INTERNATIONAL JOURNAL OF AGRICULTURE & BIOLOGY
ISSN Print: 1560–8530; ISSN Online: 1814–9596
17–0702/2018/20–4–803–810
DOI: 10.17957/IJAB/15.0567
http://www.fspublishers.org

Full Length Article

SNP-SNP Interaction Analysis on Soybean Seed Size in Multiple Years

Dayong Zhang[1†], Liqing Wu[1†], Huidong Qi[1], Xinrui Mao[1], Yan Shi[1], Zhongyu Wang[1], Hongwei Jiang[2], Qianchun Gong[1], Xiaoxia Wu[1], Dawei Xin[1*], Qingshan Chen[1*] and Zhaoming Qi[1*]

[1]*College of Agriculture, Northeast Agricultural University, Harbin 150030, Heilongjiang, People's Republic of China*
[2]*The Crop Research and Breeding Center of Land-Reclamation of Heilongjiang Province, Harbin 150090, Heilongjiang, People's Republic of China*
[*]For correspondence: qizhaoming1860@126.com; qshchen@126.com; xdawei@163.com
[†]These authors contributed equally to this work

Abstract

Seed size traits of soybean are important for seed yield. In this research, multifactor dimensionality reduction method (MDR) and the soybean SNP dataset were employed to verify SNP-SNP (single nucleotide polymorphism) interaction pairs of seed length (SL), seed width (SW) and seed length/width (SLW) in soybean for 7 years. In total, 1,962, 465 and 1,480 stable interaction pairs for SL, SW and SLW, respectively, were detected by MDR method across more than two years at $p<0.001$ level. In total, there were 37, 2 and 6 interaction pairs which showed significance for SL, SW and seed SLW, respectively. These were screened by the two ways ANOVA test at significant level of $p<0.01$. Six SNP–SNP networks have been constructed based on significant interaction pairs, 57 candidate genes were detected in the network. Two candidate genes located on the hub of network showed extremely related to the seed size, which have been verified and associated with seed size in rice or Arabidopsis. The results will be beneficial to the studies with focus on seed size traits.

Keywords: Soybean seed size; Multifactor dimensionality reduction (MDR); SNP-SNP interaction; SNP–SNP network

Introduction

Soybean (*Glycine max* (L.) Merr) is one of the most important food and oil crops in the world, as it provides a wealth of protein and oil. Many researchers have clarified that seed size traits affects seed yield (Ellis, 1992; Dargahi *et al*., 2014). Seed size traits including seed length (SL), seed width (SW), and seed length/width (SLW), are the major target of breeding, not only as a component of seed yield but also as a morphological quality trait (Wilson, 1995). In soybean, SL, SW and SLW are quantitatively inherited, which controlled by multiple genes and affected by the environment (Xu *et al*., 2011; Hu *et al*., 2013).

Epistasis refers to a non-lineal, non-additive interaction among genotypes at two or more loci (Mackay, 2014). Currently, many studies have been performed involving epistatic interaction analysis. For example, studies about heading date in rice (Qin *et al*., 2015), wheat stripe rust (Vazquez *et al*., 2015), ascochyta blight disease of pea (Timmerman-Vaughan *et al*., 2016), 100 seed weight in wild soybean (Xin *et al*., 2016), seed protein (Qi *et al*., 2016) and fatty acid concentrations (Fan *et al*., 2015). These studies only detected interaction between significant locus, thus, it may miss interaction of other locis. However, the distance of intervals of single nucleotide polymorphisms (SNPs) was narrowed down. The fine information of the SNP-SNP interaction analysis was more than the analysis of QTLs. For example, Lin *et al*. (2013) found an important gene EGFR by a gene interaction network in aggressive prostate cancer. Han *et al*. (2012) studied SNP–SNP interactions between DNA repair genes to uncover gene–gene interaction affect breast cancer risk using logistic regression models and multiple logistic regression models. Onay *et al*. (2006) used multivariate logistic models to study SNP-SNP interactions and found it increasing breast cancer risk. Therefore, genetic interaction networks base on SNP-SNP interactions worked better in expounding epistasis question.

The MDR method was the first used to study polymorphisms related to disease risk (Ritchie *et al*., 2001). A lot of SNP interactions were studied by the MDR method (Ritchie *et al*., 2001; Moore, 2014; Kuo *et al*., 2015). Their research showed that MDR may effectively reduce predictor dimensions of genotype. However, the MDR method is prone to false positive. Then some people have combined with a cross-validation/permutation procedure to optimize this shortcoming (Ritchie *et al*., 2001; Moore, 2014). However, very few researches have been conducted for soybean quantitative trait analysis. Chen *et al*. (2016) first used the MDR method analysis SNP-SNP interaction on soybean oil content, detecting many SNP interactions on oil content.

