Front. Biol. China 2007, 2(4): 397–402
DOI 10.1007/s11515-007-0060-0

RESEARCH ARTICLE

Sampling strategy for wild soybean (*Glycine soja*) populations based on their genetic diversity and fine-scale spatial genetic structure

ZHU Weiyue[2], ZHOU Taoying[1], ZHONG Ming[1], LU Baorong (✉)[1]

1 Fudan University, Shanghai 200433, China
2 Shandong Institute of Light Industry, Jinan 250353, China

Abstract A total of 892 individuals sampled from a wild soybean population in a natural reserve near the Yellow River estuary located in Kenli of Shandong Province (China) were investigated. Seventeen SSR (simple sequence repeat) primer pairs from cultivated soybeans were used to estimate the genetic diversity of the population and its variation pattern versus changes of the sample size (sub-samples), in addition to investigating the fine-scale spatial genetic structure within the population. The results showed relatively high genetic diversity of the population with the mean value of allele number (*A*) being 2.88, expected heterozygosity (*He*) 0.431, Shannon diversity index (*I*) 0.699, and percentage of polymorphic loci (*P*) 100%. Sub-samples of different sizes (ten groups) were randomly drawn from the population and their genetic diversity was calculated by computer simulation. The regression model of the four diversity indexes with the change of sample sizes was computed. As a result, 27–52 individuals can reach 95% of total genetic variability of the population. Spatial autocorrelation analysis revealed that the genetic patch size of this wild soybean population is about 18 m. The study provided a scientific basis for the sampling strategy of wild soybean populations.

Keywords **s**ampling strategy, genetic diversity, fine-scale spatial structure, wild soybean, simple sequence repeat (SSR)

1 Introduction

The annual wild soybean (*Glycine soja* Sieb. *et* Zucc.) is believed to be the direct ancestor of the cultivated soybean. It contains rich genetic diversity and useful genetic resources that have not been fully explored. China is the origin and diversification center of wild soybeans. However, the rapid development in the past several decades in the country has considerably changed the natural habitats of wild soybean populations. Due to the severe disturbances by humans, the survival of the wild soybean's natural populations is under threats and their distribution areas have considerably decreased. As a consequence, the biodiversity conservation of wild soybeans is urgently needed.

To have a reasonable and effective sampling strategy for conserving plant genetic resources, we usually need to address two essential questions. (1) How many individuals in the collected samples can adequately represent the genetic diversity of a population? (2) How should the sampled individuals be distributed in the population in terms of space to ensure that the samples contain as many genotypes as possible and to avoid the possible repetition of sampling (Jin and Lu, 2003)? The effective *ex situ* conservation of wild soybeans is dependent on the reasonable sampling strategy for the natural population. This requires the conserved samples from a natural population to contain as much genetic diversity as possible. In other words, the effective sampling strategy for wild soybeans is expected to ensure abundant (optimal) genetic diversity as much as possible with a limited number (minimum) of samples. To achieve this purpose, we should primarily estimate accurately the level of the total genetic diversity and its distribution (fine-scale spatial genetic structure) in a wild soybean population. Both theoretical studies and experiments have shown that the limitation of spatial distance and gene flow will have an impact on the mating between individuals and dispersal of seeds within the plant populations consequently resulting in a patch or cluster distribution of genotypes (Sokal and Waterberg, 1983; Furnier et al., 1987). Meanwhile, the heterogeneity at the fine-scale level within habitats can also lead to a cluster distribution of genetic variation within a population (Hamrick and Allard, 1972). For the purposes of *ex situ* and *in situ*

Translated from *Journal of Fudan University* (*Natural Science*), 2006, 45(3): 322–327 [译自: 复旦学报 (自然科学版)]

E-mail: brlu@fudan.edu.cn

conservation of plant populations, as well as the utilization of genetic resources in crop breeding, a better understanding for the fine scale spatial genetic structure of a natural plant population is very important for designing an effective sampling strategy (Mashall and Brown, 1975) to ensure the conservation and utilization of the genetic diversity of plant species to the fullest extent (He et al., 1999).

