Theor Appl Genet (2008) 117:857–871
DOI 10.1007/s00122-008-0825-0

ORIGINAL PAPER

Genetic structure and diversity of cultivated soybean (*Glycine max* (L.) Merr.) landraces in China

Yinghui Li · Rongxia Guan · Zhangxiong Liu · Yansong Ma · Lixia Wang · Linhai Li · Fanyun Lin · Weijiang Luan · Pengyin Chen · Zhe Yan · Yuan Guan · Li Zhu · Xuecheng Ning · Marinus. J. M. Smulders · Wei Li · Rihua Piao · Yanhua Cui · Zhongmei Yu · Min Guan · Ruzhen Chang · Anfu Hou · Ainong Shi · Bo Zhang · Shenlong Zhu · Lijuan Qiu

Received: 19 September 2007 / Accepted: 8 June 2008 / Published online: 28 June 2008

Abstract The Chinese genebank contains 23,587 soybean landraces collected from 29 provinces. In this study, a representative collection of 1,863 landraces were assessed for genetic diversity and genetic differentiation in order to provide useful information for effective management and utilization. A total of 1,160 SSR alleles at 59 SSR loci were detected including 97 unique and 485 low-frequency alleles, which indicated great richness and uniqueness of genetic variation in this core collection. Seven clusters were inferred by STRUCTURE analysis, which is in good agreement with a neighbor-joining tree. The cluster subdivision was also supported by highly significant pairwise F_{st} values and was generally in accordance with differences in planting area and sowing season. The cluster HSuM, which contains accessions collected from the region between 32.0 and 40.5°N, 105.4 and 122.2°E along the central and downstream parts of the Yellow River, was the most genetically diverse of the seven clusters. This provides the first molecular evidence for the hypotheses that the origin of cultivated soybean is the Yellow River region. A high proportion (95.1%) of pairs of alleles from different loci was in LD in the complete dataset. This was mostly due to overall population structure, since the number of locus pairs in LD was reduced sharply within each of the clusters compared to the complete dataset. This shows that population structure needs to be accounted for in association studies conducted within this collection. The low value of LD within the clusters can be seen as evidence that much of the recombination events in the past have been maintained in soybean, fixed in homozygous self-fertilizing landraces.

Communicated by A. Charcosset.

Electronic supplementary material The online version of this article (doi:10.1007/s00122-008-0825-0) contains supplementary material, which is available to authorized users.

Y. Li · R. Guan · Z. Liu · Y. Ma · L. Wang · L. Li · F. Lin ·
W. Luan · Z. Yan · Y. Guan · L. Zhu · X. Ning · W. Li · R. Piao ·
Y. Cui · Z. Yu · M. Guan · R. Chang · L. Qiu (✉)
The National Key Facility for Crop Gene Resources
and Genetic Improvement (NFCRI)/Key Lab of Germplasm
and Biotechnology (MOA), Institute of Crop Science,
Chinese Academy of Agricultural Sciences,
100081 Beijing, People's Republic of China
e-mail: qiu_lijuan@263.net

P. Chen · A. Hou · A. Shi · B. Zhang · S. Zhu
Department of Crop, Soil, and Environmental Sciences,
University of Arkansas, Fayetteville, AR 72701, USA

Marinus. J. M. Smulders
Plant Research International, Wageningen UR,
P.O. Box 16, 6700 AA Wageningen, The Netherlands

Introduction

It is generally accepted that soybean (*Glycine max* (L.) Merr.) originated in China (Fukuda 1933; Vavilov 1951;

Y. Ma
Soybean Research Institute of Heilongjiang Academy
of Agricultural Sciences, 150086 Harbin,
People's Republic of China

L. Li · F. Lin · W. Luan
Northwest A & F University, 712100 YangLing,
People's Republic of China

Z. Yan · Y. Guan · L. Zhu · X. Ning
Xinjiang Agricultural University, 830046 Urumqi,
People's Republic of China

W. Li · R. Piao · Y. Cui · Z. Yu · M. Guan
Soybean Research Institute, Key Laboratory of Soybean Biology,
Ministry of Education, Northeast Agricultural University,
150030 Harbin, People's Republic of China

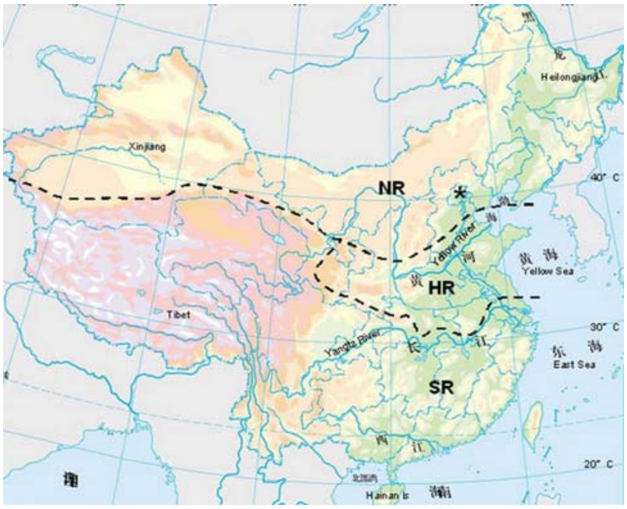

Fig. 1 Soybean landraces are divided into three large eco-regions, namely Northern eco-region (NR), Huang-Huai eco-region (HR), and Southern eco-region (SR) in order to establish primary core collection as sketched on map of China. The eco-geographic regions are roughly separated by the Yellow River and the Yangtse River

Hymowitz and Newell 1981). Soybean cultivation in China can be traced back over 3,000 years, as evidenced by the ancient books and archaeological discoveries (Morse 1950; Zhao and Gai 2004). To date, soybean is grown in China as one of the most important economic crops for its high oil and protein concentrations and is planted widely in China from Tahe county, Heilongjiang province (53°N) in the north to Ya county, Hainandao province (18°N) in the south. According to topographic distribution, soybean production in China is concentrated in three main areas: Northern region, Huang-Huai (Yellow River-Huai River) region, and Southern region (Fig. 1).

Due to differences in climate and soils each main production region can be further divided into smaller subregions (Wang 1991; Pan et al. 1994; Wang and Gai 2002). Each region has different growing conditions and locally adapted cultivars (Lu et al. 1981; Bu and Pan 1987). Soybeans can be planted during spring in all three regions, during summer in Huang-Huai and Southern regions and during fall only in Southern region.

Soybean landraces are adapted to various environments, from cold to hot, humid to drought, and fertile plains to elevated mountains. Some soybean landraces are still being used as local cultivars in southern provinces such as Guangdong, Guangxi, Yunnan, and Guizhou, although the vast majority of the production areas are planted with modern commercial varieties. These landraces have emerged as valuable germplasm for breeding modern soybean cultivars. For instance, more than 20 cultivars released in China have the superior offspring from two local soybean landrace Fengdihuang and Xiongyaoxiaolihuang, in their pedigrees (Chang et al. 1998). The Chinese landraces have also been introduced into other major soybean-producing countries including the United States, Brazil, and Japan (Chang et al. 1998). Disease resistant resources such as Peking and PI 88788 for soybean mosaic virus and soybean cyst nematode (*Heterodera glycines* Ichinohe) resistance have been used in the US for the development of new soybean cultivars (Pantalone et al. 2003; Anand et al. 2004; Wang et al. 2005).

Collection, evaluation, conservation, and utilization of soybean landraces have become one of the top priorities for agricultural crop research in China. The Chinese national collection of soybean landraces is deposited at the National Gene Bank at the Institute of Crop Science, Chinese Academy of Agricultural Sciences. With the collaboration of provincial germplasm institutions throughout China, more than 23,000 soybean germplasm lines have been collected and preserved in the national gene bank. This large number has presented challenges and demands for more efficient management and cost-effective conservation. Tremendous efforts have been taken in characterizing these soybean germplasm, mainly for phenotypic traits. In order to further identify and use the genetic diversity in these collections, we have constructed a primary core collection representing 70–80% of genetic variation present in the original entire collection (Qiu et al. 2003).

Genetic diversity analysis reveals genetic backgrounds and relationships of germplasm, and also provides strategies to establish, unitize, and manage crop core collections (Brown-Guedira et al. 2000; Roussel et al. 2004). Since 1981, soybean researchers have studied genetic diversity of cultivars, landraces and annual wild soybeans (Lu et al. 1981; Kisha et al. 1998; Cui et al. 2000; Li and Nelson 2001; Xu et al. 2002; Cui et al. 2003; Lin et al. 2003; Ude et al. 2003; Xie et al. 2003; Zhu et al. 2003; Cui et al. 2004a; 2004b; Dong et al. 2004; Mi et al. 2004; Wang et al. 2004; Chen and Nelson 2005; Li et al. 2005; Luan et al. 2005; Piao et al. 2005; Xie et al. 2005; Zhang et al. 2005; Wang et al. 2006; Guan et al. 2006). However, all of those studies compared the diversity among pre-defined populations based on tillage systems, geographical origins, and phenotypes. The traditional estimators of population structure will reduce the genetic information of germplasm if pre-defined populations do not accurately match the biological reality (Pearse and Crandall 2004). This is a serious problem for correct evaluation the genetic structure and diversity, especially in a diverse crop cultivated across a large geographical area. Take for example the single cropping system, which is classified by tillage system and sowing time or weather factors such as temperature, water, and photoperiod in China (Bu and Pan 1987; Pan et al. 1994; Wang and Gai 2002). Lu et al. (1981) and Bu and Pan (1987) considered the entire Southern China as one cropping region because multiple planting soybeans exist in this region, while Pan et al. (1994) divided cultivars from Southern China into three subgroups. Wang and Gai (2002) divided them into four ecotypes, and added a new ecotype for the cultivars from YunGui plateau region. The characterization of the correct population structure within germplasm collections is critical to identify and correctly interpret the associations between functional and molecular diversity (Pritchard and Rosenberg 1999; Buckler and Thornsberry 2002).