To cite this paper: Zhang, D., L. Wu, H. Qi, X. Mao, Y. Shi, Z. Wang, H. Jiang, Q. Gong, X. Wu, D. Xin, Q. Chen and Z. Qi, 2018. SNP-SNP interaction analysis on soybean seed size in multiple years. *Int. J. Agric. Biol*., 20: 803−810

In this research, a soybean recombinant inbred line (RIL) population were planted in 7 different years and used a high-density genetic map including 5,308 markers constructed by Qi *et al.* (2014), and used the MDR method to explore stable epistatic interactions related to soybean seed traits (SL, SW and SLW) in multiple years. Then key genes were found by epistatic interactions analysis, SNP–SNP network analysis and gene annotation in quantitative traits under multiple genes controlling. The results will be beneficial to the study of seed size traits and may help improve soybean yield traits.

Materials and Methods

Plant Materials and Trait Evaluation

The 147 RILs population (from $F_{2:16}$ to $F_{2:22}$) crossed by two soybean cultivars: 'Charleston' (♀), an American semi-draft cultivars, and 'Dongnong594' (♂), a Chinese variety, of larger seed size. This RILs populations were planted in Harbin (Harbin; at E. 126°38′ and N. 45°45′) and during from 2008 to 2014. The plants were arranged with 3 replicates in a randomized complete block design (plots were 0.5 m width and 2 m long). Three plants were randomly selected for each row of each plot. Ten seeds were selected from each plant to measure SL and SW by digital vernier caliper and as Qiu and Chang (2006) described. Value of SLW estimated as value of seed length divided by value of seed width.

Phenotypic Data Analysis

The simple correlation among SL, SW and SLW was statistically analyzed using SPSS 17.0 statistical. At $P < 0.05$, it was statistically significant.

Normal distribution test was carried out by One-Sample Kolmogorov-Simrnov Test from the SPSS17.0 statistical. When P value >0.05 the test distribution is considered normal.

Genotyping and Genetic Map Construction

The high-density genetic map was used as described by Qi *et al.* (2014).

Interaction Analysis

To identify SNP × SNP effects in this study, we used MDR method (Ritchie *et al.*, 2001). Among them, we used Pearson chi-square to assess significance ($p < 0.001$). The optimization mode was selected by the maximum Pearson chi-square (Jiang *et al.*, 2009). The chi-square value is a statistic in the non-parametric test it was used to evaluate the association between genotype (high-risk and low-risk group) and affection status (case and control group) in a two-way table. It is calculated as the sum of the square of the difference between the observed and expected frequency in each combination, divided by the expected value, across all combinations:

$$x^2 = \sum \frac{(observed - expected)^2}{expected}$$

The methods were proposed by Cheverud and Routman (1995) to calculate the epistatic interaction effects and their contribution to genetic values and variance.

Results

Phenotypic Variation and Statistical Analysis

The seed size traits (SL, SW and SLW) data of RIL population and parents across7 years are shown in Table 1. The SL of 'Dongnong594' was bigger than that of 'Charleston'. The mean values of SL, SW and SLW of RIL population across 7 years ranged from 6.83 to 7.31, 5.67 to 6.72 and 1.09 to 1.20, respectively. The standard deviation of SL and SW concentrated in 0.30 and the standard deviation of SLW concentrated in 0.05. All traits of the RIL population exhibited continuous distribution and almost showed a normal distribution with Pearson product-moment correlation coefficient ($P > 0.05$), typical of quantitative traits (Table 1).

Simple correlations among seed size traits based on the RIL population means from 2008 to 2014. There was a significant positive correlation between SL and SW, SL and SLW. However, it showed a significant negative correlation between SW and SLW (except 2014 year) in simple correlation analysis (Table 2).

MDR Analysis

The values data and genotype data of SL, SW and SLW of the RIL population across7 years were analyzed separately by the MDR method. The selection level of SNP interaction pairs was the $p<0.001$ (Table 2). In total, 204,063, 91,973 and 263,338 SNP interaction pairs of SL, LW and SLW, respectively, were detected in all years. The SNP interaction pairs of SL were above 10,000 pairs in 2008, 2010 and 2011 year. The SNP interaction pairs of SW were detected all above 10,000 pairs in 2008, 2010 and 2012. The SNP interaction pairs of SLW above 10,000 pairs have been found in 2008, 2011 and 2013.

