPs were called using the Genome Analysis Toolkit (GATK, v2.4-7g5e89f01) [73]. SNPs with MAF > 5% and missing data < 10% were imported into MEGA to build the rooted phylogenetic tree using the cultivar progenitor, Chiltepin, as an outgroup with the neighbor-joining method [74]. Population structure analysis was completed using STRUCTURE v2.3. The number of populations (K) was determined following the standard procedure [75] 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 Evanno's delta K method [38].

To acquire a dataset of genome-wide SNPs for subsequent Target SNP-seq analysis, perfect SNPs were identified using the following criteria: (i) 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) 2 alleles per locus for the SNPs.

Target SNP-seq

The Target SNP-seq procedure was completed as previously described using the SNPs identified above [35]. 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 (Additional file 9: Table S2); 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 °C for 5 min followed by 17 cycles at 95 °C for 30 s, 60 °C for 4 min, 72 °C for 4 min with a final 4 min extension at 72 °C. 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 3 M *Taq* enzyme (Molbreeding Biotechnology Company, Shijiazhuang, China), 8 μl nuclease-free water, and 1 μl of primers with the following sequences: forward 5′- AATGATACGGCGAC CACCGAGATCTACACTCTTTCCCTACACGA -CGCTCTTCCG-3′ and reverse 5′-CAAGCAGAAG ACGGCATACGAGAT -XXXXXXXXGTGACTGGAG

TTCCTTGGCACCCGAGA-3′ (barcodes are indicated by underlined sequences). The PCR procedure was 95 °C for 3 min; 7 cycles of 95 °C for 15 s, 58 °C for 15 s, and 72 °C for 30 s with a final 4 min extension at 72 °C. The PCR products were then purified with 100 μl of 80% alcohol and 23 μl Tris-HCl buffer (10 mM, pH 8.0–8.5). After that, a Target SNP-seq library was sequenced using an Illumina X Ten platform at Molbreeding Biotechnology Company (Shijiazhuang, China).

SNP genotype analysis of target SNP-seq

The raw data from the Target SNP-seq was demultiplexed to determine the exact genotypes for each variety based on the sample-specific barcodes using the Illumina bcl2fastq pipeline (Illumina, San Diego, 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. Sequence depth, alignment rate, as well as target alignment rate and uniformity for each variety were calculated as follows to evaluate the results of targeted sequencing.

$$\text{Alignment rate} = \text{the number of reads aligned on the genome/total reads.}$$

$$\text{Targetalignment rate} = \text{the number of reads aligned on the target region/total reads.}$$

Uniformity inferred to the proportion of the SNPs with a depth of > 10% of the average depth.

$$\text{Depth of each SNP} = \text{total base generated in the perfect SNP/read length.}$$

$$\text{Average depth} = S/(M \times N \times L),$$

where S indicates the total base generated from Target SNP-seq; M indicates the total number of varieties; N indicates the total number of SNPs; L indicates the average read length.

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 higher 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.

Sequence depth for each variety = Total base generated from variety / total base length of the 92 targeted genome region.

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) [76] were calculated using a Perl script with the following equation:

$$PIC = 1 - \sum_{i=1}^{l} P_i^2 - \sum_{i=1}^{l-1} \sum_{j=i+1}^{l} 2P_i^2 P_j^2$$

where l is the allele locus, and P_i and P_j represent the population frequency of the ith 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: PCA, STRUCTURE, and a phylogenetic tree. PCA was carried out using the FactoMineR package in R [77]. The Bayesian-based model procedure implemented by STRUCTURE v2.3 [75, 78] 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 [79, 80].

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 of populations and subpopulations, AMOVA and pairwise F_{st} were performed using *poppr* R package and the function pairwise.neifst in the *Hierfstat* R package, respectively [81, 82]. Randomization tests were further performed to test the significance of differentiation using the function randtest in the *ade4* package [81]. Detailed instructions for the AMOVA and randomization tests are available at https://grunwaldlab.github.io/Population_Genetics_in_R/AMOVA.html and https://rdrr.io/cran/poppr/man/poppr.amova.html.

Discrimination power of the perfect SNPs

To determine a minimal number of SNPs for distinguishing the maximum number of pepper varieties, a Perl script was developed to determine the best discrimination power for 1 to 92 perfect SNPs according to the following algorithm.