Most of the previous studies in the genetic conservation of wild soybeans have focused on the establishment of core germplasm resources, genetic diversity analyses, and the sampling strategy for wild soybean populations at a large geographical scale (Li et al., 1998; Qian et al., 1998; Choi et al., 1999; Dong et al., 2001; Li and Nelson, 2002; Cui et al., 2003). Accordingly, the experimental materials were collected from germplasm banks or wild soybean populations from different locations with a small number of seeds. For example, Li et al. (1995), and Zhou et al. (2004) used isozyme and the random applied polymorphic DNA (RAPD), respectively, to investigate different wild soybean populations, where they found high genetic diversity within the natural wild soybean populations, and thereby suggesting that the genetic diversity within wild soybean populations should also be conserved. Jin et al. (2003; 2006) investigated the genetic diversity and fine scale spatial genetic structure within wild soybean populations using the inter-simple sequence repeats (ISSR) markers, and proposed a spatial sampling strategy within wild soybean populations. However, the sample sizes in the previous studies were generally less than 100 individuals. Whether the sample size was sufficient to address the population genetic diversity is still in question. Also, the dominant markers used in former studies (Li and Nelson, 2002; Jin et al., 2003; Zhou et al. 2004; Jin et al., 2006) considerably restrained the full analysis of the genetic structures of the population at both the fine and large scales.

The simple sequence repeats (SSRs) are co-dominant markers with repeatable and consistent characteristics. These molecular markers have been widely used for genetic construct, genetic mapping, DNA fingerprinting analysis, varietal identification, genetic diversity, and evolution (Hymow, 1970; Rongwen et al., 1995; Hai et al., 2002). In this study, in order to design an appropriate conservation strategy for the effective preservation of the genetic resources of a wild soybean gene pool, we investigated a wild soybean population located in a natural reserve near the Yellow River estuary in Kenli of Shandong Province. Seventeen SSR primer pairs were used to estimate the genetic diversity of the population and its variation pattern against the changes in the sample size (sub-samples). In addition, the fine scale spatial genetic structure was studied within the population.

2 Materials and methods

2.1 Sample collection

The materials used in this study came from a natural reserve, a remote marsh near the Yellow River estuary located in Kenli of Shandong Province, with minor disturbance from human activities. There were many wild soybean populations and densely grown wild soybean individuals in the region. A wild soybean population occupying an area of about 10 000 m^2 was selected for this study. Individuals were sampled with a 2 m × 2 m grid pattern. Seeds from each individual were packed in a labeled paper seed bag, and a total of 2300 individuals were included in the samples. Within the population, 892 individuals from an area of 60 m × 60 m were selected for the analysis in this study.

2.2 Seeds germination and DNA preparation

To ensure obtaining a minimum of one seedling, two to three wild soybean seeds from each individual were germinated in an incubator at 28°C for seven days in a laboratory condition. The total genomic DNA was extracted from one of the seedlings by the CTAB (McGregor et al., 2002) protocol modified by Murray and Thompson (1980). DNA concentration was determined by a biophotometer of Epperndorf Company.

2.3 Simple sequence repeat amplification and electrophoresis

A total number of 60 SSR primer pairs, purchased from Shanghai SBS Genetech Co. Ltd., were independently screened with 20 DNA samples from wild soybean individuals to select primer pairs with sufficient polymorphism and were suitable for PCR amplification. Among these, 17 SSR primer pairs (Table 1) were selected for the amplification of all the wild soybean samples.