Bayesian or maximum-likelihood model-based clustering methods enable defining populations and assign individuals to these inferred populations based on genetic data alone (Pritchard et al. 2000). Currently, STRUCTURE is one of the most widely used genotypic clustering programs (Pritchard et al. 2000). The Bayesian method in STRUCTURE does not use sampling locations, hypothesized genetic origins of cultivars, or phenotypic information, but uses the number of genetic clusters specified before the algorithm is applied (Rosenberg et al. 2002). Typically, the program analyzes the data starting with various numbers of predefined clusters, after which the best fitting model is selected based on extent of genetic information explained and on the match with the known characteristics of the biological samples. STRUCTURE has been successfully applied on the genetic structure analyses of, for example, human (Rosenberg et al. 2002), dog (Parker et al. 2004), and maize (Remington et al. 2001; Liu et al. 2003). The program can handle large numbers of simple sequence repeat (SSR) or microsatellite markers, which are widely used because they are highly polymorphic, abundantly available and randomly distributed in the genome, and co-dominantly inherited (McGregor et al. 2000; Narvel et al. 2000).

The application of association mapping to detect quantitative trait loci (QTLs) for functional genes appears to be a promising approach to overcome the limitations of conventional linkage mapping and has received increased attention of plant geneticists during the last few years (Kraakman et al. 2004; Gupta et al. 2005). Information on the extent and structure of linkage disequilibrium (LD) within the population under consideration is a prerequisite for association mapping. A panel of a large number of SSR markers is suitable to establish LD patterns, as was shown in maize (Remington et al. 2001; Liu et al. 2003; Stich et al. 2005) and durum wheat (Maccaferri et al. 2005). Gametic phase disequilibrium can be the result of a number of factors next to recombination, and the population structure resulting from selection, population admixture, and genetic drift is more likely to remain intact over time in a selfing species such as soybean, in which recombination is very low or absent. Zhu et al. (2003) detected low genome-wide LD in 25 diverse soybean genotypes basing on sequencing of a few regions (including 66 complete GenBank genes, 50 cDNAs, 13 BAC subclones and 15 SSR flanking regions). Recently, Hyten et al. (2006) analyzed the structure of LD in three regions of the genome varying in length from 336 to 574 kb and found that LD extended from 90 to 574 kb in the three cultivated *G. max* groups, which included Asian landraces, North American Ancestor and Elite cultivars.

The objectives of this study were the following: (1) to clarify the population structure of Chinese soybean landraces; (2) to analyze the genetic diversity and genetic relationships among soybean landrace populations; and (3) to assess the extent and genomic structure of LD between pairs of SSR markers on a genome-wide scale in the populations. The results of this study will help to utilize, conserve, and manage Chinese soybean landraces effectively, and provide molecular information to broaden genetic diversity of soybean in the world.

Materials and methods

Plant materials

Sampling

In the Chinese national soybean gene bank, 23,587 soybean landraces have been collected, conserved and documented mainly with phenotypic traits that are stable during 2–3 years of observations. Based on the agronomic characteristics in the catalog (Wang 1982; Chang and Sun 1991; Chang et al. 1996), Qiu et al. (2003) compared 20 sampling strategies and proposed an optimal sampling strategy and sample size for developing a Chinese soybean core collection. All collections were first divided according to their origin into three production areas and within each of these production areas, soybean landraces were divided into ecotypes according to planting-time type (Table 1). From each ecotype, 9% of the collections were sampled to form the primary core selection. In this way, a primary Chinese soybean core collection was constructed including 2,794 cultivars from eight ecotypes in three large growing regions. These cultivars were planted and observed for 2 years again for verifying the consistency of phenotype within the cultivars. The primary core collection of summer-planting soybeans from HR represented 100% of the phenotypic and 86% of the SSR allelic variation in the ecotype (Cui et al. 2003). Because genetic recombination and artificial selection would interfere with analyses of germplasm origins, commercial varieties were excluded from this dataset. The landraces with SSR data missing for more than five loci were also removed from the study. This resulted in a sample of 1,863 gene bank accessions of landraces that were used in this study. The landraces were collected from 24 provinces, cities or municipalities in China, and are estimated to represent over 86% of the genetic diversity of 23,587 cultivars in the Chinese national soybean gene bank (Cui et al. 2003; Qiu et al. 2003).

SSR genotyping

Preliminary assessments of SSR polymorphic loci were conducted using 80 cultivars from South Autumn planting ecotype (SAu) collections with markers selected from previously reported soybean linkage groups (Cregan et al. 1999). Out of 200 SSR loci a total of 60 polymorphic and representative markers were identified (Xie et al. 2003) as the core SSR loci. These 60 markers were further confirmed to be highly polymorphic and reproducible in a different set of 190 diverse soybean cultivars (Wang et al. 2003). Here, we amplified 60 SSR loci (52 loci from the core SSR loci and eight additional loci from other tests) on the core collection of 1,863 soybean landraces. These SSR loci were located on 20 integrated genetic linkage groups, covering 1570.39 cM of soybean genome with 26.62 cM average genetic distance between adjoining loci. One SSR marker (Satt157) was removed from the dataset since it was missing in more than 186 landraces (10% of the total landraces). Table S1 (Online Supplementary material) lists the SSR loci, linkage groups and positions, and repeat motifs.

For each landrace five seeds with the same phenotype for various traits including color of seed coat and hilum and shape of seed were selected for DNA extraction. DNA was extracted from a bulk of these five seeds. DNA extraction, PCR amplification, and SSR genotyping followed the methods prescribed by Xie et al. (2003) and Wang et al. (2006). The total number of polymorphic alleles at each SSR locus was derived from calculations from all 1,863 landraces. Polymorphic SSR alleles present in only one of the 1,863 landraces were defined as unique alleles, while polymorphic alleles present in fewer than 19 (1%) of 1,863 landraces were recorded as rare alleles (details for each marker in Table S1).

Genetic structure and phylogenetic analysis

Bayesian clustering was applied on the multi-locus SSR data by using the software package STRUCTURE (Pritchard et al. 2000) in its revised version 2.1 (Falush et al. 2003). Admixture model and independent allele frequency model were used to analyze the dataset without prior population information. Five runs of STRUCTURE were done for each number of populations (*K*) (set from 1 to 10). Burn-in time and replication number were both set up to 80,000 in each run. The maximum likelihood ratio was used to assign landraces to clusters, and the cut-off probability for assignment to a cluster was 0.60. The inferred groups are referred to as ‘clusters’ to distinguish them from the predefined ecotypes as listed in Table 1. They consist of groups of landraces each sampled at a different agricultural field across China. While using STRUCTURE we have assumed that the inferred populations are in Hardy Weinberg equilibrium (HWE). Soybean is a selfing crop, but as levels of heterozygosity were low for all landraces we do not expect a bias in the results. A second assumption of STRUCTURE is that the loci are unlinked, but Pritchard and Wen (2004) indicated that linked data can be used, especially when from several linkage groups. Our markers cover all linkage groups in soybean (Table S1).

Neighbor-joining (with 1,000 bootstraps) was carried out in TREECON 1.3b (Van de Peer and Wachter 1994; available at http://bioinformatics.psb.ugent.be/software/details/). The tree was rooted using one wild species (*G. soja*) accession as outgroup.

Genetic diversity of populations

Allelic richness and gene diversity were estimated using Fstat; allelic richness was corrected for sample size (help file FSTAT293 (Goudet 2002), available at http://www.unil.ch/popgen/softwares/fstat.htm). Observed number of alleles in each population (Nt) and the number and proportion of population-specific alleles were calculated with POPGENE 1.31 (Yeh and Boyle 1997, available at http://www.ualberta.ca/~fyeh/index.htm).

Linkage disequilibrium

Linkage disequilibrium was evaluated for each pair of SSR loci using TASSEL (http://www.maizegenetics.net/bioinformatics/tasselindex.htm), both on all landraces and on the clusters as inferred by STRUCTURE. D' and r^2 LD measures modified for loci were used (Hedrick 1987; Weir 1996). Significance (*P* values) of D' for each SSR pair was determined by 100,000 permutations. For each SSR locus, the rare alleles (i.e., those present in less than 1% of the landraces) were combined into one allelic class described by Maccaferri et al. (2005). A linear regression (PROC NLIN in SAS software) was performed to investigate the decay of linkage disequilibrium with genetic distance. To compare LD at the level of (1) the whole population and (2) individual clusters while accounting for differences in population sizes, we drew random samples from the entire set of a size equal to the actual number of genotypes in the cluster (Liu et al. 2003). This procedure was repeated 30 times and the results were averaged.