1) Selection of the first SNP: a) pairwise comparison between varieties were conducted for each SNP, which included 36,585 comparisons for each SNP; b) $X_{ij} = 1$ if genotype difference existed for j^{th} pairwise comparison of ith SNP (i = 1, 2, 3, …, 92; j = 1, 2, 3, …, 36,585). $X_{ij} = 0$ if genotype difference did not exist for j^{th} pairwise comparison of ith SNP (i = 1, 2, 3, …, 92; j = 1, 2, 3, …, 36, 585); c) The SNP with a maximum value of $\sum_{j=1}^{36585} X_{ij}$ among the 92 SNPs was selected as the first SNP. 2) Selection of the best two SNPs: a) 91 SNP combinations of the first selected SNP and each of the rest 91 SNPs were formed; b) 36,583 pairwise comparisons were conducted for each of the 91 SNP combination; c) $X_{mj} = 1$ if genotype difference existed for j^{th} pairwise comparison of mth SNP combination (m = 1, 2, 3, …, 91; j = 1, 2, 3, …, 36, 583). $X_{mj} = 0$ if genotype difference did not exist for j^{th} pairwise comparison of mth SNP combination (m = 1, 2, 3, …, 91; j = 1, 2, 3, …, 36,583); d) the SNP combination with a maximum value of $\sum_{j=1}^{36585} X_{mj}$ was selected as the best two SNPs. If the SNP combinations had the same values, the SNP combination with the second SNP located at a different chromosome as first SNP was preferentially selected as the best two SNPs. 3) Selection of the best three SNPs: a) 90 SNP combinations of the best two SNPs and each of the rest 90 SNPs were formed; b), c), and d) steps were conducted similarly as that in step 2) to select the best three SNPs. If the SNP combinations had the same values, the SNP combination with the third SNP located at different chromosome as the first and second SNPs was preferentially selected as the best three SNPs. The best 4 to 92 SNPs were selected gradually as in step 3). Discrimination power for 1 to 92 best SNPs was calculated using the following formula: discrimination power = the number of varieties showing unique genotypes / 271. The saturation curve was plotted by discrimination power for 1 to 92 best SNPs (Fig. 3). High discrimination power referred to high saturation value and high SNP discernibility.

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 to 95 commercial varieties were then used to assess the potential utility of the SNP markers through the KASPar platform; fluorescence was detected as previously described [83]. Detailed instructions are available at www.kbioscience.co.uk. The Perl script, as mentioned above, was used to select a core-SNP set from the successfully verified SNP markers. Finally, the SNP markers associated with the maximum variety discrimination (the highest saturation value) were identified as a core-SNP set. The primer sequences of the core-SNP markers are shown in Additional file 11: Table S4.

Association analysis

FSI for each variety was calculated as the ratio of maximum height to maximum width. Ninety-two SNP loci (Additional file 9: Table S2) combined with 73 SSR loci (Additional file 6: Figure S6), all developed in this study and detected by target sequencing across 271 varieties, were used for association analysis. The methods used for SSRs target library construction and detection were the same as those used in Target SNP-seq. The software program TASSEL 5.2.25 was used for association analysis. The MLM that considered both the 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 more effective than other models in detecting loci [84, 85], 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 MLM. 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 [3].

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s12870-019-2122-2.

Additional file 1: 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 10.0–35.0 cm long, without shoulder and lobe, including Xian Jiao, Tiao Jiao and Mei Ren Jiao, as named in China.

Additional file 2: Figure S2. Target SNP-seq genotyping analysis results. Distribution of the average read depths (A), reads alignment rate to the pepper reference genome (B), target region alignment rate (C), and uniformity for 271 pepper varieties (D).

Additional file 3: 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).

Additional file 4: Figure S4. Kompetitive allele-specific PCR (KASPar) results of the 35 core SNP markers genotyped across 23 to 95 pepper varieties.

Additional file 5: Figure S5. Significance testing of differentiation between Pops and among 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.

Additional file 6: Figure S6. Chromosomal map of a subset of markers used in association analysis with fruit shape index (FSI). The physical position of each marker on 12 chromosomes of C. annuum Zunla-1 [3] are shown between brackets. Significantly associated markers are shown in red color.

