Ecology and Evolution

Open Access

Nucleotide diversity patterns of three divergent soybean populations: evidences for population-dependent linkage disequilibrium and taxonomic status of *Glycine gracilis*

Yunsheng Wang[1,2], Muhammad Qasim Shahid[3], Hongwen Huang[1,4] & Ying Wang[1]

[1]Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Science, Wuhan, Hubei 430074, China
[2]College of Environment and Life Science, Kaili University, Kaili, Guizhou 556011, China
[3]College of Agriculture, South China Agricultural University, Guangzhou, Guangdong 510642, China
[4]Key Laboratory of Plant Resources Conservation and Sustainable Utilization, Guangdong Provincial Key Laboratory of Applied Botany, South China Botanical Garden, Chinese Academy of Science, Guangzhou, Guangdong 510642, China

Keywords
Evolution, linkage disequilibrium, nucleotide polymorphism, population genetics, semi-wild soybean.

Correspondence
Hongwen Huang and Ying Wang, Key Laboratory of Plant Germplasm Enhancement and Specialty Agriculture, Wuhan Botanical Garden, Chinese Academy of Science, Wuhan, Hubei 430074, China.
Tel: +86 27 87510675;
Fax: +86 27 87510675;
E-mails: huanghw@mail.scbg.ac.cn; yingwang@wbgcas.cn

Funding Information
This work was supported by the National Natural Science Foundation Project (30670158), the Key Project from the Education Department of Guizhou Province [KY(2013)186], and Doctoral Project from Kaili University (BS201337).

Received: 7 March 2015; Revised: 16 May 2015; Accepted: 18 May 2015

Ecology and Evolution **2015; 5(18): 3969–3978**

doi: 10.1002/ece3.1550

Abstract

The level of linkage disequilibrium (LD) is a major factor to determine DNA polymorphism pattern of a population and to construct high-resolution maps useful in localizing and gene cloning of complicated traits. Here, we investigated LD level of three soybean populations with different genetic backgrounds and taxonomic status of *G. gracilis* by comparing the DNA polymorphism patterns of four high-diversity single-copy nuclear genes. A total of 152, 22, and 77 accessions of *G. soja*, *G. gracilis*, and *G. max* were observed. The results indicated that *G. max* retained only 75.3 (π) and 39% (θ) of the nucleotide polymorphism found in *G. soja*. Four gene loci evolved according to neutrality in both *G. max* and *G. gracilis* populations, and three gene loci evolved according to neutrality in *G. soja* population by Tajima's and Fu and Li's test. However, one gene locus deviated from neutrality by Fu and Li's test in the *G. soja* population. Further, medial level of LD (average $r^2 = 0.2426$) was found in intragene in *G. max* and *G. gracilis* populations, but unexpected low level of LD ($r^2 \leq 0.0539$) was found in *G. soja* population. Significant genetic differentiation was detected between *G. max* and *G. soja* populations and also between *G. max* and *G. gracilis* populations; however, nonsignificant genetic differentiation was found between *G. gracilis* and *G. soja* populations. The results suggest that LD level depends on genetic background of soybean population, and implicit that *G. gracilis* should be regarded as the variant of *G. soja*, not as an independent species.

Introduction

In recent years, study of linkage disequilibrium (LD) (nonrandom association of alleles) has attracted many scientists because of two reasons. First, whole-genome sequencing and high-throughput single nucleotide polymorphism (SNP) analysis enable rapid and convenient identification of haplotypes at different genetic loci. Second, in the presence of significant LD, it is possible to identify genomic regions that are associated with a particular trait of interest (e.g., disease resistance/susceptibility) and even to clone the correlative genes by a systematic and high-density genome scan of individuals from an existing population (Rafalski and Morgante 2004; Atwell et al. 2010).

There are many factors that affect the significance of LD, such as mating system, selection effect, population size, recombination rate, and population subdivision. Mating system is the most important factor (Rafalski and Morgante

2004). In general, selfing and outcrossing species have high and low level of LD, respectively. For examples, a study on *Arabidopsis thaliana*, a model species for molecular biology study and is a typical selfing species, showed that LD in most genome segments decayed within 25–50 kb (Nordborg et al. 2005). However, LD for the flowering time locus FRI in *Arabidopsis thaliana* decayed over 250 kb (Hagenblad and Nordborg 2002). Domesticated sorghum is also a typical selfing species, and the LD of sorghum in most genome segments decayed within 15 kb. Rice is an important economic crop, predominantly selfing species, and LD of which decayed over 100 kb (Garris et al. 2003). Maize is an outcrossing species, and LD decayed in only 200 bp along chromosome 1 (Tenaillon et al. 2002). But, the LD in maize was also found decaying up to over 2000 bp in populations with limited genetics background, or in the gene loci undergone the selection (Remington et al. 2001; Jung et al. 2004). For potato (*Solanum tuberosum*), an outcrossing species, its fragments show relatively fast decay of LD in the short range (r^2 = 0.208 at 1 kb) but slow decay afterward (r^2 = 0.137 at ~70 kb) (Simko et al. 2006). There are also some exceptions, for example, wild barley is an inbreeding species, with a selfing rate of ≈ 98%. However, the majority of wild barley loci, intralocus LD, decay rapidly, that is, at a rate similar to that observed in the outcrossing species, maize (Lin et al. 2002; Morrell et al. 2005).

