RESEARCH ARTICLE

High-Throughput Development of SSR Markers from Pea (*Pisum sativum* L.) Based on Next Generation Sequencing of a Purified Chinese Commercial Variety

Tao Yang[1]☯, Li Fang[1]☯, Xiaoyan Zhang[2], Jinguo Hu[3], Shiying Bao[4], Junjie Hao[2], Ling Li[5], Yuhua He[4], Junye Jiang[1], Fang Wang[1], Shufang Tian[1], Xuxiao Zong[1]*

1 The National Key Facility for Crop Gene Resources and Genetic Improvement/Institute of Crop Science, Chinese Academy of Agricultural Sciences, Beijing, China, **2** Qingdao Academy of Agricultural Sciences, Qingdao, China, **3** USDA-ARS Western Regional Plant Introduction Station, Pullman, Washington, United States of America, **4** Institute of Grain Crops, Yunnan Academy of Agricultural Sciences, Kunming, China, **5** Institute of Cash Crops, Liaoning Academy of Agricultural Sciences, Liaoyang, China

☯ These authors contributed equally to this work.
* zongxuxiao@caas.cn

Citation: Yang T, Fang L, Zhang X, Hu J, Bao S, Hao J, et al. (2015) High-Throughput Development of SSR Markers from Pea (*Pisum sativum* L.) Based on Next Generation Sequencing of a Purified Chinese Commercial Variety. PLoS ONE 10(10): e0139775. doi:10.1371/journal.pone.0139775

Editor: Shu-Biao Wu, University of New England, AUSTRALIA

Received: April 1, 2015

Accepted: September 17, 2015

Published: October 6, 2015

Data Availability Statement: The raw sequencing files were submitted to the National Center for Biotechnology Information (NCBI) short read archive under accession numbers with the accession number SRX973821. All other relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the National Natural Science Foundation of China (No. 31371695), China Agriculture Research System (No. CARS-09), Social Development of Science and Technology Plan (No. 2013BB010) from Yunnan Government, and also granted by the Agricultural

Abstract

Pea (*Pisum sativum* L.) is an important food legume globally, and is the plant species that J.G. Mendel used to lay the foundation of modern genetics. However, genomics resources of pea are limited comparing to other crop species. Application of marker assisted selection (MAS) in pea breeding has lagged behind many other crops. Development of a large number of novel and reliable SSR (simple sequence repeat) or microsatellite markers will help both basic and applied genomics research of this crop. The Illumina HiSeq 2500 System was used to uncover 8,899 putative SSR containing sequences, and 3,275 non-redundant primers were designed to amplify these SSRs. Among the 1,644 SSRs that were randomly selected for primer validation, 841 yielded reliable amplifications of detectable polymorphisms among 24 genotypes of cultivated pea (*Pisum sativum* L.) and wild relatives (*P. fulvum* Sm.) originated from diverse geographical locations. The dataset indicated that the allele number per locus ranged from 2 to 10, and that the polymorphism information content (PIC) ranged from 0.08 to 0.82 with an average of 0.38. These 1,644 novel SSR markers were also tested for polymorphism between genotypes G0003973 and G0005527. Finally, 33 polymorphic SSR markers were anchored on the genetic linkage map of G0003973 × G0005527 F_2 population.

Introduction

Pea (*Pisum sativum* L.) is one of the most popular food legumes in the world. The harvested area was approximately 6.4 million hectares and production was almost 11 million metric tons of dry peas in 2013 [1]. As one of the most important legumes, pea can be used as vegetable,

Science and Technology Innovation Program (ASTIP) in CAAS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

pulse, and feed. Moreover, pea plays a critical role in crop rotation and low-carbon agriculture for its capacity of biological fixation of atmospheric N_2 [2].

Although significant advances have been made through traditional breeding practices, resulting in semi-leafless pea, snow pea, and snap pea, progress in developing SSR markers [3–6] and marker assisted selection in pea breeding is limited. This is due mainly to the large genome size of pea (4.45 GB), which is approximately 9 times larger than that of barrel medic (*Medicago truncatula* Gaertn.) (http://www.jcvi.org/medicago/), and 4 times larger than that of soybean (*Glycine max* L. Merr.) [7].

A number of next-generation sequencing technologies such as the Roche 454, the Illumina Hiseq 2500 and the Pacific Biosciences PacBio RS II systems have been developed in recent years. These technologies are capable of generating tens of millions of short DNA sequence reads at a relatively low cost. De novo sequencing of genomes, re-sequencing of genomes and RNA-seq were popular all over the world [8–10]. However, only a few researchers utilized Next Generation Sequencing (NGS) platforms for high-throughput development of SSR markers in plant genome [11–16].

The present study aims at obtaining more SSR sequences cheaply and efficiently by using the high-throughput Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA). We report here the result of identifying over 8,899 putative SSR containing sequences, characterizing and validating 1,644 of these newly identified SSRs experimentally using 22 *P. sativum* and two *P. fulvum* genotypes, and enhancing the density of previous genetic linkage map with 33 of these newly identified markers.

Materials and Methods

Plant materials

Widely grown Chinese pea cultivar Zhongwan No. 6, numbered G0005527 in the National Genebank of China, was purified by single seed descend for three consecutive generations. DNA from the resulting plants was used for sequencing and SSR marker development.

For validating the SSRs, a diverse panel of 24 accessions, consisting of 11 entries from China, 11 from other countries and two wild relatives as out-groups, was used in the amplification experiment (Fig 1 and Table 1). These germplasm resources are maintained by the National Genebank of China at the Institute of Crop Science (ICS), Chinese Academy of Agricultural Sciences (CAAS), Beijing, China.

For SSR mapping, a segregating F_2 population of 190 individuals derived from the cross of G0003973 × G0005527 was used. The dry seed color of G0003973 (winter hardy, from Qinghai) was olivine and that of G0005527 (cold sensitive, from Beijing) was green. This population was grown in a protected field at Qingdao Academy of Agricultural Sciences, Qingdao (QdAAS), Shandong, China.

All the plant materials were planted in the protected field of experimental farm within CAAS campus (39° 57' 38" N, 116° 19' 27" E).

DNA extraction, library preparation and next-generation sequencing

Genomic DNA was extracted from 10-day old, etiolated seedlings of each genotype cleared with sterile water, using the CTAB method [17,18]. For the Illumina HiSeq 2500 run, a library was prepared with a commercial kit NEBNext Multiplex Oligos for Illumina with Index Primers Set 2 (New England Biolabs Inc., Ipswich, MA, USA) following the manufacturer's protocol (Paired-End Library Construction). The raw sequencing files were submitted to the National Center for Biotechnology Information (NCBI) short read archive under accession numbers with the accession number SRX973821.