To identify to what extent LD was shared among pairs of clusters in the form of common haplotypes we calculated multilocus SSR haplotype sharing as in Koopman et al. (2007), except that haplotypes need not be inferred but were determined directly from the genotypic data, as nearly all loci were homozygous.

Table 1 Eco-geographic distribution and description of the origin of soybean landraces in China

Ecotype code	Production area	Sub-region	Origin (Province)	Province code	Longitude region (ºE)	Latitude region (ºN)	No. of accessions
NR	Northern Region (single cropping)				82.5–132.6	28.2–51.4	796
NESp		North-East Spring planting sub-region			113.1–132.6	38.45–51.4	329
			Helongjiang	HLJ	122.0–132.6	44.3–51.4	78
			Jilin	JL	120.4–130.2	41.1–45.4	117
			Liaoning	LN	118.4–125.2	38.5–43.6	125
			Neimenggu	NMG	113.1–120.0	40.5–43.5	9
NSp		North Spring planting sub-region			82.5–119.6	28.2–44.3	467
			Neimenggu	NMG	107.3–118.4	40.3–43.4	31
			Peking City	Peking	115.6–116.5	39.4–40.3	15
			Hebei	HeB	114.1–119.6	37.4–41.6	70
			Shanxi	SX1	110.3–114.2	34.4–40.3	254
			Shaanxi	SX2	106.5–111.1	34.1–39.0	32
			Gansu	GS	99.5–116.3	28.2–39.6	43
			Ningxia	NX	105.1–107.2	37.3–38.5	20
			Xinjiang	XJ	82.5–86.3	42.0′–44.3	2
HR	Huang-Huai Region (double-cropping)				105.4–122.2	31.3–40.6	276
HSp		Huang-Huai Spring planting sub-region			114.5–119.5	33.3–34.5	45
			Jiangsu	JS	114.5–119.5	33.3–34.5	45
HSu		Huang-Huai Summer planting sub-region			105.4–122.2	31.3–40.6	231
			Hebei	HeB	114.2–117.3	36.3–40.6	39
			Henan	HeN	110.0–116.2	32.0–35.6	49
			Shandong	SD	115.2–122.2	34.5–37.5	45
			Anhui	AH	115.4–118.2	31.3–33.6	35
			Gansu	GS	105.4–106.2	33.5–33.5	7
			Jiangsu	JS	116.3–120.1	33.2–34.5	27
			Shaanxi	SX2	106.2–110.5	32.2–35.2	29
SR	Southern Region (multiple cropping)				34.5–121.5	19.4–34.5	791
CSp		Changjiang (Yangtse River) Spring planting sub-region			102.1–120.5	25.3–32.6	123
			Anhui	AH	116.2–118.3	29.5–31.3	7
			Hubei	HuB	109.3–114.6	29.3–30.4	11
			Hunan	HuN	110.3–113.3	25.3–28.5	5
			Jiangsu	JS	118.5–120.5	31.3–32.6	8
			Jiangxi	JX	114.1–118.2	26.2–28.4	6
			Sichuan	SC	102.1–117.3	26.4–32.4	80
			Zhejiang	ZJ	118.5–120.4	28.5–30.1	6
SSp		South Spring planting sub-region			104.2–121.3	21.4–28.3	143
			Fujian	FJ	117.3–118.5	23.4–25.2	7
			Guangdong	GD	110.2–116.2	21.4–25.1	25
			Guangxi	GX	105.5–109.4	21.4–25.1	6

Table 1 continued

Ecotype code	Production area	Sub-region	Origin (Province)	Province code	Longitude region (ºE)	Latitude region (ºN)	No. of accessions
			Guizhou	GZ	104.2–108.3	25.1–28.3	104
			Taiwan	TW	121.3	25.0	1
SAu		South Autumn planting sub-region			85.6–120.4	23.4–30.3	38
			Fujian	FJ	116.2–118.6	23.4–26.5	5
			Guangxi	GX	110.5	24.5	1
			Hunan	HuN	111.2–112.6	25.1–28.1	7
			Jiangxi	JX	114.6–117.6	26.2–28.6	6
			Zhejiang	ZJ	118.4–120.4	28.1–30.3	17
			Tibet	Tibet	85.6–97.3	28.1–28.4	2
SSu		South Summer planting sub-region			99.2–121.5	19.4–34.5	487
			Anhui	AH	116.1–118.6	29.5–33.2	26
			Fujian	FJ	116.2–119.5	25.5–27.6	16
			Guangdong	GD	110.2–113.2	21.4–25.1	6
			Guangxi	GX	106.1–111.2	21.4–25.6	30
			Guizhou	GZ	105.4–109.1	25.4–28.4	42
			Hainan	HN	110.0–110.2	19.4–19.4	1
			Hubei	HB	108.5–115.4	29.2–33.0	115
			Hunan	HN	109.2–113.4	25.1–29.4	40
			Jiangsu	JS	118.4–121.4	31.1–32.6	44
			Jiangxi	JX	114.2–117.6	25.4–29.2	11
			Shanghai City	SH	121.1–121.3	30.6–31.4	10
			Sichuan	SC	101.3–109.3	26.4–32.4	78
			Yunnan	YN	99.2–104.5	22.4–28.1	32
			Zhejiang	ZJ	111.1–121.5	27.3–34.5	36

The ecotypes were defined according to differences in planting-time type and various phenotypic traits [see Qiu et al. (2003) for details]

Results

Eco-geographic distribution of primary core collections

The 23,587 soybean landraces were collected from 29 of the 31 provinces or autonomous regions in China. There were no soybean collections from Qinghai and Tianjin. Qinghai province is located in the northwestern China. It has highly elevated and widely spread forage lands, on which very little soybean production has been spotted. Tianjin is a small area located to the east of Peking and encircled by the province of Hebei. Based on differences in how soybean is cultivated we have classified the sampled germplasm into three large eco-regions (NR, HR, and SR) with a few overlapping areas between bordering provinces (Fig. 1). The Northern eco-region (NR) includes the landraces from 11 provinces located above 40°N or located in the high-elevated northwestern part of China where only one full season crop of spring-type soybean is grown. Based on the climatic conditions and geographic separations, this ecotype is further divided into two ecotypes: Northeast Spring-type (NESp) and North Spring-type (NSp). In NR eco-region, we sampled a total of 796 landraces in the primary core collection (329 from NESp and 467 from NSp). Most (254) of these collections were from Shanxi province, which is located along the central valley of the Yellow River where the largest variation in cultivated and wild soybeans was identified (Xu and Lu 1983; Chang 1989). Only two samples were selected from Xinjiang, where historical soybean production has been limited by a large area of arid desert and dry conditions. However, soybean acreage has dramatically increased in that region as irrigation systems are improved, therefore, increased breeding activities and utilization of adapted soybean landraces is anticipated.

The Huang-Huai eco-region (HR) includes seven provinces in central and east China (~30 to 40°N), representing one of the most productive agricultural areas. A total of 276 landraces were selected from these eco-regions in the primary core collection [45 spring-types (HSp) and 231 summer-types (HSu)]. As rapid industrialization develops, conservation of soybean landraces becomes increasingly important and urgent in the eastern part of this eco-region.

The Southern eco-region (SR) includes 16 provinces from east, southeast, and southwest China. This eco-region has the largest eco-geographic diversity, from rainy coastal Shanghai (31°N) to subtropical Hainan Island (19°N). A total of 791 landraces were sampled in the primary core collection [266 spring-types, 487 summer-types (SSu), and 38 autumn-types (SAu)]. The spring-type was further divided into Changjiang (Yangtse River) (CSp, 123) and South (SSp, 143) types. Two landraces (SAu) from Tibet (28°N, 4,000 m above sea level) were also included in the primary core collection.

Allelic variation at SSR loci

The 59 SSR markers generated a total of 1,160 alleles among the 1,863 soybean landraces. Table S1 present the basic data on these loci including Ne (effective number of alleles), Ho (observed heterozygosity), He (expected heterozygosity) and F_{st}. The average number of alleles per locus was 19.7, ranging from 2 (Sct_188) to 41 (Satt281). The smallest and the largest alleles at each given SSR locus are listed in Table S1. Satt130 produced the largest range of alleles with fragment sizes from 229 to 396 bp. A total of 97 (8.4% of the 1,160) alleles unique for one landrace were detected in 46 of the SSR loci. An additional 485 (42% of the 1,160) rare alleles (present in less than 1% of the landraces) were observed (in all SSR loci except Sct_188). Ne ranged from 1.85 (Satt387) to 16.12 (Satt462). On average it was 6.66. Consistent with the self-pollinating nature of soybean, heterozygotes were rare. The average frequency of heterozygote samples (bulks of five individual plants) was 1.43% across marker loci. For all loci, F_{st} values and Ne values were estimated in both groups of samples (Table S1).