Major crop in the world such as rice, maize, sorghum, and wheat were all domesticated 5000–10,000 years ago (Crawley et al. 2001). In the foremost stage of domestication procession, only few wild plants were selected as the progenitors of crop species, and this led to a so-called domesticated bottleneck (Tanksley and McCouch 1997). As a result, the genetic diversity of cultivated species were commonly lower than their progenitor wild species; in particular, rare genotypes in cultivated species were far less than that in the progenitor wild species (Schneider et al. 2001; Matsuoka et al. 2002; Zhang et al. 2002). Compared with wild species, the evolutionary history of cultivated species were far short. The evolutionary dynamics of cultivated species and wild species were also different, because cultivated species were undergone forceful artificial selection. However, natural selection or neutral drift was major driving dynamics for the evolution of wild species. The different genetic background between cultivated and wild species as described above should lead to different level of LD between them. This conclusion has been validated by some other scientists. For example, Liu and Burke (2006) revealed that LD of wild sunflower population declined to negligible level within 200 bp (r^2 = 0.10), while cultivated sunflower LD decayed over 1100 bp (r^2 = 0.10). LD of *Hordeum vulgare* extended to at least 212 kb in elite barley cultivars, but it was rapidly eroded in related inbreeding ancestral populations (Caldwell et al. 2006).

Cultivated soybean (*Glycine max*) and its progenitor wild soybean (*G. soja*) are two annual species of subgenus *Soja*. Cultivated soybean, domesticated in China before 3000–4000 years ago, now has been planted in all over the world and has become one of the most important crops in the world for providing vegetable oil and protein resources to human beings (In and Inder 1997; Shahid et al. 2009). Annual wild soybean (*G. soja*) is the progenitor of the cultivated soybean, and its geographical distribution is only limited to the middle and northern region of East Asia, including China, Korea, Japan, and far-east of Russia. There is no reproductive isolation between cultivated and wild soybean; however, morphological traits are quite different. Cultivated soybean is erect or semi-erect bushy plant, with height about 1 m, and its 100-seeds weight is above 15 g. Typical annual wild soybean is vine herbage plant with the stem length about 4–6 m, and 100-seeds weight is from 0.5 to 3 g. Besides the above two species, *G. gracilis* (semi-wild soybean) is a mid-type soybean between *G. max* and *G. soja* in plant morphology. Despite advances in molecular biology, *G. gracilis* taxonomic status has not clearly been defined. Some scientist regarded *G. gracilis* as an independent species, while some incorporated it into *G. max* (Hermann 1962) or *G. Soja* (Hymowitz and Singh 1987). *G. gracilis* is also an inbreeding plant like *G. max* and *G. soja* (Lu 2004). Linkage disequilibrium of cultivated and wild soybean populations has been studied by different scientists (Zhu et al. 2003; Hyten et al. 2006), but the geographical distribution of wild soybean samples was very limited. According to our knowledge, no one has investigated the LD of special annual semi-wild soybean. This study was aimed to detect the different levels of LD in three annual soybean populations (*G. soja*, *G. max*, and *G. gracilis*) from species-wide sampling by investigating the DNA polymorphism pattern of four gene loci and to discuss the taxonomic status of *G. gracilis*.

Materials and Methods

Plant material

A total of 152 accessions of wild soybean were collected on the basis of a species-wide sampling from the natural range of wild soybean. A total of 22 semi-wild soybean accessions were collected from 12 provinces of China. In addition, 71 accessions of cultivated soybean were selected from widely different regions of three countries (China, Korea, Japan) with longest history of planting soybean and six accessions of soybean were selected from United States of America, the largest soybean planting country (Table S1).

DNA extraction

For each soybean accession, the seeds were placed on the absorbent papers in petri dish and enough water was added for seed germination. Seedlings were grown at room temperature under natural photoperiod. Young leafs about 100 mg were gathered to extract DNA by CTAB protocol (Doyle and Doyle 1987).

Gene region selection

To compare the variant pattern of homologous DNA sequences of three soybean populations, four high-diversity gene loci were investigated in this study (Table 1). Three gene loci (Locus B, Locus C, and Locus D) were selected according to Van et al. (2005), and we found an intron splice site in a soybean EST sequence (Serial Number: TC229661) published in dataset bank (The Institute for Genomic Research; http://www.tigr.org/) by intron detected software from network site (http://www.sgn.cornell.edu/cgi-bin/tools/intron_detection/find_introns.pl). We designed primers according to the DNA sequence of this intron splice site, and 500-bp fragments were obtained for each genomic accession. Finally, we analyzed the sequencing results of these PCR products.

PCR sequencing and sequence assembly

PCR system was included: buffer 5.0 μL (10×), $MgCl_2$ 4.0 μL (25 mmol/L), dNTP 1 μL (100 μmol/L), primer 0.4 μL (100 μmol/L), Tagase 2.5 μ, and add ddH_2O up to 50 μL. The following PCR protocol was used: 94°C for 5 min, 35 cycles of 94°C for 30 sec, 54°C for 45 sec for Locus A and Locus C; 56°C for Locus B and Locus D and 72°C for 90 sec, and a final extension at 72°C for 10 min. For sequencing, PCR products were purified using PCR purify kit (Takara Biotechnology (Dalian) Co., Ltd. China). Then, PCR products were directly sequenced using 3730/ABI sequencer. Alignment of homologous sequences alignment was performed with clustalx program (Thompson et al. 1997) and edited by software bioedit v7.0.5 (Hall 1999). In total, we got 1839-bp homologous genomic sequence from four genes loci (Table 1).