Stable Interaction Analysis

Stable interaction pairs were obtained by merger and de-emphasis of interaction pairs ($p<0.001$). Stable interaction pairs of SL, LW and SLW were found in different two years with 1,962, 465 and 1,480 pairs, respectively. Stable interaction pairs of SL were mainly appeared on the 2008 and 2010 years. A large quantity of stable interaction pairs of SW were mainly reappeared between different year such as the 2011 and 2012 years, the 2008 and 2010 years, and the 2009 and 2012 years. Stable interaction pairs of SLW were distributed interspersed between different year pairs.

Table 1: Phenotypic variation of seed traits of studied RIL population and parents for 7 years

Traits	Year	P_1	P_2	RIL population								
				Average	SD	CV	Steve	Kurt	Min	Max	Range	P (Sig.)
SL	2008	6.54	6.38	6.83	0.37	0.05	0.61	3.31	5.66	8.43	2.78	0.11
	2009	7.16	7.25	6.90	0.29	0.04	-0.35	0.69	5.89	7.65	1.76	0.65
	2010	7.9	7.7	7.28	0.29	0.04	-0.24	0.86	6.23	8.2	1.98	0.56
	2011	7.05	6.7	6.83	0.32	0.05	0.34	0.24	5.93	7.79	1.86	0.40
	2012	7.45	7.3	7.19	0.35	0.05	0.16	0.91	6.28	8.46	2.18	0.64
	2013	6.95	6.92	7.07	0.43	0.06	0	-0.14	5.96	8.26	2.3	0.94
	2014	7.18	7.4	7.31	0.49	0.07	0.26	0.38	6.08	8.8	2.72	0.68
SW	2008	4.92	5.58	5.68	0.23	0.04	-0.43	2.45	4.63	6.35	1.71	0.50
	2009	6.05	6.57	5.99	0.16	0.03	-0.19	-0.01	5.48	6.35	0.87	0.80
	2010	7.18	6.36	6.03	0.27	0.04	0.14	0.2	5.29	6.82	1.53	0.75
	2011	5.8	5.47	5.67	0.3	0.05	0.21	-0.12	5.02	6.58	1.56	0.82
	2012	6.53	6.5	6.25	0.31	0.05	-0.42	0.57	5.28	7.08	1.8	0.51
	2013	6.1	5.64	6.01	0.33	0.06	-0.14	0.1	4.96	6.75	1.79	0.91
	2014	6.68	6.38	6.72	0.38	0.06	0.05	0.1	5.72	7.74	2.02	0.38
SLW	2008	1.33	1.14	1.2	0.05	0.04	0.68	1.41	1.04	1.34	0.3	0.04
	2009	1.18	1.10	1.15	0.04	0.04	0.18	0.33	1.05	1.27	0.22	0.42
	2010	1.1	1.21	1.21	0.05	0.04	0.28	0.02	1.09	1.35	0.26	0.13
	2011	1.22	1.23	1.21	0.05	0.04	0.46	0.68	1.11	1.37	0.26	0.27
	2012	1.14	1.12	1.15	0.05	0.04	0.27	-0.09	1.06	1.29	0.23	0.51
	2013	1.14	1.23	1.18	0.05	0.05	0.33	0.11	1.04	1.35	0.31	0.31
	2014	1.07	1.16	1.09	0.04	0.04	0.72	1.02	1	1.23	0.23	0.06

Note: P_1-Dongnong594, P_2-Charleston, SD-standard deviation, CV-Coefficient of Variation, Steve-Skewness, Kurt-Kurtosis. P (Sig.) value is One-Sample Kolmogorov-Simrnov Test

Table 2: Simple and partial correlation coefficients for seed traits in soybean

Traits	2008SL	2008SW	2009SL	2009SW	2010SL	2010SW	2011SL	2011SW	2012SL	2012SW	2013SL	2013SW	2014SL	2014SW
2008SW	0.65**													
2008SLW	0.60**	-0.22 **												
2009SW			0.51**											
2009SLW			0.74**	-0.21**										
2010SW					0.55**									
2010SLW					0.40**	-0.59 **								
2011SW							0.78**							
2011SLW							0.20**	-0.52**						
2012SW									0.68**					
2012SLW									0.39**	-0.42**				
2013SW											0.69**			
2013SLW											0.49**	-0.30**		
2014SW													0.85**	
2014SLW													0.55**	0.24