Additional file 7: Figure S7. Manhattan plots (A) and quantile-quantile plots (B) of fruit shape index (FSI) in the 271 pepper varieties. Red dashed line indicates high probability of associated loci with FSI.

Additional file 8: Table S1. Classification of and information on the pepper varieties used in this study.

Additional file 9: Table S2. Multiplexed primers panel of the 92 perfect SNPs used for Target SNP-seq.

Additional file 10: Table S3. Characteristics of the 92 perfect SNPs and the diversity detected in the 271 pepper varieties and four fruit shape populations.

Additional file 11: Table S4. Primer sequences of the 35 core-SNP markers developed in this study.

Additional file 12: Table S5. Range, mean, and standard deviations (SD) collected for fruit shape index (FSI) in the pepper varieties.

Additional file 13: Table S6. Association regions with fruit shape index (FSI) in the reference genomes of CM334 and Zunla-1.

Additional file 14: Table S7. Specific primers used to detect functional resistance loci against four pepper diseases through Target SNP-seq.

Abbreviations

AFLP: Amplified fragment length polymorphism; AMOVA: Analysis of molecular variance; FSI: Fruit shape index; GATK: Genome Analysis Toolkit; GBS: Genotyping-by-sequencing; GBTS: Genotyping by target sequencing; GLM: General linear model; H_e: Expected heterozygosity; H_o: Observed heterozygosity; InDel: Insertion or deletion; KASPar: Kompetitive allele-specific PCR; MAF: Minor allele frequency; MLM: Mixed linear model; PCA: Principal component analysis; PIC: Polymorphism information content; RAPD: Random amplified polymorphic; RFLP: Restriction fragment length polymorphism; SNP: Single-nucleotide polymorphism; SSR: Simple sequence repeats

Authors' contributions

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

Availability of data and materials

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

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Beijing Vegetable Research Center (BVRC), Beijing Academy of Agricultural and Forestry Sciences, Beijing 100097, China. [2]Beijing Key Laboratory of Vegetable Germplasm Improvement, National Engineering Research Center for Vegetables, Beijing 100097, China. [3]Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China.

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. https://doi.org/10.17660/ActaHortic.2007.745.5.
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. 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. 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. 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. 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. 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. 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. 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. 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. https://doi.org/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. 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. 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. 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. 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