Data analysis

DNA polymorphism, recombination, neutrality test, linkage disequilibrium, and population differentiation were analyzed using DnaSP 4.0 (Rozas et al. 2003).

For DNA polymorphism, six parameters including number of polymorphic/segregating sites (S), total number of mutations (Eta), number of haplotype (h), haplotype diversity (Hd), nucleotide polymorphism (θ), and

Table 1. List of Gene loci and primers used in the study.

Gene loci	Gene source	Function	Forward primer	Reverse primer	Alignment Length (bp)		
					Total	Coding	Noncoding
Locus A	TC229661 (TIGR)	Unknown	GCGTTGGAGATTGGAGATAA	TGGGACAGTAAGCAGTTGACC	411	210	211
Locus B	AF105221 (GenBank)	*Glycine max* glutamyl-tRNA reductase precursor (gtr1) gene, complete coding sequence	GCGACGCATTCAGTACACACTACAC	GCGGCCAAAGAAAGACAAGTAGATA	483	0	483
Locus C	AJ003246 (GenBank)	*Glycine max* mRNA for putative 2-hydroxydihydrodaidzein reductase	GCGGGCAAAAAGGAAGAAAT	GCGGGGAAAAGGTGAAAATTA	516	69	437
Locus D	J02746 (GenBank)	*Glycine max* SbPRP1 gene encoding a proline-rich protein, complete coding sequence.	GCGGGGTGTTCGAGGTTTCTAAT	GCGATGCGTTGGAATTTCAGGATA	428	0	428

nucleotide diversity (π) were estimated as the measurement of DNA polymorphism within populations. *Hd* was based on these estimated haplotype frequencies as $\hat{H} = (n/(n-1)) \cdot \left[1 - \sum_{i=1}^{k} p_i^2\right]$, for k haplotypes each with frequency p_i and total chromosome count n (Nei 1987). θ showed the number of segregating sites (Watterson 1975), and π was calculated on the basis of pairwise differences between sequences of samples (Tajima 1983). Recombination parameter (*Rm*) (minimum number of intragenic recombination events) was calculated using the four-gamete test (Hudson and Kaplan 1985). Neutrality test was executed by two methods, Tajima's *D* test; Fu and Li's *D** and *F** test: Tajima's *D* test statistic was proposed by Tajima (1989) to test the neutral theory of molecular evolution (Kimura 1983). This test is based on the fact that under the neutral model estimates of the number of polymorphic/segregating sites and of the average number of nucleotide differences are correlated. *D** and *F** of Fu and Li (1993) is also to test the neutral theory of molecular evolution (Kimura 1983). However, the *D** test is based on the differences between the total number of mutations and the number of singletons (mutations appearing only once among the sequences). The *F** test is based on the differences between the average number of nucleotide differences between pairs of sequences and the number of singletons. For both test, 10,000 simulations were calculated to test the hypothesis that mutations in the gene are selectively neutral (Kimura 1983).

We estimated the degree of linkage disequilibrium (or nonrandom association between variants of different polymorphic sites), only taking into account the informative site in this analysis with the parameter R^2 (Hill and Robertson 1968). For analysis, gametes with the most or the least common variants were considered in the coupling phase (Langley et al. 1974). Both the two-tailed Fisher's exact test and the chi-square test were computed to determine whether the associations between polymorphic sites are, or are not, significant. The decay of LD with physical distance was estimated using a nonlinear regression analysis of LD between polymorphic sites versus the distance between sites in base pairs (Remington et al. 2001). Decay of LD between polymorphic sites with physical distance between sites in base pairs (Remington et al. 2001) was estimated from a logarithmic trend line fit to the data (Hyten et al. 2007) using software SPSS 14.0 (SPSS 14.0, Chicago, IL).

For interpopulations differentiation, the population differentiation statistic F_{ST} was estimated using the method of Hudson et al. (1992). Note that under a symmetric island model of migration, $F_{ST} = 1/(1 + 4Nm)$, where m is the migration rate, and the significance of population differentiation was evaluated using the S_{nn} test statistic (Hudson 2000).

Results

DNA polymorphism pattern in *G. soja*, *G. max*, and *G. gracilis*

Six parameters including number of polymorphic sites (*S*), total number of mutations (*Eta*), number of haplotype (*h*), haplotype diversity (*Hd*), nucleotide diversity (π), and theta (per site) from Eta (θ) were used to measure the DNA polymorphism of four homologous DNA sequences in three soybean populations. Single and combined data of four genes showed the highest DNA polymorphism in *G. soja*, while the lowest polymorphism was found in *G. max*. Haplotype diversity (*Hd*), nucleotide diversity (π), and theta (per site) from eta (θ) in *G. max* were 0.618, 0.00520, and 0.00439, respectively. *G. max* retained only 80.0% of 0.772 (*Hd*), 75.3% of 0.00691 (π), and 39.0% (θ) of 0.01126 that found in *G. soja*, respectively. *G. max* also retained 82.5% of 0.749 (*Hd*), 83.7% of 0.00629 (π), and 64.7% of 0.00679 (θ) that found in *G. gracilis*, respectively (Table 2). Our analysis indicated that *Rm* of Locus A, Locus B, and Locus D is similar in three soybean populations (1 and 0). However, the *Rm* of Locus C (8) and Locus D (2) in *G. soja* was higher than that of *G. gracilis* (2 and 0) and *G. max* (2 and 0). Four gene loci behaved very differently in the same soybean population (Table 2).