Reads initiative characterization

CLC Genomics Workbench 7.5 software (CLC Inc., Aarhus, Denmark) was used in the following analyses. The quality of paired-end data was checked by the Create Sequencing QC Report Module at default parameters. Subsequent quality trimming was performed with the Trim Sequences Module using quality scores limit of 0.05 and maximum number of ambiguities of 2. The Remove of Duplicate Reads Module was used to filter redundant reads at default parameters. Finally, de novo Assembly Module was used for sequences assembly. These sequences were prepared for further SSRs mining.

SSRs mining

MISA (Microsatellite identification) software, a SSRs motif scanning tool written in Perl (http://pgrc.ipk-gatersleben.de/misa/), was used for the identification and localization of SSRs or microsatellites. The identified motifs were mononucleotide to hexanucleotide, and the minimum repeat unit was defined as 10 for mononucleotide, 6 for dinucleotide, 5 for all the higher order motifs including trinucleotide, tetranucleotide, pentanucleotide and hexanucleotide. Furthermore, the maximal number of interrupting base pairs in a compound microsatellite was 20 bp. The characterizations of SSRs were obtained by statistical analysis from the MISA files. The SSRs information was extracted and statistically analyzed by in-house Perl script, plotted by R language [19].

Primer design

The high throughput primers designing pipeline contained Perl scripts p3_in.pl, p3_out.pl (http://pgrc.ipk-gatersleben.de/misa/primer3.html) and Primer 3.0 software (http://www-genome.wi.mit.edu/genome_software/other/primer3.html). Redundant primers were removed

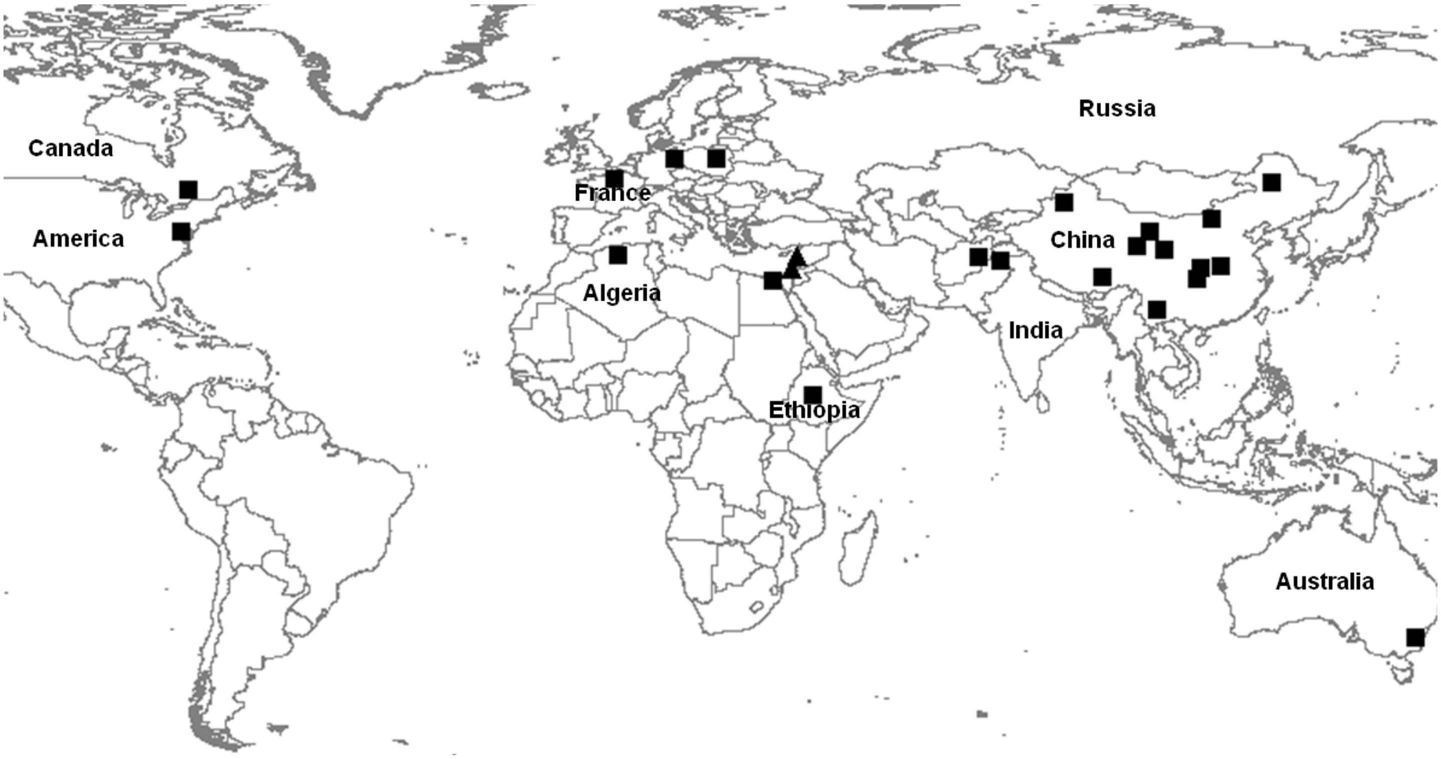

Fig 1. Geographic distribution of the 24 entries in the diverse panel for putative SSR validation (*Pisum sativum* entries are represented with squares and *Pisum fulvum* triangles).

doi:10.1371/journal.pone.0139775.g001

Table 1. Country of origin and geographic information of the 24 pea and its wild relatives used in this study.

	Voucher number	Species	Collection locale	Longitude	Latitude
1	G0000809	*Pisum sativum*	USA.	-164.5000	63.2500
2	G0000831	*Pisum sativum*	France	2.3330	48.8330
3	G0000868	*Pisum sativum*	Australia	149.1330	-35.2500
4	G0000871	*Pisum sativum*	Egypt	31.2330	30.0170
5	G0002101	*Pisum sativum*	Canada	-75.7000	45.4500
6	G0002108	*Pisum sativum*	Poland	21.0000	52.2170
7	G0002860	*Pisum sativum*	Germany	13.4170	52.5000
8	G0004353	*Pisum sativum*	Ethiopia	38.7000	9.0330
9	G0004484	*Pisum sativum*	Algeria	3.1330	36.7000
10	G0006082	*Pisum sativum*	Afghanistan	69.1830	34.4670
11	G0006170	*Pisum sativum*	Pakistan	73.1670	33.6670
12	G0000001	*Pisum sativum*	Henan, China	113.3200	32.7200
13	G0000043	*Pisum sativum*	Yunnan, China	101.6419	24.6889
14	G0000557	*Pisum sativum*	Xinjiang, China	84.8000	44.4000
15	G0000673	*Pisum sativum*	Qinghai, China	98.0872	36.2978
16	G0000783	*Pisum sativum*	Gansu, China	100.5070	38.9483
17	G0002288	*Pisum sativum*	Gansu, China	103.2105	35.6012
18	G0002305	*Pisum sativum*	Inner Mongolia, China	122.7375	48.0137
19	G0002371	*Pisum sativum*	Tibet, China	92.47	28.4200
20	G0002654	*Pisum sativum*	Inner Mongolia, China	111.6800	41.3700
21	G0003219	*Pisum sativum*	Hubei, China	108.9364	30.2910
22	G0003268	*Pisum sativum*	Hubei, China	109.7153	32.3183
23	G0005094	*Pisum fulvum*	Israel	35.2000	31.7830
24	G0005733	*Pisum fulvum*	Israel	34.4600	32.0500

doi:10.1371/journal.pone.0139775.t001

by the in-house developed script: reduce_ssr.py (data in S1 File), and the 'fine' primers were used for further study.