Note: ** Significant at 0.01 levels

Among the 20 linkage groups, for SL trait, one side of the most stable interaction pairs were located on Gm07 with others, including Gm01, Gm03, Gm06, Gm13, Gm15 and Gm20. Some of these SNPs interacted with other SNPs at a higher frequency, these locus were hot regions. For example, on Gm20, Mark538827 (2.566Mb), Mark547168 (2.482Mb), Mark582063 (2.003Mb), Mark581037 (1.896Mb), Mark554062 (1.743Mb), Mark571544 (1.221Mb), Mark578284 (0.524Mb) and Mark522605 (0.175Mb) with other SNPs constituted 225 pairs, 225 pairs, 225 pairs, 214 pairs, 143 pairs, 212 pairs, 212 pairs, and 214 pairs stable interaction pairs, respectively. For SW trait, detected stable interaction pairs were distributed scattered, however, Gm16 with Gm02 was notable. On Gm13, Mark105947 (30.064Mb) and Mark108826 (33.626Mb) with other SNPs constituted 31 pairs and 45 pairs stable interaction pairs, respectively. On Gm16, Mark1217476 (23.346Mb), Mark1202430-Mark1230181 (33.246-33.520Mb) and Mark1222957-Mark1244664 (35.204-35.414Mb) with other SNPs constituted 20 pairs, 101 pairs and 80 pairs stable interaction pairs, respectively. For SLW trait, Gm17 with Gm19 detected the most stable interaction pairs, Gm20 with others also were notable. On Gm20, Mark1158266 (1.224Mb), Mark1177650 (1.223Mb), Mark1123725 (1.295Mb), Mark1158928 (1.523Mb) and Mark1179955 (45.778Mb) with other SNPs constituted 52 pairs, 170 pairs, 82 pairs, 170 pairs and 92 pairs stable interaction pairs. In these hotspot SNPs, Mark538827, Mark547168, Mark582063, Mark581037 and Mark554062 are mapped to *qSL-7* detected by Hu *et al*. (2013), while other hot zone SNPs have not been found in QTLs found by others. There were three stable interaction pairs for SL and SW including Mark538827 with Mark105947, Mark547168 with Mark105947 and Mark582063 with Mark105947. There were no stable interaction pairs for these three traits (Fig. 1).

Epistatic Effect and Contribution Rate Analysis

Significant interaction pairs were screened by the two ways ANOVA test on epistatic interaction effects and their contribution to genetic values (at significant level of $p<0.01$). In total, there were 37, 2 and 6 SNP interaction pairs that were significant in two years, in SL, SW and SLW respectively (Table 3).

The highest epistasis value and highest contribution rate of SL were 0.0620 and 5.8756% respectively, which the corresponding interaction pair was Mark555489 with Mark1173999 in 2010. The minimum epistasis value and contribution rate of SL were 0.0059 and 0.4845% respectively, which the corresponding Mark478646 with Mark571544 in 2008. The epistasis value and contribution rate of SW were 0.0344 and 4.0383%, respectively in 2008. The epistasis value and contribution rate of SW were, 0.0285 and 3.8599% respectively in 2011. The highest epistasis value and contribution rate of SLW were 0.0008, 0.0784% respectively, which the corresponding Mark366903 with Mark1179955 in 2010. The minimum epistasis value and contribution rate of SL/SW were 0.0002, 0.0176%, respectively that the interaction pair was Mark353845 with Mark557445 in 2008 (Table 3).

Significant SNP interaction detected in this research showed no matches with previous QTL epistasis research. However, there was some stable and significant interaction pairs matched with the main effect QTL reported previously without interaction effects. Mark538827 (2.566Mb) and Mark547168 (2.482Mb) on Gm07 have been mapped seed length major QTL fragments in *Seed length 1-6* (Salas *et al*., 2006) and *qSL-7* (Hu *et al*., 2013). Some regions on Gm07 Mark562451 (5.997Mb), Mark526852 (5.222Mb), Mark566274 (5.064Mb), Mark555489 (5.260Mb), Mark525636 (5.367Mb), Mark582063 (2.003Mb) and Mark554062 (1.743Mb) all have been detected in *qSL-7* (Hu *et al*., 2013). Mark995411 (48.379Mb) on Gm02 was found in *qSW-2-3* (Xu *et al*., 2011) and in *qSW-2* (Hu *et al*., 2013).