The polymerase chain reaction (PCR) reactions analysis were performed in a 20 μL volume container: 1 × buffer, 2.0 mmol/L $MgCl_2$, 1.6 mmol/L dNTPs, 0.1 μmol/L of each of the forward and reverse primers, 1 unit Tag DNA polymerase, and 50 ng template DNA. Reactions were run on PTC-10096v thermocycler (MJ Research Inc, Watertown, Mass). The PCR conditions were: an initial denaturation of 4 min at 95°C; followed by 35 cycles of 95°C for 30 s, 47°C for30 s, 72°C for 30 s and a final extension step at 72°C for 10 min. The PCR products were stored at 4°C and were electrophoresed on 6% polyacrylamide denaturing gel for about 1.5–3 h at a of stable voltage of 400 v. Then the products were visualized by a modified silver staining method (Song et al., 2001) and photo-documentations were taken for each gel. A pUC19 DNA/Msp I (Hpa II) Marker was used as a molecular standard weight.

2.4 Data analysis

2.4.1 Genetic diversity analysis

As SSR markers are co-dominant molecular markers, the amplified DNA polymorphic fragments by each SSR primer pair represent diverse alleles of the same locus. Accordingly, diverse alleles on one locus of the wild soybean were scored as double letters: homozygous genotypes are scored with

Table 1 SSR primer pairs used for the wild soybean genetic analysis in this study

Code of primer	5′-3′	3′-5′	No. of DNA band scored
Satt200	GCGATAAATGGTTAATGTAGATAA	GCGAAAGGACAGATAGAAAGAGA	2
Satt424	CAACCTGTATTCCACAAAAAATCTCACC	GCGCCCCAATTTGACTATAAATAAAAGT	2
Sat009	CACACGTATTGTCTTACCAC	CTCCGAGAAGCACGTA	2
Satt197	CACTGCTTTTTCCCCTCTCT	AAGATACCCCCAACATTATTTGTAA	3
SoyGPATR	GGAAGAAAGTATTGGTCTGT	AGGAGAGAGTGGAGAGATTA	2
Satt286	GCGGCGTTAATTTATGCCGGAAA	GCGTTTGGTCTAGAATAGTTCTCA	2
Satt005	TATCCTAGAGAAGAACTAAAAAA	GTCGATTAGGCTTGAAATA	5
Sat022	GCGGCCTTTTCTGACTGTTAA	GCGCAGTGACTAAAACTTACTAT	3
Satt045	TGGTTTCTACTTTCTATAATTATTT	ATGCCTCTCCCTCCT	3
Sat168	TGTGGATAAAAGAGCATTCAAAATG	GCGATCCTTGTTTATCTCAAAAAAGTGT	3
Satt434	GCGTTCCGATATACTATATAATCCTAAT	GCGGGGTTAGTCTTTTTATTTAACTTAA	3
Satt571	GGGTAGGGGTGGAATATAAG	GCGGGATCCGCGGATGGTCAAAG	4
Satt431	GCGTGGCACCCTTGATAAATAA	GCGCACGAAAGTTTTTCTGTAACA	3
Satt373	TCCGCGAGATAAATTCGTAAAAT	GGCCAGATACCCAAGTTGTACTTGT	3
Satt590	GCGCGCATTTTTTAAGTTAATGTTCT	GCGCGAGTTAGCGAATTATTTGTC	3
Satt022	GGGGGATCTGATTGTATTTTACCT	CGGGTTTCAAAAAACCATCCTTAC	3
Satt243	GCGCATTGCACATTAGGTTTTCTGTT	GCGGTAAGATCACGCCATTATTTAAGA	3

double same letters, e.g. AA, BB, CC; and heterozygous genotypes with two different letters which stand for diverse genotypes, AB, AC, BC. The data matrix of the SSR genotype was assembled to evaluate the genetic diversity in the wild soybean population, with the following four kinds of the most popular genetic diversity indexes, using PopGen32 (Ver.1.31) software (Yeh et al., 1999). The genetic diversity indexes were: (1) allele number (A); (2) expected heterozygosity (H_e), $H_e = 1 - \sum p_i^2$, where p_i stands for the frequency of the *ith* allele; (3) Shannon diversity index (I), $I = -\sum p_i ln p_i$, where p_i stands for the frequency of the *ith* allele (Nei, 1973); (4) frequency of polymorphic loci (P), $P =$ (polymorphic loci number/loci number) $\times 100\%$.