Two methods (Tajima's *D* test, Fu and Li's *D** and *F** test) were used to execute neutrality test. The results revealed that the D value of all four gene loci departed nonsignificantly from zero in *G. max* and *G. gracilis*. These results showed that four gene loci followed neutral evolution in these two populations. Loci B, C, and D also followed neutral evolution by Tajima's *D* test, and Fu and Li's *D** and *F** test in *G. soja* population. However, Locus A followed neutral evolution by Tajima' test but deviated from neutral evolution by Fu and Li's test, indicating that Locus A had undergone weak selection effect.

Linkage disequilibrium analysis

Our data indicated 372 pairwise comparisons of all four gene loci in *G. max* population (Table 3). The LD of 162 pairwise comparisons (43.5%) of segregating sites was significant based on Fisher's exact test, and LD of 175 pairwise comparisons (47.0%) was significant by chi-square test. The average r^2 value of total 372 pairwise comparisons in *G. max* population was 0.2426 with the minimum and maximum values of 0.0010 (Locus A) and 0.4095 (Locus B), respectively. A total of 531 pairwise comparisons from all four gene loci were found in *G. gracilis* population, and LD of 125 (23.5%) and 148 (27.9%) pairwise comparisons were found significant by Fisher's

Table 2. DNA Polymorphism of four gene loci in three soybean populations.

DNA polymorphism	*S*	*Eta*	*h*	*Hd*	π ($\times 10^{-2}$)	θ ($\times 10^{-2}$)	*Rm*	Tajima's D test	Fu and Li's D* test	Fu and Li's F* test *
Gm Locus A	2	2	4	0.551	0.150	0.099	1	ns	ns	ns
Locus B	5	5	4	0.585	0.031	0.212	0	ns	ns	ns
Locus C	27	27	9	0.647	1.147	1.121	2	ns	ns	ns
Locus D	5	5	5	0.689	0.396	0.238	0	ns	ns	ns
Total	39	39	22	0.618	0.520	0.439				
Gg Locus A	6	6	7	0.762	0.358	0.401	1	ns	ns	ns
Locus B	5	5	5	0.597	0.197	0.286	0	ns	ns	ns
Locus C	32	32	10	0.879	1.541	1.752	2	ns	ns	ns
Locus D	5	5	4	0.758	0.383	0.320	0	ns	ns	ns
Total	48	48	26	0.749	0.629	0.679				
Gs Locus A	15	15	14	0.828	0.376	0.654	1		$P < 0.02$**	$P < 0.02$**
Locus B	9	9	10	0.540	0.202	0.422	0	ns	ns	ns
Locus C	62	66	41	0.935	1.592	2.632	8	ns	ns	ns
Locus D	13	13	19	0.785	0.467	0.588	2	ns	ns	ns
Total	99	103	84	0.772	0.691	1.126				

Gm, *Glycine max*; *Gg*, *Glycine gracilis*; *Gs*, *Glycine soja; S*, number of polymorphic (segregating) sites; *Eta*, total number of mutations; *h*, number of haplotype; *Hd*, haplotype diversity; π, nucleotide diversity; θ, theta (per site) from *Eta*; *Rm*, minimum number of recombination events; ns, non-significant.

**Significantly different from zero at $0.05 < P < 0.01$.

Table 3. The level of LD of four gene loci in three soybean populations.

LD	Locus A	Locus B	Locus C	Locus D	Total
Gm population					
Number of pairwise comparisons	1	10	351	10	372
Fisher's exact test (number of significant)	0	6	150	6	162
chi-square test (number of significant)	0	6	163	6	175
r^2	0.0010	0.4095	0.2394	0.2099	0.2426
Gg population					
Number of pairwise comparisons	15	10	496	10	531
Fisher's exact test (number of significant)	1	1	120	3	125
chi-square test (number of significant)	2	3	140	3	148
r^2	0.0805	0.2075	0.2034	0.3662	0.2030
Gs population					
Number of pairwise comparisons	105	36	1653	78	1872
Fisher's exact test (number of significant)	10	1	259	14	284
chi-square test (number of significant)	10	4	297	16	327
r^2	0.0472	0.0348	0.0539	0.0431	0.0527

LD, linkage disequilibrium; *Gm*, *Glycine max*; *Gg*, *Glycine gracilis*; *Gs*, *Glycine soja*.

exact test and chi-square test, respectively. The average r^2 value of total 531 pairwise comparisons in *G. gracilis* population was 0.2030. The level of linkage disequilibrium did not decay in intralocus (<500 bp) (Fig. 1A and B) of both *G. max* and *G. gracilis* populations; however, low level of linkage disequilibrium was found. In *G. soja* population(s), 1872 pairwise comparisons were found and LD of 282 (15.2%) and 327 (17.5%) pairwise comparisons were significant based on Fisher's exact test and chi-square test, respectively. The average r^2 values were lower than 0.1 at all loci in *G. soja* (Fig. 1C), and minimum (0.0348) and maximum (0.0539) r^2 values were found at Locus B and Locus C, respectively. It is concluded from above analysis that the level of linkage disequilibrium not only depended on gene loci but also had a strong relationship with the genetic background of soybean populations.