17. mapping the QTLs affecting the initiation of flower primordia in pepper (*Capsicum* spp.). Plos One. 2015;10:e0119389. https://doi.org/10.1371/journal.pone.0119389.
17. 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. Plant Mol biol Report. 2012;30:1470–9. https://doi.org/10.1007/s11105-012-0466-y.
18. 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. https://doi.org/10.1007/s11105-012-0466-y.
19. 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. https://doi.org/10.3732/ajb.0800155.
20. 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. https://doi.org/10.1007/s10722-011-9744-z.
21. 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 Harb Symp Quant Biol. 2003;68:69–78. https://doi.org/10.1101/sqb.2003.68.69.
22. Steemers FJ, Gunderson KL. Whole genome genotyping technologies on the BeadArray™ platform. Biotechnol J. 2007;2:41–9. https://doi.org/10.1002/biot.200600213.
23. 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. https://doi.org/10.1101/gr.4.6.357.
24. Kang JH, Yang HB, Jeong HS, Cheo P, Kwon JK, Kang BC. Single nucleotide polymorphism marker discovery from transcriptome sequencing for marker-assisted backcrossing in *Capsicum*. Kor J Hortic Sci Technol. 2014;32:535–43. https://doi.org/10.7235/hort.2014.14109.
25. 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 Genom. 2016;17:943. https://doi.org/10.1186/s12864-016-3297-7.
26. Nimmakayala P, Abburi VL, Saminathan T, Almeida A, Davenport B, Davidson J, Reddy CV, Hankins G, Ebert A, Choi D, Stommel J. Genome-wide divergence and linkage disequilibrium analyses for *Capsicum baccatum* revealed by genome-anchored single nucleotide polymorphisms. Front PlantSci. 2016;7:1646. https://doi.org/10.3389/fpls.2016.01646.
27. Nimmakayala P, Abburi VL, Saminathan T, Alaparthi SB, Almeida A, Davenport B, Nadimi M, Davidson J, Tonapi K, Yadav L, Malkaram S, Vajja G, Hankins G, Harris R, Park M, Choi D, Stommel J, Reddy UK. Genome-wide diversity and association mapping for Capsaicinoids and fruit weight in *Capsicum annuum* L. Sci Rep. 2016;6:38081. https://doi.org/10.1038/srep38081.
28. Taitano N, Bernau V, Jardón-Barbolla L, Leckie B, Mazourek M, Mercer K, McHale L, Michel A, Baumler D, Kantar M, van der Knaap E. Genome-wide genotyping of a novel Mexican Chile Pepper collection illuminates the history of landrace differentiation after *Capsicum annuum* L. domestication. Evol Appl. 2018;12:78–92. https://doi.org/10.1111/eva.12651.
29. Pereira-Dias L, Vilanova S, Fita A, Prohens J, Rodríguez-Burruezo A. Genetic diversity, population structure, and relationships in a collection of pepper (*Capsicum* spp.) landraces from the Spanish centre of diversity revealed by genotyping-by-sequencing (GBS). Horticulture Res. 2019;6:54. https://doi.org/10.1038/s41438-019-0132-8.
30. Colonna V, D'Agostino N, Garrison E, Albrechtsen A, Meisner J, Facchiano A, Cardi T, Tripodi P. Genomic diversity and novel genome-wide association with fruit morphology in *Capsicum*, from 746k polymorphic sites. Sci Rep. 2019;9:10067. https://doi.org/10.1038/s41598-019-46136-5.
31. 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 Res. 2017;45. https://doi.org/10.1093/nar/gkx093.
32. 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. https://doi.org/10.1073/pnas.1619268114.

Du et al. BMC Plant Biology   (2019) 19:578

Page 15 of 16

33. 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. https://doi.org/10.1186/1471-2164-15-1105.

34. 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. https://doi.org/10.1007/s11032-019-0940-4.

35. 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. https://doi.org/10.3389/fpls.2019.00531.

36. Zhang XM, Zhang ZH, Gu XZ, Mao SL, Li XX, Chadoeuf J, Palloix A, Wang LH, Zhang BX. Genetic diversity of pepper (Capsicum spp.) germplasm resources in China reflects selection for cultivar types and spatial distribution. J Integr Agric. 2016;15:1991–2001. https://doi.org/10.1016/S2095-3119(16)61364-3.

37. Meng CY, Wei XC, Zhao YY, Yuan YX, Yang SJ, Wang ZY, Zhang XW, Sun JW, Zheng XL, Yao QJ, Zhang Q. Genetic diversity analysis of Capsicum genus by SSR markers. Mol Plant Breed. 2017;8:70–8. https://doi.org/10.5376/mpb.2017.08.0008.

38. 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. https://doi.org/10.1111/j.1365-294X.2005.02553.x.

39. 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. https://doi.org/10.1007/s00122-003-1467-x.

40. 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. https://doi.org/10.1186/s12864-016-3297-7.

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. https://doi.org/10.1371/journal.pone.0056200.

42. Solomon AM, Han K, Lee J-H, Lee H-Y, Jang S, Kang B-C. Genetic diversity and population structure of Ethiopian Capsicum germplasms. PLoS One. 2019;14:e0216886. https://doi.org/10.1371/journal.pone.0216886.

43. 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. https://doi.org/10.1007/s10722-013-0006-0.

44. 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. https://doi.org/10.1007/s00122-009-1102-6.

45. Chaim AB, Paran I, Grube RC, Jahn M, van Wijk R, Peleman J. QTL mapping of fruit-related traits in pepper (Capsicum annuum). Theor Appl Genet. 2001; 102:1016–28. https://doi.org/10.1007/s001220000461.

46. Han K, Jeong HJ, Yang HB, Kang SM, Kwon JK, Kim S, Choi D, Kang BC. An ultra-high-density bin map facilitates high-throughput QTL mapping of horticultural traits in pepper (Capsicum annuum). DNA Res. 2016;23:81–91. https://doi.org/10.1093/dnares/dsv038.