PCR amplification

Polymerase chain reactions (PCR) were performed in 10 μl reaction volumes containing 5 μl 2 x TaqPCR MasterMix (Hooseen, Beijing, China), 1 μl primer pair (10 μM), 1.5 μl of genomic DNA (30 ng) and 2.5 μl of dd H_2O. Microsatellites were amplified on a K960 Thermal Cycler (Jingle, Hangzhou, China) with the following cycle: 5 min initial denaturation at 95°C, 35 cycles of 30 s at 95°C, 30 s at the optimized annealing temperature, 45 s of elongation at 72°C, and a final extension at 72°C for 10 min. The PCR products were separated on 8% non-denaturing polyacrylamide gel electrophoresed at 280 V and 50 W and visualized by 0.1% silver nitrate staining.

Polymorphic validation and genetic diversity assessment

The number of alleles and polymorphism information content (PIC) of the alleles revealed by each primer pair were calculated by Powermarker v3.25 [20] with the genotype data among 24 accessions. A cluster analysis was conducted based on the unweighted pair group method on arithmetic averages (UPGMA) algorithm using Powermarker v3.25, and a dendrogram was drawn by Powermarker v3.25 [20] and modified by MEGA4 [21].

STRUCTURE V2.3.3 was used to analyze population structure and differentiation [22,23]. Simulations were run with a burn-in of 100,000 iterations and from K (the number of populations) = 1 to 10. Runs for each K were replicated 160 times and the true K was determined according to the method described by Evanno [24].

Linkage map construction and blast mapped SSR markers to *Medicago truncatula*

The distorted segregation of the markers against the expected Mendelian segregation ratio was tested with Chi-squared analysis ($P < 0.05$) by QTL ICIMapping V3.2 software [25]. The information of SSR markers were filled into Map Manager QTXb 20 software [26]. For the F_2 population, the male allele was recorded as "A" and the female allele as "B", "H" was recorded when a locus was heterozygous, and "-" when there was a missing or null allele. The linkage map was constructed using the Map Manager QTXb 20 software with the parameter of Kosambi function ($P < 0.0001$) and marker distances in centiMorgans (cM). Finally, the linkage map was presented by JoinMap 4.0 software [27]. Putative location of flanking sequences of mapped SSRs onto chromosomes of *Medicago truncatula* for synteny-based comparison was conducted by using blast method (http://phytozome.jgi.doe.gov/pz/portal.html#!search?show=BLAST&method=Org_Mtruncatula).

Results

Illumina paired-end sequencing

In this study, a total of 17.5 GB of paired-end raw sequencing data, comprising 173,245,234 reads from a 500 bp insert DNA library, was generated by Illumina Hiseq2500 system. After trimming the adaptors and removal of possible contaminations, the remaining 170,865,238 high quality read sequences were used for further analysis. Adenine was the most abundant type, accounting for 29.1% of total nucleotides, followed by thymine (28.9%), cytosine (21.0%) and guanine (21.0%). The CG content was about 42% and the average read length was 94.7 bp.

Duplicated reads removing and genome de novo assembling

The trimmed reads were used for duplicated sequences analysis under the Remove Duplicate Reads Module in CLC Genomics Workbench 7.5 software. As a result, there were 505,740 (0.3%) duplicate reads and 170,359,498 (99.7%) unique ones. After the de novo assembly, the number of contigs (including scaffolded regions) was 343,849. The average length of contigs and the N50 was about 370 bp and 359 bp, respectively.

Mining for SSRs

MISA software was used for SSRs search based on contigs. The total number of SSR containing sequences was 8,899, and these sequences contained 10,207 SSRs (Table 2). In this study, mono- and di- nucleotide motifs occurred at the highest rate (accounting for 40.86% and 32.68%, respectively). Trinucleotide motifs accounted for 25.29%, while tetra-, penta-, and hexa-nucleotide motifs accounted for 1.17%. $(A/T)_n$, $(AC/GT)_n$ and $(AG/CT)_n$ were the relatively more frequent motifs in our study.

Primer design

A total of 3,275 non-redundant primer pairs were designed by Primer 3.0 software and reduce_ssr.py (in house developed programs) based on criteria of melting temperature, CG content, lack of secondary structure and length of amplification bands. The expected length of target bands was between 110 bp and 210 bp.

Table 2. MISA result in this study.

Category	Numbers
Total number of sequences examined	343,849
Total size of examined sequences (bp)	127,283,564
Total number of identified SSRs	10,207
Number of SSR containing sequences	8,899
Number of sequences containing more than one SSR	671
Number of SSRs present in compound formation	450

doi:10.1371/journal.pone.0139775.t002

Validation of the SSR markers

A subset of 1,644 SSR markers was randomly selected for validation. Among them 841 (51.16%) markers (S2 File) produced reliable polymorphic bands between 22 pea accessions (*Pisum sativum*) and two wild relatives (*Pisum fulvum*). Meanwhile, the monomorphic markers were listed in S3 File. The allele number per locus ranged from 2 to 10 with an average of 3.22. The polymorphism information content (PIC) with an average of 0.38, ranged from 0.08 to 0.82 (S2 File). The dendrogram clearly showed that the 24 pea and its wild relative accessions fell into three distinct clusters based on 841 polymorphic SSR markers (Fig 2). Cluster I consisted of overseas accessions except G0002305; Cluster II consisted Chinese accessions; Cluster III consisted of wild relatives.

The population structure of this diverse panel of cultivated pea and its wild relative was inferred by using STRUCTURE V2.3.3 with the dataset of 841 SSR markers. Three sub-populations were identified, based on ΔK values (Fig 3, [24]). The rational for thisΔK is to make salient the break in slope of the distribution of *L*(K) at the true K. The entries from China, from other countries and the wild species were separated into 3 sub-populations (Fig 4), in good according with the three clusters in the UPGMA dendrogram. The results were in accordance with those published earlier [28].