Interpopulation differentiation

The differentiation index (*Fst*) and significance test (*Snn*) were executed to investigate the extent of genetic differen-

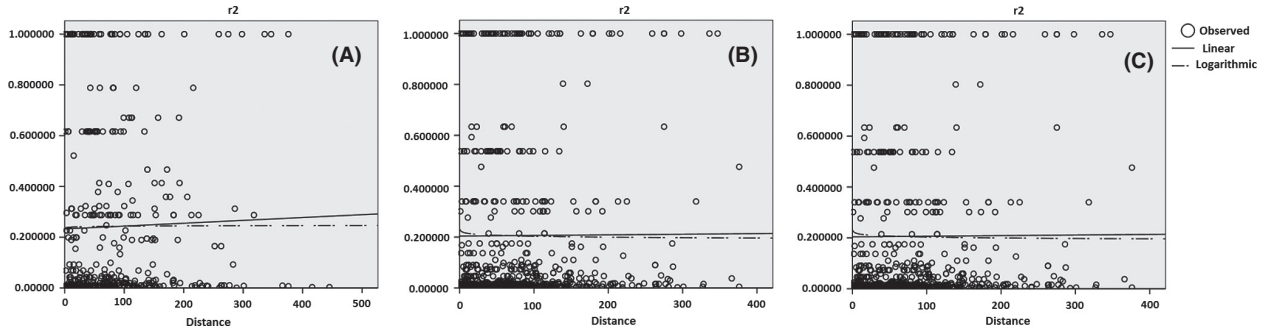

Figure 1. Linkage disequilibrium (r^2) versus distance within loci in three divergent soybean populations. The line is a logarithmic trend line fit to the data by SPSS 14.0. (A) A total of 372 pairwise estimates of r^2 were calculated from four loci across the genome of *G. max*; (B) A total of 531 pairwise estimates of r^2 were calculated from four loci across the genome of *G. gracilis*; (C) A total of 1872 pairwise estimates of r^2 were calculated from four loci across the genome of *G. Soja*.

tiation between four loci of three soybean populations (Table 4). The data of four gene loci showed significant genetic differentiation between *G. max* and *G. soja* populations, as well as between *G. max* and *G. gracilis* populations. However, there was nonsignificant genetic differentiation between *G. gracilis* and *G. soja* populations. Besides Locus A, other three loci and combined data indicated that the extent of genetic differentiation between *G. max* and *G. soja* populations was more than that between *G. max* and *G. gracilis* populations.

Discussion

DNA polymorphism in *G. soja*, *G. gracilis*, and *G. max* populations

Homologous DNA polymorphism pattern is an important indicator of population genetic structure and also an important premise to investigate the relationship between the population genetic structure and evolutionary factors (Gaut and Clegg 1993). Our data showed that the nucleotide polymorphism in *G. max* (landrace) population was 75.3% (π), where 39.0% (θ) of polymorphism existed in *G. soja*. In contrast to these results, Hyten et al. (2006) reported that the landraces retained only 66% (π), where 49% (θ) of nucleotide polymorphism existed in *G. soja*. We speculated that there might be two reasons for this difference: First, the gene loci selected in these two studies were different from each other. Second, the samples selected in two studies were different from each other. A large number of gene loci were used in Hyten's study, but geographical locations or number of samples was limited, especially for *G. soja* (Hyten et al. 2006). For example, China is the major native geographical area of *G. soja*, and studies have shown that significant genetic differentiation happened among different *G. soja* populations

Table 4. Genetic differentiation between three soybean populations.

Genetic differentiation	Population 1	Population 2	*Fst*	*Snn*
Locus A	*Gg* population	*Gs* population	−0.00928	0.78169 (ns)
	Gg population	*Gm* population	0.23068	0.79003***
	Gs population	*Gm* population	0.17590	0.65614***
Locus B	*Gg* population	*Gs* population	−0.00687	0.77446 (ns)
	Gg population	*Gm* population	0.09094	0.70440**
	Gs population	*Gm* population	0.17450	0.71704***
Locus C	*Gg* population	*Gs* population	−0.00443	0.77085 (ns)
	Gg population	*Gm* population	0.00647	0.70862**
	Gs population	*Gm* population	0.06424	0.69512***
Locus D	*Gg* population	*Gs* population	0.03847	0.78731 (ns)
	Gg population	*Gm* population	0.03863	0.73287***
	Gs population	*Gm* population	0.11660	0.74881***
Combination data	*Gg* population	*Gs* population	0.00643	0.78277 (ns)
	Gg population	*Gm* population	0.04622	0.82595***
	Gs population	*Gm* population	0.11953	0.90146***

Gm, *Glycine max*; *Gg*, *Glycine gracilis*; *Gs*, *Glycine soja;* ns, nonsignificant.
Fst (Hudson et al. 1992), *Snn* (Hudson 2000). *$0.01 < P < 0.05$; **$0.001 < P < 0.01$; ***$P < 0.001$.

(Kuroda et al. 2006), so large number of samples should be used to investigate the DNA polymorphism of whole *G. soja* species. However, the number of *G. soja* samples from China was very limited in Hyten's study. Our sampling strategy based on species distribution has surmounted this limitation, but the number of gene loci was low in our study.