Interestingly, landraces from the same group tended to have alleles with similar fragment sizes for the same SSR loci, providing additional evidence for the differentiation among clusters. This was confirmed by the finding that R_{st} values were very similar to F_{st} values (not shown).

Clusters of soybean landraces in China

STRUCTURE was run for $K = 1$–10 based on the distribution of 1160 different alleles at 59 SSR loci among 1,863 soybean landraces (Fig. 2). The value of log Pr(X/K) (i.e., the log likelihood of the model given that K clusters are present in the data) was the highest at $K = 7$. When landraces were subdivided into seven clusters, 81.7% of landraces were strongly assigned to one population or another and the proportions assigned to each group are asymmetric, which strongly indicated that there existed real population structure. Anyway, from $K = 7$ to $K = 8$, the frequency of landraces which could not be clearly assigned to any one of these clusters was increased largely (from 19.3 to 57.5%) when the maximum likelihood ratio was set 0.60 or more (Fig. 2). Based on Pritchard and Wen's (2004) suggestion we chose seven clusters. This number is in good accordance with planting ecotypes and sowing seasons of cultivated soybean in China, and indeed landraces from these ecotypes ended up in different clusters. For clarity we denote the seven clusters with the name of the major constituent followed by 'M' (modeled).

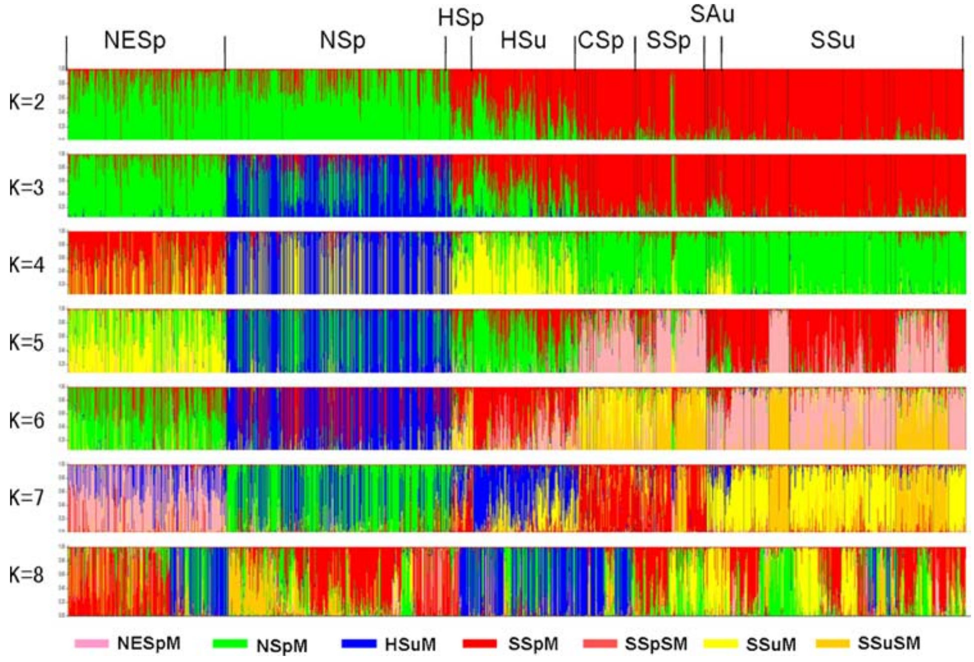

Fig. 2 Inferred population structure of soybean landraces in China. Each landrace is represented by a single vertical line, which is partitioned into K (2–8) colored segments. Each color represents one cluster, and the length of the colored segment shows the landrace's estimated proportion of membership in that cluster as calculated by STRUCTURE in a typical run at that value of K. The original ecotypes (Table 1) are indicated above the figure and inferred seven clusters ($K = 7$) are indicated below the figure

Cluster 1, denoted NESpM (short for Northeast, spring, modeled), consisted of 202 landraces of which 199 (98.5%) were from the ecotype NESp, notably from the most Northern ecotype (39.42–51.43°N) in China including Heilongjiang, Jilin, Liaoning provinces and a part of the eastern ecotype in Neimongol municipality where soybeans are sown in spring and harvested in autumn.

Cluster 2 was named NSpM since all of its 304 landraces were from the ecotype NSp, collected from the mid-down stream in Yellow River valley in China including Gansu, Hebei, Neimongol, Ningxia, Shanxi, as well as Shaanxi and Peking City (99.46–119.56 E, 34.00–43.36°N). They are cultivated as a spring sowing type.

Cluster 3, denoted HSuM, contained 275 landraces, of which almost half (43.6%) were landraces from the HSu ecotype, namely from Gansu, Hebei, Henan, Jiangsu, Shandong, and Shanxi provinces. They are cultivated as double cropping planting type (soybean planted in summer). This cluster contained more than half of all HSu ecotype landraces. Another important part of this cluster (42.9%) was formed by one quarter of all landraces from the NSp ecotype. Geographically the planting regions of HSu and NSp ecotypes do overlap, but this was previously overlooked. In fact, summer soybean/winter wheat double cropping system is the main cultivation system in the middle plain area and southern region in Shanxi Province. Indeed, among the 85 landraces from Shanxi province that were assigned into HSuM, 72 landraces came from these regions. Therefore we can consider that planting characteristics were previously misclassified and that the most proper name of these summer ecotypes is indeed HSuM.

Cluster 4 contained 82 landraces drawn from all predefined Spring-sowing ecotypes—48.8% from CSp, 20.7% from SSp, 15.9% from HSp, 12.2% from NESp and 2.4% from NSp. They were from Jiangsu, Fujian, Guangdong, and Sichuan provinces. Since these spring-planting landraces were mainly (69.5%) collected from south China, we named it as SSpM.

Cluster 5 consisted of 116 landraces (57 from CSp and 57 from SSp, 2 from SSu). As was the case for cluster SSpM, most (98.3%) of the landraces came from spring-planting ecotypes in south China. Since 77.6% of the landraces were collected in the Southwest region of China (Sichuan and Guizhou provinces), we named this second spring-planting cluster SSpSM.

Cluster 6 was comprised of 365 landraces that were cultivated in Anhui, Fujian, Guangdong, Guangxi, Henan, Hebei, Jiangsu, Jiangxi, Shanghai, and Zhejiang provinces. Among them, 292 landraces (80% of the total) were collected from summer-planting ecotypes in south China outside Guizhou, Yunnan and Sichuan provinces. Therefore we named this cluster SSuM. In addition, it includes 31 (81.6%) of the landraces from ecotype SAu.

Cluster 7 consisted of 160 summer sowing type landraces, of which 135 from SSu (mainly from Guizhou, Yunnan and Sichuan provinces). We named this cluster SSuSM.

In addition to the landraces that were clearly assigned (probability > 0.60) to a single population, 359 landraces (19.3% of the total) were not clearly assigned to any one of these clusters. Only 8.8% of the landraces of NSp remained unassigned, but as many as 42.2% of the landraces of HSu.

Genetic relationships among clusters

Overall F_{st} among clusters is 0.130 (95% confidence interval 0.113–0.150) with F_{st} for each locus ranging from 0.034 to 0.322. Pairwise comparison on the basis the values of F_{st} can be interpreted as standardized population distances between two populations (Li and Nelson 2001; Chen and Nelson 2005). The pairwise F_{st} values among seven clusters ranged from 0.048 between the southern clusters SSuSM and SSuM to 0.194 between NSpM from northern China and SSpSM from southern China (average pairwise $F_{st} = 0.122$). All values were significantly different from zero ($P < 0.001$) (Table 2).

Table 2 Pairwise estimates of F_{st} based on 59 SSR loci among the seven model-based clusters of Chinese soybean landraces

Subdivision	NESpM	NSpM	HSuM	SSpM	SSpSM	SSuM	SSuSM
NESpM		10	23	12	14	13	9
NSpM	0.188	0.114	18	9	13	12	12
HSuM	0.109	0.164	0.092	19	19	17	19
SSpM	0.140	0.194	0.122	0.100	*27*	*30*	*30*
SSpSM	0.169	0.165	0.089	0.058	0.081	*28*	*31*
SSuM	0.152	0.170	0.101	0.071	0.074	0.048	*35*
SSuSM	0.158						

Below the diagonal are pairwise F_{st} values between clusters. All pairwise F_{st} values are significant at 0.001 level (obtained after 21,000 permutations and adjusted for multiple comparisons). Above the diagonal are the number of loci which have the same highest frequency allele (in italic and underlined the comparisons of clusters from Southern China)

Alleles were consistently more common in one cluster than another. The highest-frequency alleles among 59 loci in each cluster were listed in Table S1. We calculated, for each pair of clusters, the number of loci for which the same allele had the highest frequency (Table S1). In accordance with the result of F_{st} analysis, more loci had the same most frequent allele among the four Southern clusters, up to 35 loci between SSuM and SSuSM.