Using novel SSR markers to enhance the density of genetic linkage map

A segregating F_2 population derived from the cross between G0003973 and G0005527 was used for mapping the newly validated SSR markers. Among the 1,644 SSRs used in genetic diversity analysis, 63 were polymorphic between the two parents. Being amplified in the population, 22 of the 63 SSRs showed significant segregation distortion ($P < 0.05$) in S4 File. These distorted markers were excluded from linkage map construction. The Map Manager QTXb 20 was used to add the newly developed SSR markers to the genetic linkage map which had been published [29]. Consequently, 41 polymorphic markers that segregated in appropriate Mendalian ratios were used to run Map Manager QTXb 20 software, of which 33 markers were mapped to the existing linkage groups. However, the remaining eight markers were not linked to any mapped markers on the linkage map. The new map contained 199 markers including the 33 newly added markers (Table 3) in 13 linkage groups with an average genetic distance of 9.5 cM between neighboring markers and covered 1890.88 cM (Fig 5).

Discussion

SSR markers are excellent genetic markers because they are co-dominant, multi-allelic and reproducible. In genetics, SSRs have been widely used for diversity analysis [30], linkage map construction [31], QTL mapping [32] and association mapping [33].

Pea is important in genetics, because of the work of J.G. Mendel [34]. However, the pea genome is very large, which seriously hindered pea genomic research. The nuclear genome size

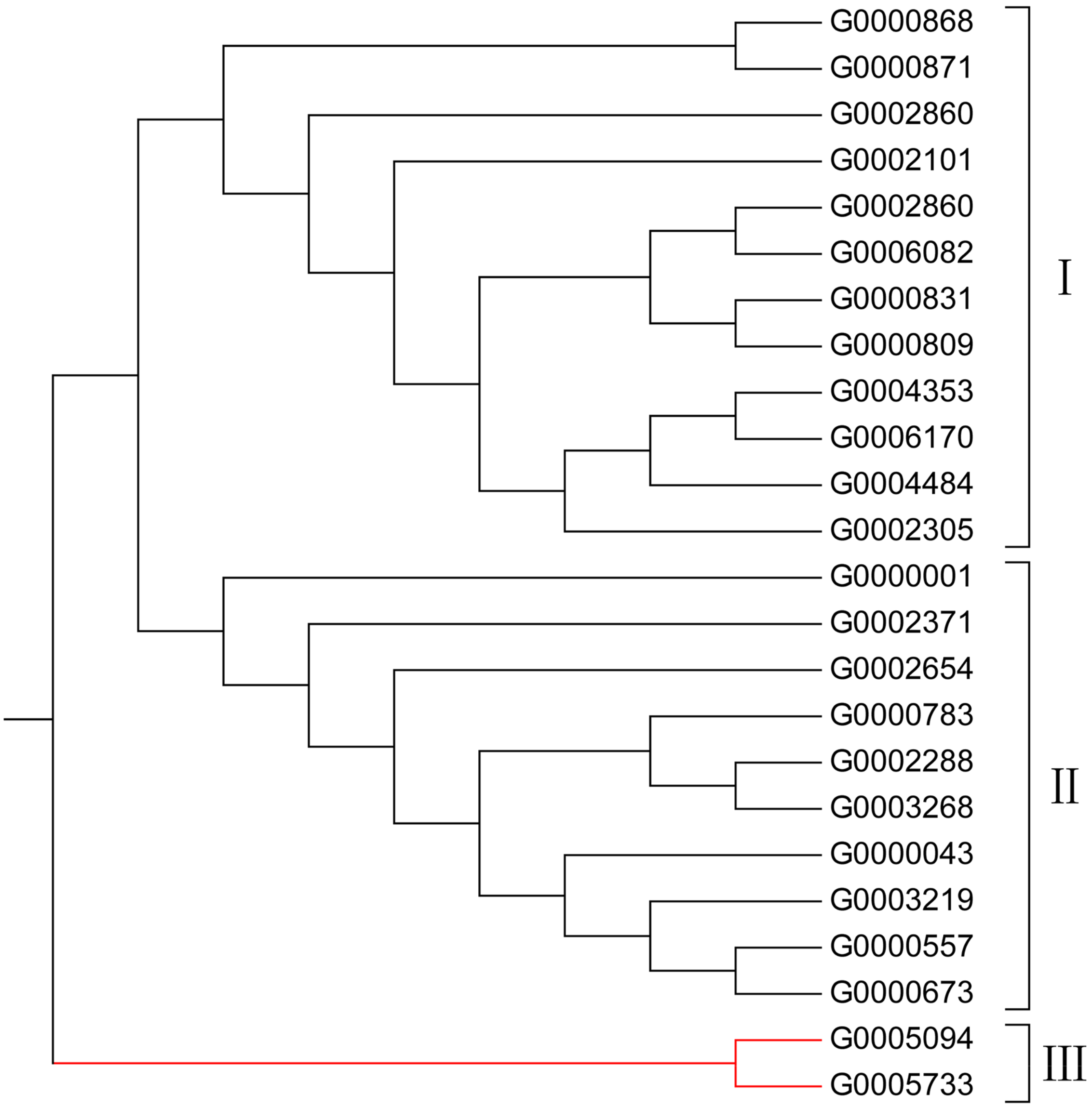

Fig 2. A UPGMA dendrogram of 24 accessions of pea and its wild species based on 841 polymorphic SSR markers amplified (Cluster I consisted of overseas accessions except G0002305; Cluster II consisted Chinese accessions; Cluster III consisted of wild relatives).

doi:10.1371/journal.pone.0139775.g002

of pea was estimated to be 9.09 pg DNA/2C, which corresponds to a haploid genome size (1C) of 4.45 Gbp [35], one and half times larger than the human genome of 3Gb [36]. Compared with other legume crops such as soybean (*Glycine max*) of 1.1 Gb [7] and barrel medic (*Medicago truncatula*) of 0.47 Gb (http://www.jcvi.org/medicago/), More efforts are needed to develop molecular tools especially for SSR and SNP (single nucleotide polymorphism) markers in order to build a solid foundation for its genomic research in peas.

Using NGS technology for the identification of SSR markers is effective

Consistent with previous reports [37–39], results from this study demonstrated that Illumina paired-end sequencing offers an opportunity for high-throughput identification of SSRs with diverse motifs from economically important crop plant species. Within a relatively short time