In order to reveal the correlation patterns of the genetic diversity in the population against the changes in the sample size or the number of samples (sub-samples), ten different sub-samples with various numbers of wild soybean individuals (i.e. 3, 5, 8, 12, 18, 27, 52, 78, 117, and 178) were randomly drawn from the population by computer simulation, and with the sampling for each sub-sample repeated for 30 times. In other words, we generated ten groups (sub-samples) of wild soybeans from this population with various sample sizes, with each sub-sample having 30 replicates. The above four genetic diversity indexes of these sub-samples were analyzed by the software PopGen ver. 32. Then, the mean values of the genetic diversity indexes of each group were calculated and transformed into the percentage of the indexes of the total samples (892 individuals). Finally, the regression model of the four diversity indexes, in percentage value with the change in the sample sizes, was computed by the statistic software SPSS. To achieve the best regression results, the *S*-curves were used for allele number (A), and *I*-curve for the rest indexes: H_e, I, P. The formula for the *S*-curve was $\mathrm{Ln}(Y) = b_0 + (b_1/t)$, and $Y = b_0 + (b_1/t)$ for *I*-curve, where the b_0, b_1 were the regression parameters, t was the sample size, and Y was the percentage of the genetic diversity indexes for the sub-samples from the total sample.

2.4.2 Fine scale spatial genetic structure analysis

The spatial autocorrelation coefficient r, which represented the correlation between the special distances and genetic distances of individual samples, was calculated to evaluate the fine scale spatial genetic structure of the wild soybean population. The calculation was performed by use of the computer software GeneAlEx V (Peakall and Smouse, 2001).

3 Results

3.1 Genetic diversity of the Kenli wild soybean population

A total of 48 alleles were detected in the wild soybean population with 892 individuals by the 17 selected primer pairs (loci). Each SSR primer pair identified 2–5 alleles, with an average of 2.88 (A). The results showed relatively high genetic diversity of the population with the average value of expected heterozygosis (*He*) being 0.431, Shannon diversity index (I) 0.699, and percentage of polymorphic loci (P) 100%.

Different sizes of the sub-samples (ten groups) were randomly drawn from the population and their genetic diversity was calculated by computer simulation (Table 2). The mean values and their standard deviation of A, H_e, I and P from the ten groups of sub-samples were also calculated. The regression results are listed in Fig. 1.

3.2 Spatial genetic structure of the Kenli wild soybean population

Spatial autocorrelation analysis on the Kenli wild soybean population revealed that genes in this wild soybean population significantly aggregated into cluster. Figure 3 shows the results of the spatial autocorrelation analysis using the

Table 2 The mean values and their standard deviation of A, H_e, I and P from the ten groups of sub-samples (from 3 to 176 individuals)

Sample size	A	H_e	I	P /%
3	1.657±0.169	0.281±0.073	0.409±0.105	61.57±16.20
5	1.925±0.191	0.341±0.073	0.518±0.109	77.26±12.77
8	2.176±0.153	0.397±0.048	0.616±0.076	87.84±6.61
12	2.288±0.144	0.395±0.040	0.624±0.060	90.98±4.74
18	2.441±0.121	0.410±0.029	0.624±0.038	90.98±5.02
27	2.510±0.093	0.415±0.024	0.663±0.035	94.90±4.22
52	2.667±0.093	0.425±0.019	0.684±0.026	97.65±2.88
78	2.747±0.059	0.422±0.020	0.685±0.028	99.02±2.19
117	2.800±0.060	0.423±0.016	0.689±0.021	99.41±1.76
176	2.835±0.041	0.428±0.012	0.695±0.695	100.00±0.00
892 (Total)	2.880	0.431	0.699	100.00

distance size class of 3 m. The correlation between genetic and spatial distances among individuals was found and the correlation was significantly positive in the distance interval of 18.8 m with an intercept at the distance of 24.5 m (Fig. 2).