A Neighbor-joining tree with bootstrap support values (>50%, based on 1,000 bootstraps) for the 1,863 soybean landraces in China (Fig. 3) showed good agreement with the STRUCTURE analysis. Except those landraces (with blue color) at the bottom of the graph which were assigned into HSuM rather than clusters in South China most landraces from one cluster were grouped together. Some landraces from clusters in south China grouped together in the tree, consistent with smaller genetic differences among some of the clusters from southern China. All of the F_{st} values were statistically significant (Table 2), but this may, at least partly, be the result of the large number of loci and landraces involved. Indeed, when the F_{st} analysis between the overlapping south China clusterss SSuM and SSuSM was rerun using subsamples of 30 landraces and 20 loci, they were not significant, while those from other comparisons generally were (not shown).

Genetic diversity of clusters of populations

All 59 SSR loci were polymorphic in each cluster, with always a large excess of homozygotes. Of a total of 1,160 alleles 552 were found in all seven clusters and 291 were unique for one cluster.

Gene diversity among the seven clusters were not significantly different ($P > 0.05$), but allelic richness was ($P < 0.001$), as well as the number of alleles that were population-specific (Table 3). Among the seven clusters, SSuM, whose landraces have two planting types (summer and autumn), had the highest allelic richness (on average 10.5 alleles per locus) and the largest number of cluster-specific alleles (71 in total, or 8.5% of the alleles present) followed by HSuM. HSuM had the highest gene diversity and the second highest allelic richness and number of cluster-specific alleles. The difference in gene diversity and allelic richness between SSuM and HSuM was not significant. NESpM, NSpM and SSpSM had lower allelic richness regardless of the sample size analyzed.

Linkage disequilibrium: effect of population stratification

The extent of LD was assessed among all 1,711 pairs of SSRs loci for all landraces as well as for the seven clusters separately (Table 4). Across all landraces, as many as 95.2% of total pairs of marker pairs were in LD (based on D', $P < 0.05$) after Bonferroni-correction. The value of 96.7% for pairs of markers from the same chromosome (within the same linkage group) was only slightly higher

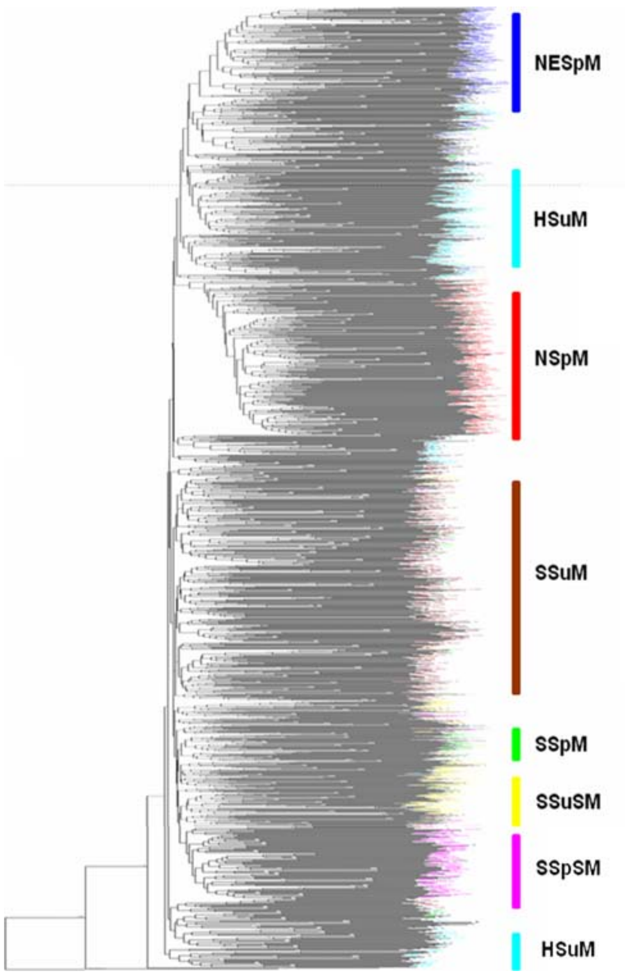

Fig. 3 Neighbour-joining tree with bootstrap support values (>50%, based on 1,000 bootstraps) for the 1863 soybean landraces in China. The tree was rooted using one wild species (*G. soja*) accession as outgroup.The *colored bars* indicated the position of the landraces which were assigned to seven inferred clusters

Table 3 Summary statistics of the seven model-based clusters in China

Cluster	No. of landraces	Nt	Allelic richness[a]	No. of cluster-specific alleles	Gene diversity
NEspM	202	516	7.5	25	0.711
NspM	304	491	7.4	31	0.691
HsuM	275	762	10.3	47	0.753
SspM	82	539	9.1	23	0.748
SspSM	116	465	7.5	17	0.685
SsuM	365	837	10.5	71	0.730
SSuSM	160	598	8.9	18	0.725

Nt Observed number of alleles for 59 SSR loci in each cluster

[a] Calculated using rarefaction on a sample of 46 landraces per cluster

Table 4 Percentage of SSR locus pairs in significant ($P < 0.05$) linkage disequilibrium (LD) and LD statistics D' and r^2 of soybean landrace populations in China

Samples	No. of pairs of loci in LD							The extent of LD					
	Markers on the same linkage group		Markers from different linkage groups		Total		Expected fraction of locus pairs ± Standard Error	Markers on the same linkage group		Markers from different linkage groups		Total	
	No. of locus pairs in LD[a]	Fraction of locus pairs	No. of locus pairs in LD[a]	Fraction of locus pairs	No. of locus pairs in LD	Fraction of locus pairs		D'	r^2	D'	r^2	D'	r^2
All	58 (60)	0.967	1,441 (1515)	0.951	1,499	0.952		0.309	0.008	0.308	0.008	0.308	0.008
NESpM	31 (62)	0.500	687 (1591)	0.432	718	0.434	0.579 ± 0.0327	0.316	0.020	0.293	0.014	0.294	0.015
NSpM	10 (57)	0.175	233 (1567)	0.149	243	0.150	0.581 ± 0.0362	0.257	0.009	0.271	0.010	0.271	0.010
HSuM	16 (59)	0.271	336 (1520)	0.221	352	0.223	0.433 ± 0.0393	0.413	0.020	0.413	0.018	0.413	0.018
SSpM	9 (59)	0.153	283 (1536)	0.184	292	0.183	0.325 ± 0.0402	0.523	0.030	0.522	0.030	0.522	0.030
SSpSM	12 (60)	0.200	281 (1549)	0.181	293	0.182	0.217 ± 0.0475	0.480	0.042	0.470	0.036	0.471	0.036
SSuM	11 (57)	0.193	222 (1459)	0.152	233	0.154	0.202 ± 0.0449	0.386	0.013	0.379	0.013	0.379	0.013
SSuSM	8 (54)	0.148	233 (1477)	0.158	241	0.154	0.206 ± 0.0359	0.455	0.026	0.466	0.027	0.466	0.027

[a] Based on average percentage of all locus pairs showing significant D' in random samples containing the same number of landraces. Between brackets the actual number of locus pairs used in analysis

than that of 95.1% for markers from different chromosomes.

The frequency of pairs of loci with significant ($P < 0.05$) LD (based on D') was reduced by more than half when LD was calculated within each cluster. The lowest percentage of locus pairs in LD (15.0%) was found in the NSpM cluster despite it being the largest population. Cluster NESpM had the highest percentage (43.4%). To evaluate this further, we used the method of Liu et al. (2003) to correct for the fact that larger population size leads to an increased power for detecting locus pairs in LD. The expected percentage of SSR locus pairs in LD in the total population was calculated for the population size of each cluster and compared to the percentage observed within that cluster. This showed that population size could not explain all, as the observed percentage of locus pairs in LD within each cluster was lower than the expected percentage. This suggests that genetic structure influences LD in soybean landraces.

Linkage disequilibrium: effect of genetic distance between loci

Within clusters, the LD found could be due partly due to the existence of (physical) linkage between loci located several centiMorgans away on the same chromosome. Therefore, we undertook two different analyses on the subset of pairs of loci that were located on the same linkage groups.

First, if physically long haplotypes would be maintained in the cluster, then one would expect that the statistical frequency of pairs of loci in high LD would depend on the distance between the two markers. Indeed, a scatter plot of the LD statistics as a function of the inter-marker distance (Fig. 4a) showed that D' decreased with increasing distance in some clusters, notably in the southern China clusters SSpM and SSuSM, but not so in others. No statistically significant correlation was found between inter-marker distance and D' or r^2 in seven model-based clusters. However, consistent with results above both r^2 and D' were slightly higher in individual clusters than in the complete set of landraces (average D' across clusters was 0.402 compared to 0.308 overall).