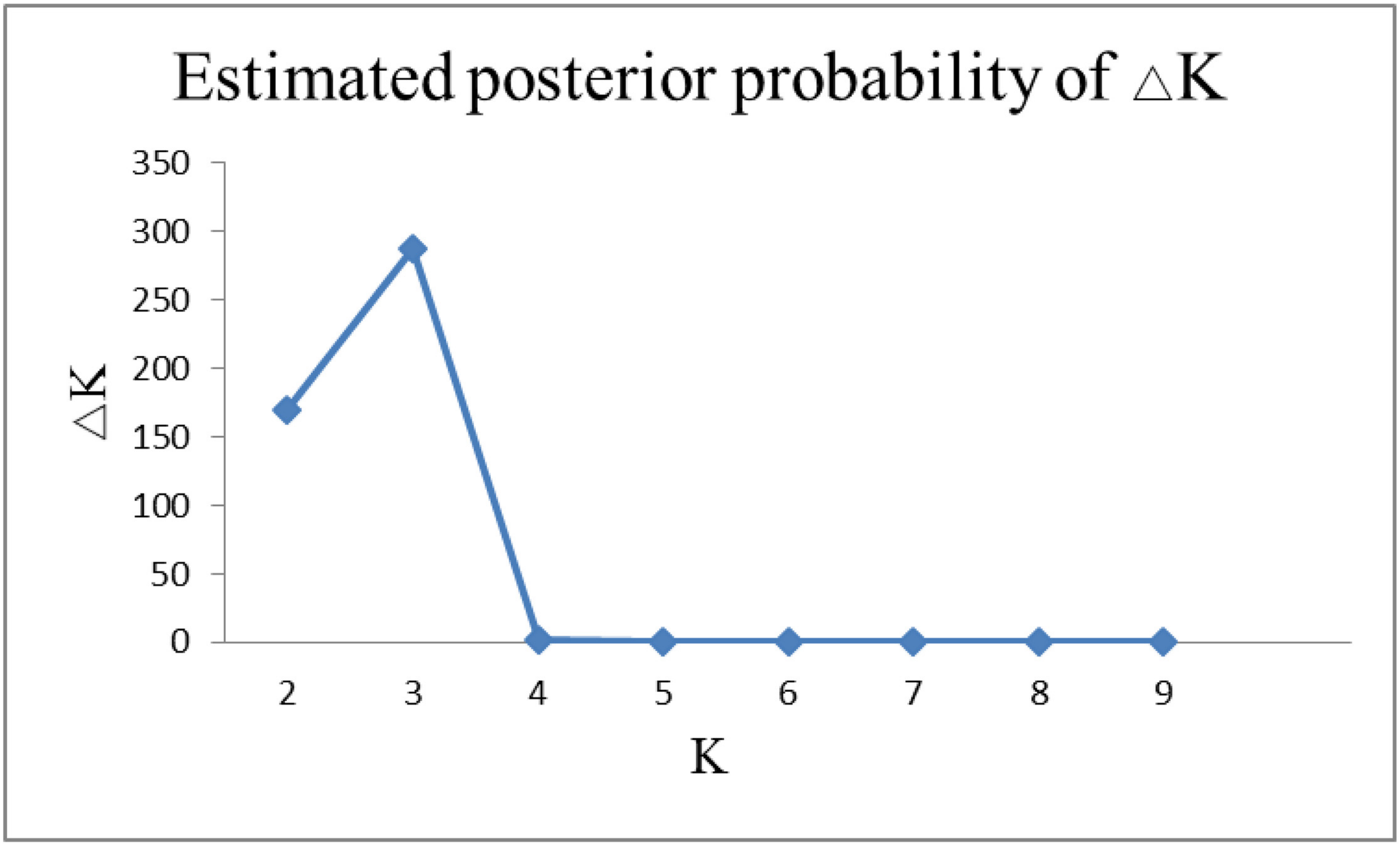

Fig 3. ΔK was used to determine the most appropriate K value for population structure.

doi:10.1371/journal.pone.0139775.g003

period, our sequencing experiment generated a total of 17.5 GB of raw paired-end sequencing data. From this raw data, 343,849 contigs were effectively assembled and used for SSR markers development. A total of 3,275 non-redundant primers were designed and nearly half of them

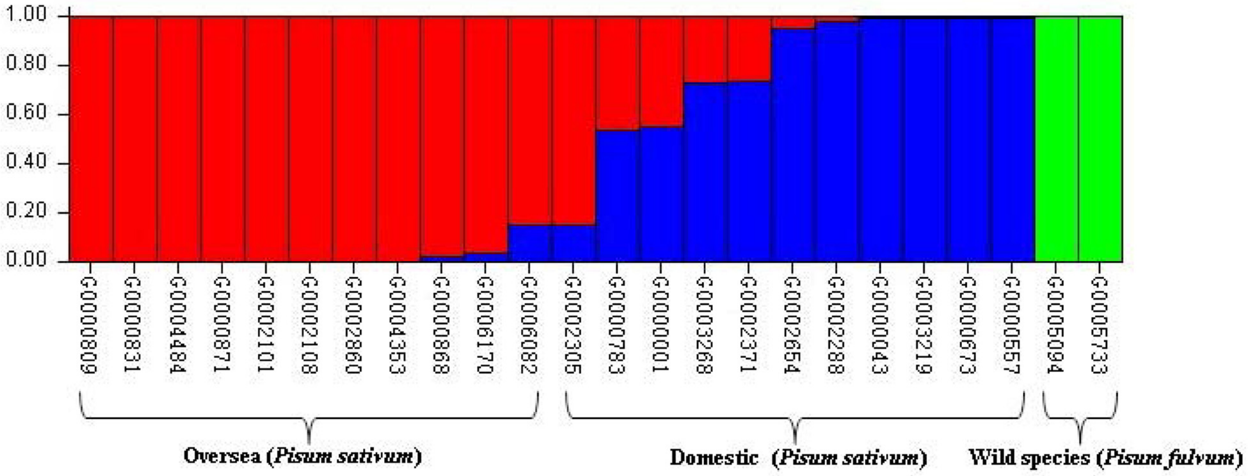

Fig 4. Population structure of K = 3 inferred by Bayesian clustering approaches based on SSR markers.

doi:10.1371/journal.pone.0139775.g004

Table 3. Primer sequence and linkage group information of the 33 mapped SSR markers.