4 Discussion

4.1 Gene diversity of the Kenli wild soybean population

The results from this study showed a relatively high level of genetic diversity in the Kenli wild soybean population. For a self-pollination species, the level of genetic diversity in the Kenli wild soybean population is unexpectedly high. For example, the value of the expected heterozygosity (H_e) in previous studies on soybean populations was reported to vary between 0.221–0.463 as estimated by isozyme and inter-simple sequence repeat (ISSR) (Li et al., 1995; Jin et al., 2003, 2006). The results obtained from this study may be attributed to the long history of the Kenli population, which is a natural wild soybean population, and thus abundant genetic polymorphisms were developed there. Therefore the conservation of the Kenli wild soybean population is necessary. Meanwhile, a large sample size (with 892 individuals) in this study may also be another reason for the high level of the genetic diversity. Accordingly, we believe with confidence that the Kenli wild soybean population in this study possessed abundant genetic variation in an adequate number of samples. Thus, this sampled population can be used for the estimation of the variation pattern of genetic diversity among its sub-populations with changes in sample sizes. This guarantees that the variation pattern obtained from this study with such a sample size can truly represent the genetic diversity of wild soybean populations.

Through computer simulated analysis of the randomly-drawn sub-populations, we can see that, with the increase of the sample number, the values of the genetic diversity of various sub-populations quickly increased within the number of 50 individuals. The rates of increase gradually slowed down and approached a stable level that was close to the value of the total population. As a result, for the indexes of H_e,

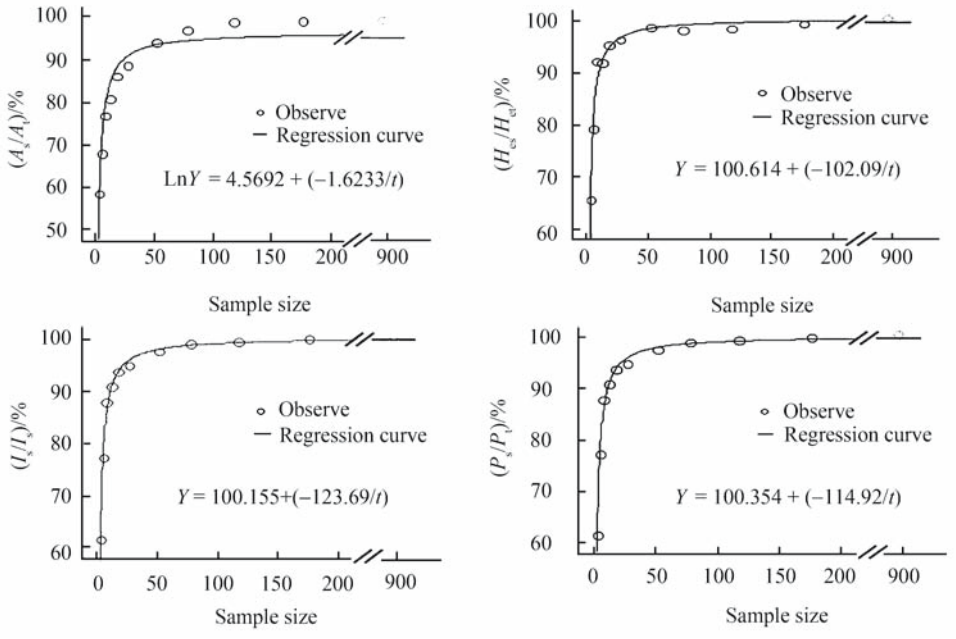

Fig. 1 Regression chart generated based on the % values of allele number (A), expected heterozygosity, (H_e), Shannon diversity index (I), and percentage pf polymorphic loci (P) with different sample sizes
Note: The X axis indicates the sample size, and the Y axis indicates average % values of the four genetic indexes against the genetic indexes of the total sample. The rings represent the average of observed values (%), and the line indicates the regression model of observed values (%) from groups with different sample sizes.