The second analysis was done to determine whether specific multilocus microsatellite haplotypes (cf. Koopman et al. 2007) could be detected in multiple landraces from one cluster, compared to finding the same long haplotype shared between landraces from different clusters. Cluster NESpM contained the highest number of shared pairs of loci of all clusters, and also the highest level of shared pairs with other clusters. In contrast, no single pair of loci in LD was shared between SSpM and SSpSM and between SSpSM and SSuSM. Moreover, the number of shared locus

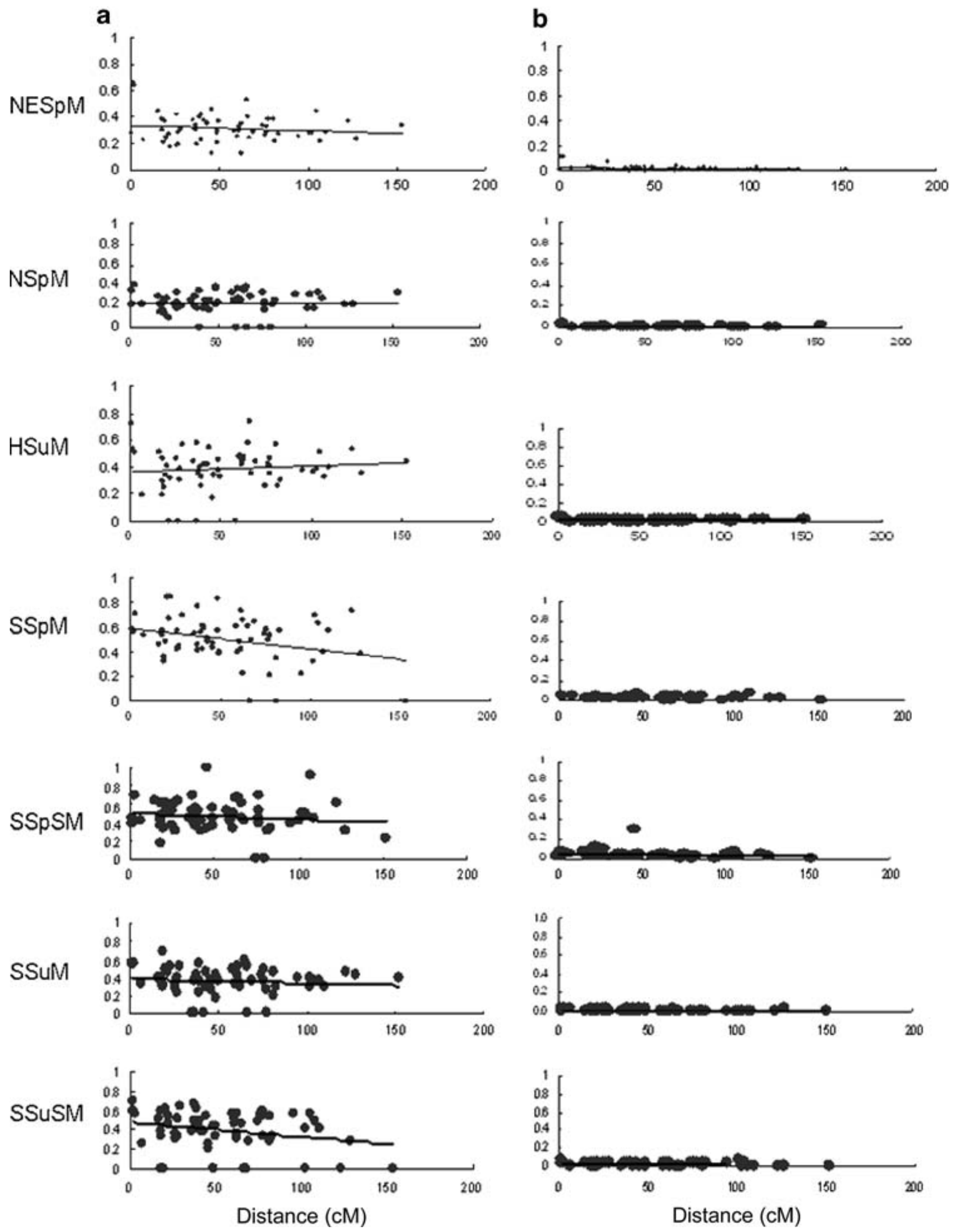

Fig. 4 Scatterplot of the LD statistics D' (**a**) and r^2 (**b**) within seven model-based clusters of soybean landraces in China

pairs in LD among all southern clusters was lower than those of other pairwise clusters.

Discussion

The abundant allelic variation per locus in this study (19.7) exceeded the previous reported 5.0–15.8 alleles per locus (Maughan et al. 1995; Diwan and Cregan 1997; Narvel et al. 2000; Abe et al. 2003). This may be caused by our much extended sampling. However, it does indicate a high level of genetic diversity among the Chinese landraces.

The unique or rare alleles in this study were always detected at those loci with more polymorphic alleles, which is in agreement with the findings reported in rice (Jain et al. 2004). As these unique or rare alleles are likely caused by natural mutation and selection (Mousadik and Petit 1996), they can be used not only in the specific categorization of germplasm collections, but also in their subsequent utilizations in breeding and plant development as unique markers. Rareness at SSR loci, especially in plants with rare alleles at multiple loci, may indicate that these landraces have had limited genetic exchange with other landraces, and therefore may have rare alleles for various (functional) traits as well. As genetic improvement of a trait often involves finding useful variants which may be present at a low frequency (Chamberlain and Bernard 1968; Van Duyn et al. 1971; Anand and Gallo 1984; Anand et al. 1988; Coordinative

Group of Evaluation of SCN 1993), this could be a means of identifying plants with possible useful alleles. Hyten et al. (2006) found that 79% of the low-frequency sequence variants in the Asian Landrace collection were not present in the elite cultivars, which suggests that the variation useful for genetic improvement is not yet present in the elite cultivars. These studies suggested that identification of the larger number of unique and rare alleles is important, and thus confirmed the necessity and perspectives of our research with a large number of landraces. Conservation of unique germplasm collections is part of the process of maintaining the largest amount of genetic diversity in the gene pool.

Genetic structure

Previously, soybean landraces have been classified into ecotypes on the basis of region only. Genetic differences between regions were determined to some extent, but it was not known whether this classification was optimal. In addition, the classification was identical for each landrace from a certain region, ignoring differences among landraces, possible substructure, and admixture through trade of seeds. We have obtained genetic information for 59 single sequences repeats (SSR) loci and based solely on this genetic information we have used Bayesian clustering to group 1,863 Chinese cultivated soybean landraces into seven clusters. The seven model-based clusters, NESpM, NSpM, HSuM, SSpM and SSuM, which contained 1,504 landraces with a probability >0.60 to one of the clusters, do correspond to geographically defined ecotypes, with mostly (but not exactly) the same composition of landraces. It was noted that HSu landraces collected between 32.0 and 41.6°N along the mid-stream of the Yellow River valley were assigned to one cluster, termed HSuM, while the majority of the HSu landraces from 31.3 to 34.5°N (adjacent to South China) were assigned to SSuM. This indicates that the former HSu ecotype consists of two genetically distinct groups of landraces, which have a distinct geographical distribution.

The other two southern clusters do not have such a clear geographically distinct distribution. SSpSM and SSuSM were defined as new independent clusters which were distinct in composition of landraces from SSpM and SSuM, respectively. The landraces of SSpSM mainly came from Sichuan and Guizhou province, while the landraces of SSuSM were mainly from Sichuan, Guizhou and Yunnan province. Similarly, Gai and Wang (2001) already defined landraces from YunGui plateau region, the mountainous areas around the Sichuan and the plateau of western Sichuan as a distinct cluster according to phenotypes (mainly based on growing period) and the type of natural and cultivating condition. Yunnan, Sichuan and Guizhou provinces, which are adjacent to each other, have a special climatic environment by a surrounding mountain chain. Combined with large differences in elevation, selection may have led to the formation of these special cultivated soybean populations. Such environmentally distinct populations can be very useful for broadening the genetic base of soybean cultivars.

Among ecotype populations in soybean in China, SAu had attracted attention for its low number of landraces, its limited distribution region, and the occurrence of very widely divergent phenotypes. As much as 81.6% of the SAu landraces were assigned to SSuM, supporting the opinion that SAu landraces are a small group that derived from SSuM. This may at least partly be due to the fact that in SAu region some summer-planting varieties (which is common for SSuM landraces) were used as autumn-planting varieties (sowed in the last ten days of July to the first 10 days of August in South China).

In our analyses, landraces from each geographically predefined ecotype were assigned to at least four model-based clusters, in some cases (e.g., SSp) even to six of the seven clusters. This major reshuffling may have at least two reasons. First, our analysis identified those landraces which were introduced to a new region based on genetic similarity to the landraces in the original area. An example is the 7.6% NESp landraces that were assigned into HSuM. Second, some planting characteristics were ignored during the previous classifications, especially when it concerned minority characteristics in that region. For example, 118 (25.3%) of the NSp landraces were assigned into HSuM, with which they indeed share the characteristic planting in summer, while the majority of NSp is spring-sowing type. Third, previous classifications used morphological information to infer genetic relationships, but there generally is no direct relationship between morphological and genetic distance, as large morphological differences in a few conspicuous traits may be based on differences in a few genes only. This is consistently found in variety assessment studies as well (Pearse and Crandall 2004).

In addition to the landraces that were clearly assigned to a single population, 359 landraces (19.3%) did not have a clear probability (>0.60) for any of the seven clusters. The frequencies of these unassigned landraces assignment into the mixed cluster (Fig. 2) ranged from 8.8% (landraces from ecotype NSp) to as many as 42.2% (ecotype HSu). Both gene diversity and allelic richness of the group of unassigned landraces, if treated as a group, were higher than those of the seven model-based clusters, suggesting that it indeed is not one homogeneous group. As many as 940 alleles (81% of the total) were observed, consistent with a highly diverse origin. Rosenberg et al. (2002) proposed that a mixed cluster might reflect continuous gradations in allele frequencies across regions or admixture of neighboring groups, indicating that it is not one homogeneous group.