Average nucleotide polymorphism (θ) in *G. max* (landrace) and *G. soja* was 0.00115 and 0.00235, respectively (Hyten et al. 2006). Except Locus A, the nucleotide polymorphism of other three loci (Locus B, Locus C, and Locus D) was higher than the average value in *G. max* and *G. soja*. For Locus A, nucleotide polymorphism was higher than the average value in *G. max* and lower than the average value in *G. soja*, and the results depicted that Locus A may undergo the effect of adaptation selection in *G. soja* population. These results were consistent with the neutrality test. High-diversity genes represent an important class of loci in organism genomes, as elevated levels of nucleotide variation are a key component of the molecular signatures for balancing selection or local adaptation. Earlier study has shown that most of the high-diversity loci in *Arabidopsis thaliana* departed from neutral evolution and undergone balance selection (Cork and Purugganan 2005). However, Locus A in cultivated and semi-wild populations and other three loci in all three soybean populations exhibited neutral evolution and suggested that balancing selection or local adaptation may not be the major evolution dynamics of high-diversity genes in soybean populations. To our knowledge, this is the first report on DNA polymorphism in *G. gracilis*, and it revealed that the level of DNA polymorphism in *G. gracilis* population was higher than *G. max* and *lower than G. soja*.

Linkage disequilibrium in *G. soja*, *G. gracilis*, and *G. max* populations

Mating system, recombination rate, population subdivision, population size, effect of selection, demographic events, etc. all could affect the level of linkage disequilibrium (Rafalski and Morgante 2004). As *G. soja, G. gracilis, and G. max* are all inbreeding species (Lu 2004), high level of linkage disequilibrium is expected. The results showed that the average r^2 value in *G. max* population of four gene loci data was 0.2426, and LD did not decay in short distance (<500 bp). These results are in agreement with the conclusion of Zhu et al. (2003) who reported that average r^2 value was approximately 0.2 within short distance (>500). However, unexpectedly low level of LD in *G. soja* population was found, even in very short distance, and the average r^2 value in *G. soja* was lower than 0.1 and declines slowly in short distance (<500 bp). In conclusion, population-dependent linkage disequilibrium existed in soybean population that led to different level of LD between *G. soja* and *G. max* populations. We speculated that there might be some reasons for these results: (1) population size: large population size leading to the decay of LD level; in this study, the number of *G. soja* samples were approximately twofold higher than that of *G. max*. (2) Genetic drift effect: Cultivated soybean (*G. max*) was domesticated from *G. soja;* in the process of domestication, few genotypes were selected as the progenitor of cultivated soybean and genetic bottleneck came by this way. In addition, very little seeds were selected to breed generation by generation, as humans eat majority of seeds as food. The genetic drift and selective breeding would lead to an increase in the level of LD in cultivated soybean. (3) Mating system: *G. max* and *G. soja* are inbreeding species (Lu 2004), and the outcrossing rate of *G. max* was lower than 1.0% (Bai and Gai 2003). However, the outcrossing rate in native *G. soja* could be up to 13.0% (Fujita et al. 1997), and this "high" outcrossing rate may lead to low level of LD in *G. soja*. (4) Demographic history: The evolutionary history of *G. soja* is far longer than *G. max*, and genetic differentiation among *G. soja* populations is stronger than *G. max* populations. Therefore, the genetic foundation of *G. soja* is wider than *G. max* based on species-wide sampling.

Studies on the cultivated and wild barley, cultivated and wild sunflower also have shown that the levels of LD in cultivated crops were higher than their wild progenitor species (Caldwell et al. 2006; Liu and Burke 2006), as the mating system of cultivated barley, sunflower, and soybean was same as of wild progenitor species; therefore, we proposed that the genetic background is the major factor for different levels of LD between cultivated species and their wild progenitor species. However, the genetic background consisted of multifactors, so we could not measure the exact level of LD affected by a given factor.

Genetic differentiation among three soybean populations and taxonomic status of *G. gracilis*

The genus *Glycine* contains two subgenera: subgenus *Glycine* and subgenus *Soja*. Subgenus *Glycine* comprises at least 16 perennial species that largely distributed in Australia and its adjunctive island, extended northward to east–south of Chinese sea border and Japanese south island (Lu 2004). Generally, subgenus *Soja* could be divided into two annual species: cultivated soybean (*G. max*) and its progenitor wild soybean (*G. soja*). *G. soja* has a vining or trailing growth habit, while *G. max* has erect and busy growth habit. There are few mutant individuals named as semi-wild soybean (*G. gracilis*) with medial plant morphology

between typical *G. max* and typical *G. soja*. Some scientist regarded the semi-wild soybean as a new species, *G. gracilis* (Skvortzow 1927). Some scholars revealed that it is wrong to regard the semi-wild soybean as a new species (Hermann 1962) and should be merged it into *G. max,* while other scholars reported that semi-wild soybean should be merged into *G. soja* (Hymowitz and Singh 1987; Doyle et al. 2003). Some scientists proposed that these three kinds of soybean should be regarded as one species (Hui et al. 1996) or one biology species (Lu 2004). Our data showed nonsignificant genetic differentiation between *G. soja* and *G. gracilis* populations. However, significant genetic differentiations were found between *G. gracilis* and *G. max* and also between *G. soja* and *G. max.* In spite of common characteristics, significant differentiations happened among different populations in same species, but plant morphology, physiological traits, and growth habit between *G. soja* and *G. max* were different from each other. Moreover, both *G. soja* and *G. max* can reproduce steady and heritable populations. Therefore, we speculated that it is not correct to regard semi-wild soybean as an independent species. We strongly suggest that semi-wild soybean possesses the wild characteristics and is the plant morphological mutant of typical *G. soja.* These results are in consistent with earlier study, which also reported semi-wild soybean as a variant of *G. soja* based on genome re-sequencing of semi-wild soybean (Qiu et al. 2014).