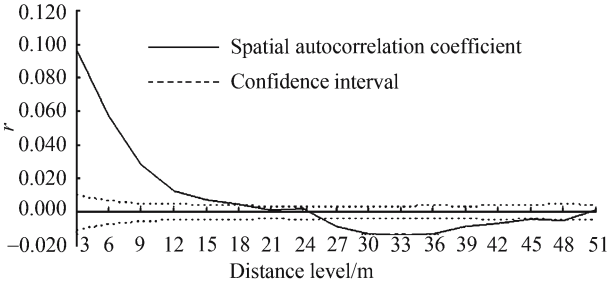

Fig. 2 The result of spatial autocorrelation analysis showing the size of genetic patches in the wild soybean population

I, and *P*, only 27 individuals reached 95% of the total genetic diversity of the population; and for the index of *A*, about 52 individuals reached the same level of genetic diversity as that of the total genetic diversity of the population. This indicates that the variation of the index *A* depends significantly on the sample size at a certain level. Based on the regression charts of the four genetic diversity indexes against the changes of sample sizes, the values of the genetic diversity indexes became very stable after the sample size exceeded 50 individuals, and the increasing rates of the regression curves remained almost unchanged. The regression analysis also indicated that those sub-populations with 27–52 randomly-drawn individuals can represent the total genetic diversity of the Kenli wild soybean population. This result is well accordant with the previous study (Jin et al., 2003).

4.2 Fine scale spatial genetic structure of the Kenli wild soybean population

As the result of the inbreeding system and the limited dispersal of seeds and pollen, in addition to the heterogeneity of microhabitats, the distribution of genes and genotypes within a population are commonly uneven or in clusters, instead of having a random distribution. This circumstance is defined as population genetic structuring at the fine scale. In this study, the spatial autocorrelation analysis showed that the correlation between the wild soybean individuals was positive and significant up to 18.8 m, indicating a high level of genetic structuring and a close genetic relationship (or the same genotype) among the individuals within patches (18 m in diameter) in the Kenli wild soybean population. The micro-environment in the Kenli wild soybean population is more or less homogeneous and therefore, the influence from the habitats that may cause the formation of a fine scale genetic structure (or patches) can be excluded. The wild soybean is a self-pollinating and annual herbal species with extremely limited pollen flow. Seed dispersion of wild soybeans is accomplished by pod explosion at maturity, which discharges the mature seeds to a limited distance (within 1–3 m per generation) (Li et al., 1997). Given the facts of its inbreeding characteristics and the nature of its seed dispersal, the long-distance gene flow between individuals within the wild soybean population is very occasional resulting in the formation of genetic patches within populations. Individuals within the genetic patches usually have little genetic differentiations. In other words, these individuals have significant and positive correlation; whereas individuals among the patches within the population have much more significant genetic differentiations. The conclusion serves as the scientific foundation for determining the spatial intervals for sampling wild soybean individuals within a population. As a consequence, we suggest the sampling of wild soybean individuals within a population with a spatial distance beyond 18 m to ensure the inclusion of different genotypes for *ex situ* conservation. This way, abundant genetic diversity of wild soybean germplasm can be captured with a limited number of samples thereby reducing the risk of statistical errors caused by the random sampling of a limited number of samples with low genetic diversity in the collected samples.

Acknowledgements This work was supported by the National Basic Research Program of China (No. 2006CB403305).

References

Choi I Y, Kang J H, Song H S, Kim N S (1999). Genetic diversity measured by simple sequence repeat variations among the wild soybean, Glycine soja, collected along the riverside of five major rivers in Korea. Genes and Genetic System, 74: 169–177

Cui Y H, Qiu L J, Chang R Z (2003). Plant core germplasm research headway. Plant Genetic Resource Journal, 4(3): 279–284 (in Chinese)

Dong Y S, Zhuang B C, Zhao L M, Sun H, He M Y (2001). The genetic diversity of annual wild soybeans grown in China. Theoretical and Applied of Genetics, 103: 98–103