No.	Marker Name	Primer sequences(5'-3')	Linkage Group	Map Positions (cM)	*Medicago truncatula* Chromosome	E-Value
1	18339	F:TGGTTGAACTGGAACGAGTG R: TGAAATTGCAATGTAAGCATGA	LG I	32.67	Chr 5	8.9E-20
2	16881	F:ATGGGCTTTAGGGGAAGAAA R: AAAAGCAGCACATGGAGGAC	LG I	77.78	Chr 5	7.1E-02
3	18011	F:GACCAACGACTTGGACATCA R: GGTGAGTTCCTAAGATGAATCAGA	LG I	115.80	Not found	Not found
4	18591	F:AGGGCCGAATGCTAAGTGAT R: TTTTGAACCCTGGAGGGAGT	LG I-1	8.44	Chr 5	1.3E-18
5	16549	F:CAATGAGATGCTGGCGATAA R: GTTCGGTGTTGTGGGTTTTT	LG I-1	228.11	Chr 4	9.8E-07
6	18391	F:CCATCCTCCACGTGTCTCTT R: TCGCATATCCAAATGCAAAC	LG I-1	303.15	Chr 7	3.3E-19
7	16758	F:CCCTTCAACAAAGCCTAACG R: AGGGTGCGAAGGAGGTTAGT	LG II	205.92	Chr 6	2.7E-06
8	18438	F:GATTGAGCCGTGCCAATATC R: GATCCCACCCTAGAGGAAAAA	LG II	240.35	Chr 4	5.7E-10
9	18363	F:CATGCATGGAGTTGGAAGAG R: GTCCCAAAATGCAGCCAATA	LG III	4.22	Chr 2	9.1E-20
10	16524	F:CCAGAGGATGTGAACCAGGTA R: TTCAACCAAGCTGAACCCTTA	LG III	37.88	Chr 6	2.1E-02
11	18135	F:CTTCAACCAACTGCGAGTGA R: TCATTTGAGTTTTGCCATGTTC	LG III	75.26	Chr 3	2.2E-20
12	17066	F:TGGGATGAAAATGTTATGAATG R: CAAAACCACCCTTTCCGATA	LG VII	0.00	Chr 8	3.2E-20
13	18013	F:TCAATTCCGAACCACCTTTC R: CGGCAGAATTAGGGTTTTGA	LG VII	4.60	Chr 8	1.8E-05
14	17431	F:TTCACAATTCACCACCAATCA R: CCAACGTCAGGTACGATTCA	LG VII	39.48	Chr 1	9.5E-03
15	19075	F:CACGAGTACAACATGGAGTGAAG R: CAAGCTCAACCTCCTCATACC	LG A	0.00	Chr 6	3.1E-02
16	19252	F:CAATATTGATCGGAATTTGTTTC R: TGCGGTTTGATTGAGTTTGA	LG A	16.34	Chr 7	6.8E-11
17	16534	F:TTGCAAATATACCAATTCCAAAA R: ATTGGAGCCTGGTGAAGACC	LG A	30.23	Not found	Not found
18	18049	F:ACCCCTCTTTGCTAGGGTGA R: ACCACACATCTCGCACACAT	LG B	65.61	Chr 1	2.3E-04
19	18358	F:CCTGAACCGATTTTGGTGAT R: ATTCCGCCCTCTTTCACTTC	LG B	100.89	Chr 4	1.7E-03
20	16512	F:TAAGCCCGACGCTTCTATTC R: GTGCCTCAGTTTCCGTTTGT	LG B	153.32	Chr 6	1.4E-04
21	18533	F:TCCAAAATGCGTGTCATCAT R: TGACCGACACATTCATCTTCA	LG B	177.69	Chr 4	4.3E-18
22	17754	F:AGCAACGGGCAACCTTATAG R: CCTTTTGTTTGGAAGCTCAA	LG C	0.00	Chr 2	9.5E-02
23	17531	F:TGCAGGGGTGTGTGTTACAT R: TGAACATGGTGAAATGGATTG	LG C	15.31	Not found	Not found
24	18237	F:GGGATATGAGAAGGCGATACC R: TGGTTGTAGGATGTGGGATTT	LG C	25.77	Chr 3	2.5E-07
25	18272	F:CCCCAACATTTCTCTAGGTAACA R: TTCTTCGCAGCTCGGTAAGT	LG C	57.91	Chr 1	4.7E-04
26	17713	F:AAAAAGGGGAAAGCAGGAGA R: TTGACTGTGAGGCTGGTTTG	LG C	157.15	Chr 2	4.1E-06

(*Continued*)

Table 3. (*Continued*)

No.	Marker Name	Primer sequences(5'-3')	Linkage Group	Map Positions (cM)	*Medicago truncatula* Chromosome	E-Value
27	16570	F:CAAACACCAACCACCACAGT R: AAGGGGAGACGAAGTGGAGT	LG C	194.92	Chr 6	1.8E-03
28	16588	F:CGGTCTGAGGTTGTTGTGAA R: TTGTAAGACCGACTCGTCCA	LG E	0.00	Chr 2	1.6E-29
29	18928	F:TGAATGTGGAAAGGAGGAATG R: AGGGTCACCACTTTGGAGAG	LG G	14.69	Chr 5	4.6E-06
30	16452	F:CGATGGTTGCTGTTGTGAGA R: ACCCCAAACAAACACCAATG	LG G	29.64	Chr 5	5.6E-03
31	17605	F:CGCCCTTCATCATCATCTTC R: AGAGTCGGTCCCTCCAACAT	LG G	46.14	Chr 8	5.5E-04
32	17628	F:GGTTTTGTTTGCCGTTGATT R: CCACCCCCAAACTTCCTTAT	LG G	56.22	Chr 5	7.4E-28
33	18323	F:CAGACAATGGCAATTATTTGGTAA R: CTGCTGTTGCTTCGATTTCA	LG H	51.84	Chr 3	1.5E-10

doi:10.1371/journal.pone.0139775.t003

(1,644 primers) were validated in two different ways, a diversity panel of 24 accessions and a segregating F2 population. For SSR markers development, NGS research strategy is very efficient.