Genetic diversity

The region with the highest genetic diversity is generally considered as the center of origin of a species (Vavilov 1951; Chang 1989; Dong et al. 2004). Among the seven clusters, SSuM had the highest allelic richness and largest number of unique (i.e., cluster-specific) alleles, while HSuM had the highest gene diversity. The HSuM landraces cultivated in the central and downstream area of the Yellow River valley come from Hebei, Henan, Anhui, Gansu, Shaanxi and Shanxi provinces. Many studies have hypothesized that the central and downstream part of the Yellow River valley is the centre of origin of cultivated soybean, which is a kind of rain-fed crop (Vavilov 1951; Hymowitz and Newell 1981; Chang 1989; Dong et al. 2004; Zhao and Gai 2004). Dong et al. (2004) found that the cultivated soybean landraces in this region exhibited higher diversity based on ten quality traits and five quantitative traits. Our results on HSuM landraces are also consistent with these earlier studies. In contrast, no one has proposed that the landraces we clustered into SSuM were the first cultivated soybeans originating from wild soybean. Dong et al. (2004) found that the diversity at agronomic traits for southern and southeastern landraces was much lower. Our cluster SSuM, which is genetically highly diverse at our SSR loci, is a specific subset of landraces from south China, namely summer-planting soybean in the south of the Yangtse River. Harlan (1975) considered that secondary centers of origin might possess richer genetic diversity than primary centers. SSuM is a candidate secondary centre of origin. The high genetic diversity within SSuM may be the result of several factors, including large area (from 100.35 to 121.52°E, 19.42 to 34.50°N), complex tilling systems, diverse environments and multiple types of usage.

Linkage disequilibrium

SSR markers have been used for primary evaluation of LD across the genome of crops, including maize (Remington et al. 2001; Liu et al. 2003; Stich et al. 2005) and durum wheat (Maccaferri et al. 2005). As a selfing species, soybean was supposed to have a high level of LD, but Zhu et al. (2003) detected low genome-wide LD in soybean (*Glycine max* L. Merr.) using 49 fragments with three or more single-nucleotide polymorphisms (SNPs), for example, at very short distances. In this study, extensive LD was found within the whole set of landraces: nearly all pairs of loci (95.2%) were in significant LD within soybean in China. The *P* value for the pairwise estimates of LD among all 59 SSR loci were evaluated by the permutation version of Fisher's exact test. This test is affected by population size. Since the collection of almost 2,000 landraces is much larger than that of all previous studies, this could be part of the reason why such higher LD values have been estimated here. More importantly, however, population structure was shown to produce a large part of the overall LD, since the number of 598 locus pairs in LD was reduced sharply within each of clusters compared to the complete dataset. This shows that population structure needs to be accounted for in association studies (Gupta et al. 2005) conducted with this collection. The low value of LD within the clusters can be seen as evidence that much of the recombination events in the past have been maintained in soybean fixed in homozygous self-fertilizing landraces.

Acknowledgments This research was supported by the State Key Basic Research and Development Plan of China (973) (No.2004CB117203), National Key Technologies R&D Program in the 11th Five-Year Plan (No.2006BAD13B05), the National Natural Science Foundation of China (No.30490251 and 30471096) and the Academy and Institute Foundation for Basic Scientific Research in Institute of Crop Science, Chinese Academy of Agricultural Sciences. Dr. Wim J.M. Koopman (Plant Research International, Wageningen UR, The Netherlands) is thanked for a batch program for NTSYS, and Dr. Alain Charcosset and two anonymous reviewers for useful suggestions.

References

Abe J, Xu DH, Suzuki Y, Kanazawa A, Shimamoto Y (2003) Soybean germplasm pools in Asia revealed by nuclear SSRs. Theor Appl Genet 106:445–453

Anand SC, Gallo KM (1984) Identification of additional soybean germplasm with resistance to race3 of the soybean cyst nematode. Plant Disease Report 68:593–595

Anand SC, Gallo KM, Baker IA, Hartwig EE (1988) Soybean plant introductions with resistance to races 4 or 5 of soybean cyst nematode. Crop Sci 28:563–564

Anand SC, Shannon JG, Wrather JA, Arelli PR, Sleper DA, Young LD (2004) Registration of S97–1688 soybean germplasm line high in protein content and resistant to soybean cyst nematode. Crop Sci 44:698–699

Brown-Guedira GL, Thompson JA, Nelson RL, Warburton ML (2000) Evaluation of genetic diversity of soybean introductions and north American ancestors using RAPD and SSR markers. Crop Sci 40:815–823

Bu MH, Pan TF (1987) Soybean Improvement and Cultivation in China. Agricultural Press, Beijing, pp 58–105

Buckler ES, Thornsberry JM (2002) Plant moleculardiversity and applications to genomics. Curr Opin Plant Biol 5:107–111

Chang RZ (1989) Studies on the origin of the cultivated soybean (*Glycine max* (L) merr). Oil Crop Sci 1:1–6

Chang RZ, Sun JY (1991) Catalogues of Chinese Soybean Germplasm and Resources: Continuation I. China Agricultural Press, Beijing

Chang RZ, Sun JY, Qiu LJ, Chen YW (1996) Catalogues of Chinese Soybean Germplasm and Resources: Continuation II. China Agricultural Press, Beijing

Chang RZ, Sun JY, Qiu LJ (1998) The development of soybean germplasm in China. Crops 3:7–9

Chamberlain DW, Bernard RL (1968) Resistance to brown stem rot in soybeans. Crop Sci 8:728–729

Chen Y, Nelson RL (2005) Relationship between origin and genetic diversity in Chinese soybean germplasm. Crop Sci 45:1645–1652

Coordinative group of evaluation of SCN (1993) Evaluation of soybean germplasm for resistance to race 1, 3 and 4 of the soybean cyst nematode. Soybean Sci 12(2):91–99

Cregan PB, Jarvik T, Bush AL, Shoemaker RC, Lark KG, Kahler AL, Kaya N, Van Toai TT, Lohnes DG, Chung J, Specht JE (1999) An integrated genetic linkage map of the soybean genome. Crop Sci 39:1464–1490

Cui YH, Qiu LJ, Chang RZ, Lu WH (2003) Examination of representiveness of the primary core collection in Huanghuai summer sowing soybean (*Glycine max*) using SSR. J Plant Genetic Resour 4(1):9–15

Cui YH, Qiu LJ, Chang RZ, Lu WH (2004a) A study of genetic diversity of Huanghuai summer sowing soybean in China. Sci Agricult Sin 37:15–22

Cui YH, Qiu LJ, Chang RZ, Lu WH (2004b) Representative test for primary core collection of summer sowing soybeans in Huanghuai region of China. Acta Agron Sin 30:284–288

Cui ZL, Carter TE, Burton JW (2000) Genetic diversity patterns in Chinese soybean cultivars based on coefficient of parentage. Crop Sci 40:1780–1793

Diwan N, Cregan PB (1997) Automated sizing of fluorescent-labeled simple sequence repeat (SSR) markers to assay genetic variation in soybean. Theor Appl Genet 95:723–733

Dong YS, Zhao LM, Liu B, Wang ZW, Jin ZQ, Sun H (2004) The genetic diversity of cultivated soybean grown in China. Theor Appl Genet 108:931–936

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

Fukuda Y (1933) Cytogenetical studies on the wild and cultivated Manchurian soybeans (*Glycine L*). Jap J Bot 6:489–506

Gai JY, Wang YS (2001) A study on the varietal eco-regions of soybeans in China. Sci Agricult Sin 34:139–145

Goudet J (2002) FSTAT: A program to estimate and test gene diversities and fixation indices V2932 Available online at http://www.unilch/popgen/softwares/fstathtm Department of Ecology & Evolution, Biology Building, UNIL, CH-1015 LAUSANNE, Switzerland

Guan Y, WD E, Wang LX, Guan RX, Liu ZX, Chang RZ, Qu YY, Qiu LJ (2006) Analysis of factors influencing the genetic diversity evaluation using two soybean [*Glycine max* (L.) Merr.] collections from Hunan and Hubei. Acta Agron Sin 33:461–468

Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485

Harlan JR (1975) Geographic patterns of variation in some cultivated plants. J Hered 66:84–101

Hedrick PW (1987) Gametic disequilibrium measures: proceed with caution. Genetics 117:331–341

Hymowitz T, Newell CA (1981) Taxonomy of genus *Glycine*, domestication and uses of soybeans. Econ Bot 35:272–288

Hyten DL, Song Q, Zhu Y, Choi IY, Nelson RL, Costa JM, Specht JE, Shoemaker RC, Cregan PB (2006) Impacts of genetic bottlenecks on soybean genome diversity. Proc Natl Acad Sci USA 103:16666–16671

Jain S, Jain R, Mccouch S (2004) Genetic analysis of Indian aromatic and qualityrice (*Oryza sativa* L.) germplasm using panels of fluorescently-labled microsatellite markers. Theor Appl Genet 109:965–977