Conclusion

This study revealed that population-dependent linkage disequilibrium existed in soybean populations that led to different level of LD among *G. soja*, *G. gracilis,* and *G. max* populations. We also offered the viewpoints of the taxonomic status of *G. gracilis* by comparing the nucleotide diversity patterns of three divergent soybean populations. This study will help us to understand the taxonomic status of semi-wild soybean and genetic differentiation between *G. soja* and *G. max.* Based on the above genetic differentiation analysis, we proposed that subgenus *soja* contains two different species: wild soybean (*G. soja*) and cultivated soybean (*G. max*). Semi-wild soybean (*G. gracilis*) could be merged into *G. soja* and regarded it as the variant of *G. soja*.

Acknowledgments

We are very thankful to Dr Lijuan Qiu, Dr Xinan Zhou, Dr Xianghua Li, and ADUSA for providing soybean seeds. This work was supported by the National Natural Science Foundation Project (30670158), the Key Project from the Education Department of Guizhou Province [KY(2013)186], and Doctoral Project from Kaili University (BS201337).

Conflict of Interest

The authors declare no conflict of interest.

References

Atwell, S., Y. S. Huang, B. J. Vilhja'lmsson, G. Willems, Y. Li, D. Meng, et al. 2010. Genome-wide association study of 107 phenotypes in *Arabidopsis thaliana* inbred lines. Nature 465:627–631.

Bai, Y. N., and J. Y. Gai. 2003. Development of soybean cytoplasmic nuclear male sterile line NJ CMSA and restorability of its male fertility. Sci. Agric. Sin. 36: 740–745.

Caldwell, K. S., J. Russell, P. Langridge, and W. Powell. 2006. Extreme population-dependent linkage disequilibrium detected in an inbreeding plant species, *Hordeum vulgare*. Genetics 172:557–567.

Cork, J. M., and M. D. Purugganan. 2005. High-diversity genes in the Arabidopsis genome. Genetics 170:1897–1911.

Crawley, M. J., S. L. Brown, R. S. Hails, D. D. Kohn, and M. Rees. 2001. Biotechnology-transgenic crops in natural habitats. Nature 409:682–683.

Doyle, J. J., and J. L. Doyle. 1987. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19:11–15.

Doyle, J. J., J. L. Doyle, J. T. Rauscher, and A. H. Brown. 2003. Diploid and polyploid reticulate evolution throughout the history of the perennial soybeans (*Glycine* subgenus *Glycine*). New Phytol. 161:121–132.

Fu, Y. X., and W. H. Li. 1993. Statistical tests of neutrality of mutations. Genetics 133:693–709.

Fujita, R., M. Ohara, K. Okazaki, and Y. Shimamoto. 1997. The extent of natural cross-pollination in wild soybean (*Glycine soja*). J. Hered. 88:124–128.

Garris, A. J., S. R. McCouch, and S. Kresovich. 2003. Population structure and its effect on haplotype diversity and linkage disequilibrium surrounding the *xa5* locus of rice (*Oryza sativa* L.). Genetics 165:759–769.

Gaut, B. S., and M. T. Clegg. 1993. Molecular evolution of the Adhl locus in the genus *Zea*. Proc. Natl Acad. Sci. USA 90:5095–5099.

Hagenblad, J., and M. Nordborg. 2002. Sequence variation and haplotype structure surrounding the flowering time locus FRI in *Arabidopsis thaliana*. Genetics 161:289–298.

Hall, T. A. 1999. BIOEDIT: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acids Symp. Ser. 41:95–98.

Hermann, F. J. 1962. A revision of the genus Glycine and its immediate allies. United States Department of Agriculture Technical Bulletin 1268: 1–79.

Hill, W. G., and A. Robertson. 1968. Linkage disequilibrium in finite populations. Theor. Appl. Genet. 38:226–231.

Hudson, R. R. 2000. A new statistic for detecting genetic differentiation. Genetics 155:2011–2014.

Hudson, R. R., and N. L. Kaplan. 1985. Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics 111:147–164.

Hudson, R. R., M. Slatkin, and W. P. Maddison. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics 132:583–589.

Hui, D. W., B. C. Zhuang, and S. Y. Chen. 1996. Phylogeny of genus *Glycine* reconstructed by RAPD fingerprinting. Acta Genet. Sin. 23:460–468.

Hymowitz, T., and R. J. Singh. 1987. Taxonomy and speciation. Pp. 23–48 *in* J. R. Wilcox, ed. Soybeans, improvement, production, and uses. American Society of Agronomy, Madison, WI.

Hyten, D. L., Q. Song, Y. Zhu, I. Y. Choi, L. N. Randall, J. M. Costa, et al. 2006. Impacts of genetic bottlenecks on soybean genome diversity. Proc. Natl Acad. Sci. USA 105:16667–16671.

Hyten, D. L., I. Y. Choi, Q. Song, R. C. Shoemaker, R. L. Nelson, J. M. Costa, et al. 2007. Highly variable patterns of linkage disequilibrium in multiple soybean populations. Genetics 175:1937–1944.