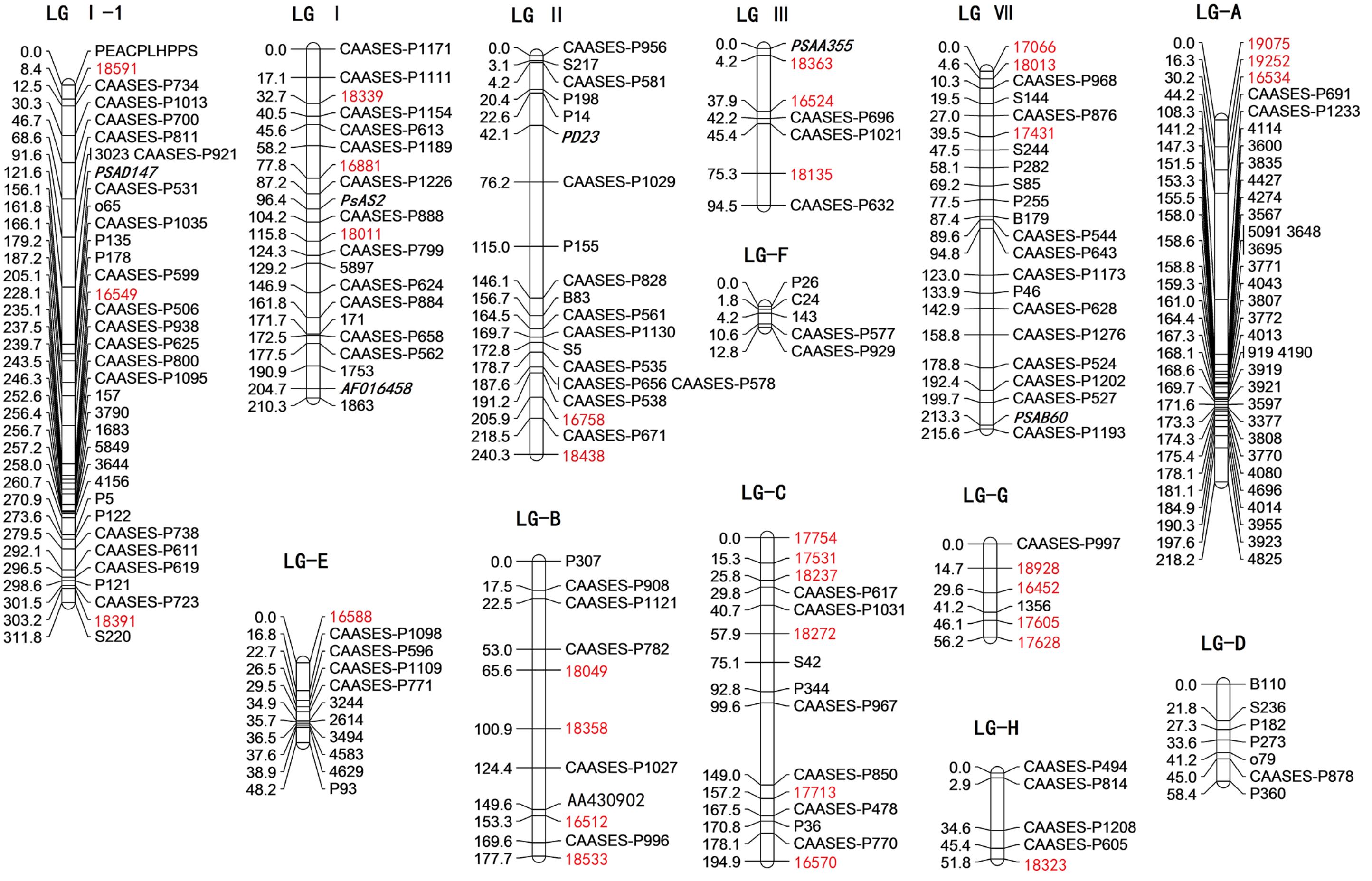

Fig 5. The enhanced genetic linkage map of the G0003973 × G0005527 F_2 population with newly developed SSR markers (red).

doi:10.1371/journal.pone.0139775.g005

More reliable validation of the NGS based SSR markers was conducted

In the published studies of other plant species [40–41], only a small proportion of newly developed SSR markers was tested. In this study, more than half (1,644 markers) was carefully tested in two different ways. One way was genetic diversity analysis, the other way was the mapping of the novel markers to a linkage map based on an existing mapping population. More than 51% tested SSRs involved in this study were polymorphic among 24 accessions and clearly divided into 3 sub-groups. Meanwhile, 33 novel SSRs were anchored onto a previous genetic linkage group.

Chinese pea germplasm differs from that of other countries

The comparison of the diversity of Chinese and foreign peas by using 841 polymorphic SSR markers in our study identified a significant degree of diversity (Figs 2 and 4). This result coincided with a previous study by using 21 informative SSRs to assess and compare the genetic diversity of 1,243 Chinese pea genotypes from 28 provinces to 774 pea genotypes that represented a globally diverse germplasm collection, and the Chinese pea germplasm was found genetically distinct from the global gene pool sourced outside China [28].

On the other hand, our genotype data did reveal an exception. In our experiment, G0002305 is an accession collected from Inner Mongolia. The cluster analysis grouped this accession into Cluster I with germplasm accessions collected outside China (Figs 2 and 4). Analysis of population structure also confirmed that this Chinese accession shared more than 80% of kinship with accessions collected from outside China, especially with G0006082 from Afghanistan and G0006170 from Pakistan (Fig 4). In addition, six Chinese accessions shared variable percentages (approximately 5 to 50%) of closeness with accessions collected outside China and two accessions collected outside China share a small percentage of closeness with the Chinese accessions (Fig 4). Both cluster and population structure analyses clearly separated the cultivated pea from its wild relative accessions (Figs 2 and 4). These results implies the usefulness of the newly developed SSRs.

More SSR markers were anchored on a genetic linkage map

There was no genetic linkage map based on Chinese accessions previously. In 2014, we constructed the first Chinese pea linkage map constructed with 157 SSR markers [29]. In this study, the existing linkage map has been more saturated. The new map contained 199 markers including the 33 newly added markers. We anticipate that with more effective SSR markers, QTL mapping and association study as well as marker-assisted selection in pea will become available in the near future.

Supporting Information

S1 File. The code of reduce_ssr.py.
(TXT)

S2 File. Polymorphic SSR markers of *Pisum sativum* L.
(DOCX)

S3 File. Monomorphic SSR markers of *Pisum sativum* L.
(DOCX)

S4 File. Twenty two distorted segregation markers.
(DOCX)

Acknowledgments

We acknowledge F. Dugan for critical reading of the manuscript. This work was supported by the National Natural Science Foundation of China (No. 31371695), China Agriculture Research System (No. CARS–09), Social Development of Science and Technology Plan (No. 2013BB010) from Yunnan Government, and also granted by the Agricultural Science and Technology Innovation Program (ASTIP) in CAAS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Author Contributions

Conceived and designed the experiments: XZ. Performed the experiments: LF ST FW. Analyzed the data: TY SB LL YH JJ. Contributed reagents/materials/analysis tools: LF XZ JH. Wrote the paper: TY JH.