Kisha TJ, Diers BW, Hoyt JM, Sneller CH (1998) Genetic diversity among soybean plant introductions and North American germplasm. Crop Sci 38:1669–1680

Koopman WJM, Li YH, Coart E, Van De Weg WE, Vosman B, Roldán-Ruiz IR, Smulders MJM (2007) Linked vs. unlinked markers: multilocus microsatellite haplotype-sharing as a tool to estimate gene flow and introgression. Molec Ecol 16:243–256

Kraakman ATW, Niks RE, Van den Berg PMMM, Stam P, Van Eeuwijk FA (2004) Linkage disequilibrium mapping of yield and yield stability in modern spring barley cultivars. Genetics 168:435–446

Li LH, Qiu LJ, Chang RZ, He XL (2005) Differentiation and genetic diversity of SSR molecular markers for Huanghuai and southern summer sowing soybean in China. Acta Agron Sin 31:777–783

Li ZL, Nelson RL (2001) Genetic diversity among soybean accessions from three countries measured by RAPDs. Crop Sci 41:1337–1347

Lin FY, Qiu LJ, Chang RZ, He BR (2003) Genetic diversity of landrace and bred varieties of soybean in Shanxi. Chin J Oil Crop Sci 25(3):24–28

Liu KJ, Goodman M, Muse S, Smith JS, Buckler ED, Doebley J (2003) Genetic structure and diversity among maize inbred lines as inferred from DNA microsatellites. Genetics 165:2117–2128

Lu SL, Cheng SH, Cheng CJ (1981) A study on the classification of soybean cultivation regions in China. J Shanxi Agric Univ 1(1):9–17

Luan WJ, Liu ZX, Guan RX, Chang RZ, He BR, Qiu LJ (2005) Representativeness of Northeast China spring soybeans and their genetic diversity at SSR loci. Appl Ecol 16:1469–1476

Maccaferri M, Sanguineti MC, Noli E, Tuberosa R (2005) Population structure and long-range linkage disequilibrium in a durum wheat elite collection. Mol Breeding 15:271–289

Maughan PJ, Saghai Maroof MA, Buss GR (1995) Microsatellite and amplified sequence length polymorphism in cultivated and wild soybean. Genome 38:715–723

McGregor CE, Lambert CA, Greyling MM, Louw JH, Warnich L (2000) A comparative assessment of DNA fingerprinting techniques (RAPD, ISSR, AFLP and SSR) in tetraploid potato (*Solanum tuberosum* L) germplasm. Euphytica 113:135–144

Mi SJ, Qiu LJ, Chang RZ, Hao ZB, Guan RX (2004) Genetic diversity analysis of varieties of *Glycine max* (L.) Merr resistant to Soybean mosaic virus by SSR fingerprints. Acta Phyto Pathol Sin 34:244–253

Morse WJ (1950) History of soybean production. In: Markley KS (ed) Soybean and soybean products. Interscience Publishers, Inc, New York, pp 3–59

Mousadik A, Petit RJ (1996) High level of genetic differentiation for allelic richness among populations of the argan tree [*Argania spinosa* (L.) Skeels] endemic to Morocco. Theor Appl Genet 92:832–835

Narvel JM, Fehr WR, Chu W, Grant D, Shoemaker RC (2000) Simple sequence repeat diversity among soybean plant introductions and elite genotypes. Crop Sci 40:1452–1458

Pan TF, Zhang DR, Zhang WG (1994) A study on the climatic regions of soybeans in China. Soybean Sci 3:169–182

Pantalone VR, Allen FL, Landau-Ellis D (2003) Registration of '5601T' soybean. Crop Sci. 43:1123–1124

Parker HG, Kim LV, Sutter NB, Carlson S, Lorentzen TD (2004) Genetic structure of the purebred domestic dog. Science 304:1160–1164

Pearse DE, Crandall KA (2004) Beyond F_{ST}: analysis of population genetic data for conservation. Conserv Genet 5:585–602

Piao RH, Liu ZX, Guan RX, Chang RZ, Hao ZB, Qiu LJ (2005) Genetic diversity of southern summer soybean in Chinese Coastal Revealed by SSR markers. Agricult Biotechnol 13:435–440

Pritchard JK, Rosenberg NA (1999) Use of unlinked genetic markers to detect population stratification in association studies. Am J Hum Genet 65:220–228

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

Pritchard JK, Wen W (2004) Documentation for *structure* software: Version 2. http://pritchbsduchicagoedu/software/readme_structure2_1pdf

Qiu LJ, Cao YS, Chang RZ, Zhou XA, Wang GX, Sun JY, Xie H, Zhang B, Li XH, Xu ZY, Liu LH (2003) Establishment of Chinese soybean (*G. max*) core collection: Sampling strategy. Sci Agricult Sin 36:1442–1449

Remington DL, Jeffry MT, Matsuoka Y, Wilson LM, Whitt SR, Doebley J, Kresovich S, Goodman MM, Buckler ES (2001) Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc Natl Acad Sci USA 98:1147–11484

Roussel V, Koenig J, Bechert M, Balfouriter F (2004) Molecular diversity in French bread wheat accessions related to temporal trends and breeding programmes. Theor Appl Genet 108:920–930

Rosenberg NA, Prichard JK, Weber JL, Cann HM, Kidd KK, Zhivotovsky LA, Feldman MW (2002) Genetic structure of human populations. Science 298:2381–2385

Stich B, Melchinger AE, Frisch M, Maurer HP, Heckenberger M, Reif JC (2005) Linkage disequilibrium in European elite maize germplasm investigated with SSRs. Theor Apple Genet 111:723–730

Ude GN, Kenworthy WJ, Costa JM, Cregan PB, Alvernaz J (2003) Genetic diversity of soybean cultivars from China, Japan, North America, and North American Ancestral Lines determined by amplified fragment length polymorphism. Crop Sci 43:1858–1867

Van de Peer Y, De Wachter Y (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci 10:569–70

Van Duyn JW, Turnipseed SG, Maxwell JD (1971) Resistance in soybeans to the Mexican Bean Beetle I: source of resistance. Crop Sci 11:572–573

Vavilov NI (1951) The origin variation immunity and breeding of cultivated plants (translation by K Star Chester). Chron Bot Ronald Press, NewYork

Wang B, Chang RZ, Yan L, Tao L, Guan RX, Yan L, Feng ZF, Qiu LJ (2003) Identification of SSR primer numbers for analyzing genetic diversity of Chinese cultivated soybean. Molec Plant Breeding 1:82–88

Wang GX (1982) Catalogues of Chinese Soybean germplasm and resources. China Agricultural Press, Beijing

Wang LX, Li YH, Li W, Zhu L, Guan Y, Ning XC, Guan YX, Liu ZX, Chang RZ, Qiu LJ (2004) Establishment of a core collection of Changjiang spring sowing soybean. Biodiv Sci 12:578–585

Wang LX, Guan RX, Liu ZX, Chang RZ, Qiu LJ (2006) Genetic diversity of Chinese cultivated soybean revealed by SSR markers. Crop Sci 46:1032–1038

Wang JL (1991) Soybean ecotypes. Agricultural Press, Beijing

Wang Y, Hobbs HA, Hill CB, Domier LL, Hartman GL, Nelson RL (2005) Evaluation of ancestral lines of U.S. soybean cultivars for resistance to four soybean viruses. Crop Sci 45:639–644

Wang YS, Gai JY (2002) Study on the ecological regions of soybean in China: Ecological environment and representative varieties. Chin J Appl Ecol 13:71–75

Weir BS (1996) Genetic data analysis II. Sinaur, Sunderland

Xie H, Chang RZ, Cao RS, Zhang MH, Fen ZF, Qiu LJ (2003) Selection of core SSR loci by using Chinese autumn soybean. Sci Agricult Sin 36:360–366

Xie H, Guan RX, Chang RZ, Qiu LJ (2005) Genetic diversity of Chinese summer soybean germplasm revealed by SSR markers. Chin Sci Bull 50:526–535

Xu B, Lu B (1983) Soybean ecology study: day length and temperature study of wild soybean in different latitude. Soybean Sci 2(3):8–9

Xu DH, Abe J, Gai JY, Shimamoto Y (2002) Diversity of chloroplast DNA SSRs in wild and cultivated soybeans: evidence for multiple origins of cultivated soybean. Theor Appl Genet 105:645–653

Yeh FC, Boyle TJB (1997) Population genetic analysis of co-dominant and dominant markers and quantitative traits. Belgian J Bot 129:157

Zhang HY, Guan RX, Li YH, Wang LX, Luan WJ, Chang RZ, Liu ZX, Qiu LJ (2005) Genetic diversity analysis and marker assisted identification of salt tolerant soybean by using SSR marker. Plant Genetic Resour 6(3):251–255

Zhao TJ, Gai JY (2004) The Origin and evolution of cultivated soybean [*Glycine max* (L) Merr]. Sci Agricult Sin 37:954–962

Zhu YL, Song QJ, Hyten DL, Van Tassell CP, Matukumalli LK, Grimm DR, Hyatt SM, Fickus EW, Young ND, Cregan PB (2003) Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134