Furnier G R, Knowles P, Ciyde M A, Dancik B P (1987). Effect of avian seed dispersal on the genetic structure of whitebark populations. Evolution, 41: 607–612

Hai L, Wang K J, Yang K (2002). Genetic diversity analysis of semi-wild-soybean genetic resourse with SSR marker. Northwest Plant Journal, 22(4): 751v–757 (in Chinese)

Hamrick J L, Allard R W (1972). Microgeographical variation in allozyme frequencies in Avena barbata. Proceedings of the National Academy of Sciences of USA, 69: 2100–2104

He T H, Yang J, Rao G Y (1999). Spatial autocorrelation analysis of plant population diversity. Botany Report, 16: 636–641 (in Chinese)

Hymow I Z (1970). On the domestication of the *soybean*. Econ Bot, 24: 408–421

Jin Y, Lu B R (2003). Genetic diversity sample strategy. Biology Diversity, 11(2): 155–161 (in Chinese)

Jin Y, He T H, Lu B R (2003). Fine scale genetic structure in a wild soybean (*Glycine soja*) population and the implications for conservation. New Phytologist, 159: 513–519

Jin Y, He T H, Lu B R (2006). Genetic spatial clustering: significant implications for conservation of wild soybean (*Glycine soja*: Fabaceae). Genetica, in press

Li J, Tao Y, Zheng S Z, Zhou J L (1995). Isozymatic differentiation in local population of *Glycine soja* sieb. et Zucc. Acta Bet Sin, 37(9): 669–676 (in Chinese)

Li J, Zheng S Z, Qian J, Ren W W (1997). Study on the seeds rain of wild soybean. Chinese Journal of Applied Ecology, 8(4): 372–376 (in Chinese)

Li J, Qian B, Zheng S Z (1998). Pilot study for genetic diversity of wild soybean seeds bank with isozyme. Applied Ecology Journal, 9(2): 145–149 (in Chinese)

Li Z L, Nelson R L (2002). RAPD marker diversity among cultivated and wild soybean accessions from four Chinese provinces. Crop Science, 42: 1337–1347

Mashall D R, Brown A D H (1975). Crop Genetic Recourses for Today and Tomorrow. Cambridge: Cambridge University Press

McGregor C E, Lamber C A, Greyling M M, Louw J H, Warnich L (2002). A comparative assessment of DNA fingerprinting techniques (RAPD, ISSR, AFLP and SSR) in tetraploid potato (*Solanum tuberosum* L.). Euphytica, 113: 135–144

Murray M G, Thompson W F (1980). Rapid isolation of high molecular weight plant DNA. Nucleic Acids Research, 8: 4321–4325

Nei M (1973). Analysis of gene diversity in subdivided populations. Proc Nat Acad Sci USA, 70(12): 3321–3323

Qian J, Cheng Y, Zheng S Z (1998). Genetic diversity between wild soybean population of various latitude. Fudan Journal, 37(2): 208 (in Chinese)

Peakall R, Smouse P E (2001). Genetic Analysis in Excel. Population genetic software for teaching and research. Canberra: Australian National University

Rongwen J, Akkaya M S, Bhagwat A A, Lavi U, Cregan P B (1995). The use microsatellite DNA markers for soybean geno-type identification. Theoretical and Applied of Genetics, 90: 43–48

Sokal R R, Waterberg D E (1983). A test of spatial autocorrelation analysis using an isolation-by-distance model. Genetics, 105: 219–237

Song Z P, Xu X, Wang B, Chen J K, Lu B R (2003). Genetic diversity in the northernmost *Oryza rufipogon* populations estimated by SSR markers. Theoretical and Applied of Genetics, 107: 1492–1499

Yeh F C, Yang R C, Boyle T (1999). Microsoft Window-based freeware for population genetic analysis (POPGENE32 ,Ver.1.31)

Zhou X F, Zhuang B C, Wang Y M, Zhao H K (2001). Study on differentiation within wild soybean population with RAPD. Songliang Academic Magazine (Nature Science Version), 4: 1–4 (in Chinese)