References

1. FAO (2013) Statistical Database, Food and Agriculture Organization (FAO) of the United Nations.
2. Phillips DA (1980) Efficiency of Symbiotic Nitrogen Fixation in Legumes. Annual Review of Plant Physiology 31: 29–49.
3. Loridon K, McPhee K, Morin J, Dubreuil P, Pilet-Nayel M, Aubert G, et al. (2005) Microsatellite marker polymorphism and mapping in pea (*Pisum sativum* L.). Theoretical and Applied Genetics 111: 1022–1031. PMID: 16133320
4. Burstin J, Deniot G, Potier J, Weinachter C, Aubert G, Barranger A (2001) Microsatellite polymorphism in *Pisum sativum*. Plant Breeding 120: 311–317.
5. Xu S-C, Gong Y-M, Mao W-H, Hu Q-Z, Zhang G-W, Fu W, et al. (2012) Development and characterization of 41 novel EST-SSR markers for *Pisum sativum* (Leguminosae). American Journal of Botany 99: e149–e153. doi: 10.3732/ajb.1100445 PMID: 22447986
6. Kaur S, Pembleton LW, Cogan NOI, Savin KW, Leonforte T, Paull J, et al. (2012) Transcriptome sequencing of field pea and faba bean for discovery and validation of SSR genetic markers. BMC Genomics 13:104. doi: 10.1186/1471-2164-13-104 PMID: 22433453
7. Schmutz J, Cannon SB, Schlueter J, Ma JX, Mitros T, Nelson W, et al. (2010) Genome sequence of the palaeopolyploid soybean. Nature 465: 120–120.
8. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10: 57–63. doi: 10.1038/nrg2484 PMID: 19015660
9. Schuster SC (2008) Next-generation sequencing transforms today's biology. Nat Meth 5: 16–18.
10. Huang XH, Wei XH, Sang T, Zhao QA, Feng Q, Zhao Y, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. Nature Genetics 42: 961–976. doi: 10.1038/ng.695 PMID: 20972439
11. Yang T, Bao S-y, Ford R, Jia T-j, Guan J-p, He Y-h, et al. (2012) High-throughput novel microsatellite marker of faba bean via next generation sequencing. BMC Genomics 13: 602. doi: 10.1186/1471-2164-13-602 PMID: 23137291
12. Tangphatsornruang S, Somta P, Uthaipaisanwong P, Chanprasert J, Sangsrakru D, Seehalak W, et al. (2009) Characterization of microsatellites and gene contents from genome shotgun sequences of mungbean (*Vigna radiata* (L.) Wilczek). BMC Plant Biology 9: 137. doi: 10.1186/1471-2229-9-137 PMID: 19930676
13. Csencsics D, Brodbeck S, Holderegger R (2010) Cost-Effective, Species-Specific Microsatellite Development for the Endangered Dwarf Bulrush (*Typha minima*) Using Next-Generation Sequencing Technology. Journal of Heredity 101: 789–793. doi: 10.1093/jhered/esq069 PMID: 20562212
14. Zhu H, Senalik D, McCown BH, Zeldin EL, Speers J, Hyman J, et al. (2012) Mining and validation of pyrosequenced simple sequence repeats (SSRs) from American cranberry (*Vaccinium macrocarpon* Ait.). Theor Appl Genet 124: 87–96. doi: 10.1007/s00122-011-1689-2 PMID: 21904845
15. Malausa T, Gilles A, MeglÉCz E, Blanquart H, Duthoy S, Costedoat C, et al. (2011) High-throughput microsatellite isolation through 454 GS-FLX Titanium pyrosequencing of enriched DNA libraries. Molecular Ecology Resources 11: 638–644. doi: 10.1111/j.1755-0998.2011.02992.x PMID: 21676194
16. Yang T, Jiang JY, Burlyaeva M, Hu JG, Coyne CJ, Kumar S, et al. (2014) Large-scale microsatellite development in grasspea (*Lathyrus sativus* L.), an orphan legume of the arid areas. BMC Plant Biology 14: 65. doi: 10.1186/1471-2229-14-65 PMID: 24635905

17. Dellaporta S, Wood J, Hicks J (1983) A plant DNA minipreparation: Version II. Plant Molecular Biology Reporter 1: 19–21.

18. Doyle J, Doyle J (1990) A rapid total DNA preparation procedure for fresh plant tissue. Focus 12: 13–15.

19. RCoreTeam (2012) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria ISBN 3-900051-07-0, URL http://wwwR-projectorg/.

20. Liu KJ, Muse SV (2005) PowerMarker: an integrated analysis environment for genetic marker analysis. Bioinformatics 21: 2128–2129. PMID: 15705655

21. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. Molecular Biology and Evolution 24: 1596–1599. PMID: 17488738

22. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945–959. PMID: 10835412

23. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: Linked loci and correlated allele frequencies. Genetics 164: 1567–1587. PMID: 12930761

24. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software STRUCTURE: a simulation study. Molecular Ecology 14: 2611–2620. PMID: 15969739

25. Li HH, Ye GY, Wang JK (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175: 361–374. PMID: 17110476

26. Manly KF, Cudmore RH, Meer JM (2001) Map manager QTX, cross-platform software for genetic mapping. Mammalian Genome 12: 930–932. PMID: 11707780

27. Van Ooijen JW (2006) JoinMap 4.0, software for the calculation of genetic linkage maps in experimental populations.

28. Zong XX, Redden R, Liu QC, Wang SM, Guan JP, Liu J, et al. (2009) Analysis of a diverse global *Pisum* sp. collection and comparison to a Chinese local *P. sativum* collection with microsatellite markers, Theoretical and Applied Genetics 118: 193–204. doi: 10.1007/s00122-008-0887-z PMID: 18815768

29. Sun X, Yang T, Hao J, Zhang X, Ford R, Jiang J, et al. (2014) SSR genetic linkage map construction of pea (*Pisum sativum* L.) based on Chinese native varieties. The Crop Journal 2: 170–174.

30. Ni JJ, Colowit PM and Mackill DJ. (2002) Evaluation of genetic diversity in rice subspecies using microsatellite markers. Crop Science 42: 601–607.

31. Hwang TY, Sayama T, Takahashi M, Takada Y, Nakamoto Y, Funatsuki H, et al. (2009) High-density integrated linkage map based on SSR markers in soybean. DNA Res. 16: 213–225. doi: 10.1093/dnares/dsp010 PMID: 19531560

32. Prasad M, Kumar N, Kulwal PL, Röder MS, Balyan HS, Dhaliwal HS, et al. (2003) QTL analysis for grain protein content using SSR markers and validation studies using NILs in bread wheat. Theor Appl Genet. 106: 659–667. PMID: 12595995

33. Zhang QQ, Wu CL, Ren FL, Li Y, Zhang CQ (2012) Association analysisof important agronomical traits of maize inbred lines with SSRs. Australian Journal of Crop Science. 6: 1131–1138.

34. Mendel GJ (1866) Versuche über Pflanzen-Hybriden. Verhandlungen der naturfoschung Vereins, Abhandlungern, Brünn 4: 3–47.

35. Dolezel J, Greilhuber J (2010) Nuclear Genome Size: Are We Getting Closer? Cytometry Part A 77A: 635–642.

36. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. (2001) The Sequence of the Human Genome. Science. 291 (5507): 1304–1351. PMID: 11181995

37. Wang ZY, Fang BP, Chen JY, Zhang XJ, Luo ZX, Huang LF, et al. (2010) De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). BMC Genomics 11: 726. doi: 10.1186/1471-2164-11-726 PMID: 21182800

38. Castoe TA, Poole AW, de Koning APJ, Jones KL, Tomback DF, Oyler-McCance SJ, et al. (2012) Rapid Microsatellite Identification from Illumina Paired-End Genomic Sequencing in Two Birds and a Snake. PLoS ONE 7(2): e30953 doi: 10.1371/journal.pone.0030953 PMID: 22348032

39. Duan CX, Li DD, Sun SL, Wang XM, Zhu ZD (2014) Rapid Development of Microsatellite Markers for *Callosobruchus chinensis* Using Illumina Paired-End Sequencing. PLoS ONE 9(5): e95458. doi: 10.1371/journal.pone.0095458 PMID: 24835431

40. Li DJ, Deng Z, Qin B, Liu XH, Men ZH (2012) De novo assembly and characterization of bark transcriptome using Illumina sequencing and development of EST-SSR markers in rubber tree (*Hevea brasiliensis* Muell. Arg.). 13: 192.

41. Du F, Wu Y, Zhang L, Li X-W, Zhao X-Y, Wang W-H, et al. (2015) De Novo Assembled Transcriptome Analysis and SSR Marker Development of a Mixture of Six Tissues from Lilium Oriental Hybrid 'Sorbonne'. Plant Molecular Biology Reporter 33: 281–293.