In, F., and B. Inder. 1997. Long-run relationships between world vegetable oil prices. Aust. J. Agric. Resour. Econ. 41:445–470.

Jung, M., A. Ching, D. Bhattramakki, M. Dolan, S. Tingey, M. Morgante, et al. 2004. Linkage disequilibrium and sequence diversity in a 500-kbp region around the adh1 locus in elite maize germplasm. Theor. Appl. Genet. 109:681–689.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, MA.

Kuroda, Y., A. Kaga, N. Tomooka, and A. Vaughan. 2006. Population genetic structure of Japanese wild soybean (*Glycine soja*) based on microsatellite variation. Mol. Ecol. 15:959–974.

Langley, C. H., Y. N. Tobari, and K. Kojima. 1974. Linkage disequilibrium in natural populations of *Drosophila melanogaster*. Genetics 78:921–936.

Lin, J. Z., P. L. Morrell, and M. T. Clegg. 2002. The influence of linkage and inbreeding on patterns of nucleotide sequence diversity at duplicate alcohol dehydrogenase loci in wild barley (*Hordeum vulgare* ssp. *spontaneum*). Genetics 162:2007–2015.

Liu, A., and J. M. Burke. 2006. Patterns of nucleotide diversity in wild and cultivated sunflower. Genetics 173:321–330.

Lu, B. R. 2004. Conserving biodiversity of soybean gene pool in the biotechnology era. Plant Species Biol. 19:115–125.

Matsuoka, Y., S. E. Mitchell, S. Kresovich, M. Goodman, and J. Doebley. 2002. Microsatellites in Zea – variability, patterns of mutations, and use for evolutionary studies. Theor. Appl. Genet. 104:436–450.

Morrell, P. L., D. M. Toleno, K. E. Lundy, and M. T. Clegg. 2005. Low levels of linkage disequilibrium in wild barley (*Hordeum vulgare* ssp. *spontaneum*) despite high rates of self-fertilization. Proc. Natl Acad. Sci. USA 102:2442–2447.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York, NY.

Nordborg, M., T. T. Hu, Y. Ishino, J. Jhaveri, C. Toomajian, H. Zheng, et al. 2005. The pattern of polymorphism in *Arabidopsis thaliana*. PLoS Biol. 3:1289–1299.

Qiu, J., Y. Wang, S. Wu, Y. Y. Wang, C. Y. Ye, X. Bai, et al. 2014. Genome re-sequencing of semi-wild soybean reveals a complex *Soja* population structure and deep introgression. PLoS ONE 9:e108479.

Rafalski, A., and M. Morgante. 2004. Corn and humans: recombination and linkage disequilibrium in two genomes of similar size. Trends Genet. 20:103–111.

Remington, D. L., J. M. Thornsberry, Y. Matsuoka, L. M. Wilson, S. R. Whitt, J. Doebley, et al. 2001. Structure of linkage disequilibrium and phenotypic associations in the maize genome. Proc. Natl Acad. Sci. USA 98:11479–11484.

Rozas, J., J. C. Snchez-DelBarrio, X. Messegyer, and R. Rozas. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19:2496–2497.

Schneider, K., B. Weisshaar, D. C. Borchardt, and F. Salamini. 2001. SNP frequency and allelic haplotype structure of Beta vulgaris expressed genes. Mol. Breed. 8:63–74.

Shahid, M. Q., M. F. Saleem, H. Z. Khan, and S. A. Anjum. 2009. Performance of soybean (*Glycine max* L.) under different phosphorus levels and inoculation. Pak. J. Agric. Sci. 46:1–5.

Simko, I., K. G. Haynes, and R. W. Jones. 2006. Assessment of linkage disequilibrium in potato genome with single nucleotide polymorphism markers. Genetics 173:2237–2245.

Skvortzow, B. V. 1927. The soybean - wild and cultivated in Eastern Asia. Proc. Manchurian Res. Soc. Publ. Ser. A. Natural History. History Sect. No 22:1–8.

Tajima, F. 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics 105:437–460.

Tajima, F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585–595.

Tanksley, S. D., and S. R. McCouch. 1997. Seed banks and molecular maps: unlocking genetic potential from the wild. Science 277:1063–1066.

Tenaillon, M. I., M. C. Sawkins, L. K. Anderson, S. M. Stack, J. Doebley, and B. S. Gaut. 2002. Patterns of diversity and recombination along chromosome 1 of maize (Zea mays ssp. mays L.). Genetics 162:1401–1413.

Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 24:4876–4882.

Van, K., E. Y. Hwang, M. Y. Kim, H. J. Park, S. H. Lee, and P. B. Cregan. 2005. Discovery of SNPs in soybean genotypes frequently used as the parents of mapping populations in the United States and Korea. J. Hered. 96:529–535.

Watterson, G. A. 1975. On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. 7:188–193.

Zhang, L., A. S. Peek, D. Dunams, and B. S. Gaut. 2002. Population genetics of duplicated disease defense genes, hm1 and hm2 in maize (*Zea mays* ssp. mays L.) and its wild ancestor (*Zea mays* ssp. parviglumis). Genetics 162:851–860.

Zhu, Y. L., Q. J. Song, D. L. Hyten, C. P. Van Tassell, L. K. Matukumalli, D. R. Grimm, et al. 2003. Single-nucleotide polymorphisms in soybean. Genetics 163:1123–1134.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Table S1. Cultivated and wild soybean accessions and their origin.