47. Yarnes SC, Ashrafi H, Reyes-Chin-Wo S, Hill TA, Stoffel KM, VanDeynze A. Identifcation of QTLs for capsaicinoids, fruit quality, and plant architecture-related traits in an interspecifc Capsicum RIL population. Genome. 2013;56: 61–74. https://doi.org/10.1139/gen-2012-0083.

48. 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. https://doi.org/10.1007/s00122-002-1132-9.

49. Rao GU, Chaim AB, Borovsky Y, Paran I. Mapping of yield-related QTL in pepper in an interspecific cross of Capsicum annuum and C. frutescens. Theor Appl Genet. 2003;106:1457–66. https://doi.org/10.1007/s00122-003-1204-5.

50. Zygier S, Chaim AB, Efrati A, Kaluzky G, Borovsky Y, Paran I. QTL mapping for fruit size and shape in chromosomes 2 and 4 in pepper and a comparison of the pepper QTL map with that of tomato. Theor Appl Genet. 2005;111: 437–45. https://doi.org/10.1007/s00122-005-2015-7.

51. Barchi L, Lefebvre V, Sage-Palloix AM, Lanteri S, Palloix A. QTL analysis of plant development and fruit traits in pepper and performance of selective phenotyping. Theor Appl Genet. 2009;118:1157–71. https://doi.org/10.1007/s00122-009-0970-0.

52. Hill TA, Chunthawodtiporn J, Ashrafi H, Stoffel K, Weir A, Van Deynze A. Regions underlying population structure and the genomics of organ size determination in Capsicum annuum. Plant Genome. 2017. https://doi.org/10.3835/plantgenome2017.03.0026.

53. 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. https://doi.org/10.3835/plantgenome2016.12.0125.

54. 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. https://doi.org/10.3389/fpls.2014.00227.

55. 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. https://doi.org/10.1038/s41467-018-07216-8.

56. 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. https://doi.org/10.1073/pnas.1308933111.

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

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

59. 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.

60. 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. https://doi.org/10.1016/j.scienta.2018.03.012.

61. 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;220:360–3. https://doi.org/10.1111/nph.15413.

62. Lee HY, Ro NY, Jeong HJ, Kwon JK, Jo J, Ha Y, Jung A, Han JW, Venkatesh J, Kang BC. Genetic diversity and population structure analysis to construct a core collection from a large Capsicum germplasm. BMC Genet. 2016;17:142. https://doi.org/10.1186/s12863-016-0452-8.

63. 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. https://doi.org/10.1007/s10722-013-0069-y.

64. 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. https://doi.org/10.21273/JASHS.129.6.0826.

65. 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. https://doi.org/10.1371/journal.pone.0116276.

66. 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. https://doi.org/10.1007/s11032-011-9666-7.

67. 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. https://doi.org/10.3835/plantgenome2014.03.0011.

68. Liu WY, Kang JH, Jeong HS, Choi HJ, Yang HB, Kim KT, Choi D, Choi GJ, Jahn M, Kang BC. Combined use of bulked segregant analysis and microarrays reveals SNP markers pinpointing a major QTL for resistance to Phytophthora capsici in pepper. Theor Appl Genet. 2014;127:2503–13. https://doi.org/10.1007/s00122-014-2394-8.

69. 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;318:645–8. https://doi.org/10.1126/science.1144958.

70. 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. https://doi.org/10.1007/s00122-005-0120-2.

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

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

73. 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. https://doi.org/10.1101/gr.107524.110.

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

75. 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.

76. 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.

77. 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/

78. 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.

79. 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.

80. 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. https://doi.org/10.7717/peerj.281.

81. Excoffier L, Smouse PE, Quattro JM. Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics. 1992;131:479–91.

82. de Meeus T, Goudet J. A step-by-step tutorial to use HierFstat to analyse populations hierarchically structured at multiple levels. Infect Genet Evol. 2007;7:731–5. https://doi.org/10.1016/j.meegid.2007.07.005.

83. 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. https://doi.org/10.1007/s11032-018-0795-0.

84. 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. https://doi.org/10.1186/s12864-015-1226-9.

85. 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. https://doi.org/10.3835/plantgenome2017.03.0026.

Publisher's Note