SNP–SNP Network Analysis and Candidate Genes Mining

There were based on significant interaction pairs to construct networks affecting soybean seed size traits. Three SNP epistatic interaction subnets containing more than one node based on significant interaction pairs are shown in Fig. 2. Subnet A, B, C and D are SNP–SNP Network of SL, subnet E is SNP–SNP Network of SW and subnet F is SNP–SNP Network of SLW. Subnet A contained SNP pairs from five linkage groups, which is the largest number of linkage groups in all the subnet. Mark571544 on Gm07 with 15 two-way interactions, the maximum degree, could be considered the hub site of subnet A. Mark670797 (on Gm01)/Mark522605 (Gm07), Mark582063 (Gm07),

Fig. 1a: Stable SNP interactions related to SL for7 years ($p<0.001$), **(b):** Stable SNP interactions related to SW for7 years ($p < 0.001$) and **(c):** Stable SNP interactions related to SLW for7 years ($p < 0.001$)

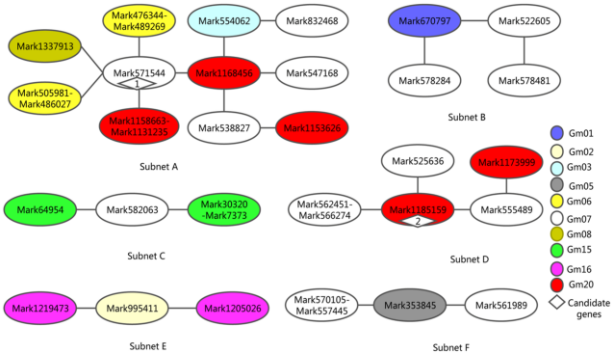

Fig. 2: Epistatic interaction network based significant SNP interaction pairs affecting seed traits. (Note: The figure shows the construction of significant SNP pairs for subnets A, B, C, D, E and F. Nodes are colored according to linkage groups as follows: Gm01, blue; Gm02 light yellow; Gm05, black; Gm06, yellow; Gm07, white Gm08, orange; Gm15, green; Gm16, purple; Gm20, red. Each edge corresponds to a two-way interaction; the degree of each node refers to the number of connecting edges. 1and 2 were Glyma07g01840 and Glyma20g36690, respectively)

Table 3: Significant SNP interactions for seed traits (Seed length, Seed width and seed length/ width)

Traits	Significant Interaction pairs								Interaction years		E^2	I^2	E^{2a}	I^{2a}
	SLAF Marker	LG	Physical interval (bp)		SLAF Marker	LG	Physical interval (bp)							
SL	Mark670797	Gm01	38998419	38998739	Mark522605	Gm07	175113	175424	2008	2010	0.0174	2.34%	0.0318	4.15%
	Mark670797	Gm01	38998419	38998739	Mark578284	Gm07	524274	524593	2008	2010	0.0326	4.82%	0.0265	3.90%
	Mark832468	Gm03	42316707	42317024	Mark554062	Gm07	1743015	1743335	2008	2010	0.0186	2.39%	0.0272	3.35%
	Mark505981	Gm06	45242787	45243078	Mark571544	Gm07	1221503	1221792	2008	2010	0.0089	0.75%	0.0125	1.12%
	Mark478531	Gm06	45266519	45266819	Mark571544	Gm07	1221503	1221792	2008	2010	0.0089	0.75%	0.0125	1.12%
	Mark489820	Gm06	44730716	44730992	Mark571544	Gm07	1221503	1221792	2008	2010	0.0089	0.75%	0.0125	1.12%
	Mark478646	Gm06	46085306	46085587	Mark571544	Gm07	1221503	1221792	2008	2010	0.0059	0.48%	0.0099	0.91%
	Mark486027	Gm06	47585104	47585354	Mark571544	Gm07	1221503	1221792	2008	2010	0.0107	0.83%	0.0172	1.50%
	Mark476344	Gm06	47630602	47630889	Mark571544	Gm07	1221503	1221792	2008	2010	0.0085	0.66%	0.0156	1.37%
	Mark445104	Gm06	47624809	47625076	Mark571544	Gm07	1221503	1221792	2008	2010	0.0124	0.93%	0.0126	1.15%
	Mark478575	Gm06	47005658	47005960	Mark571544	Gm07	1221503	1221792	2008	2010	0.0116	0.96%	0.0131	1.15%
	Mark459140	Gm06	46947520	46947781	Mark571544	Gm07	1221503	1221792	2008	2010	0.0149	1.17%	0.0101	0.92%
	Mark510387	Gm06	46775195	46775485	Mark571544	Gm07	1221503	1221792	2008	2010	0.0106	0.77%	0.0095	0.92%
	Mark489269	Gm06	46651267	46651561	Mark571544	Gm07	1221503	1221792	2008	2010	0.0095	0.80%	0.013	1.17%
	Mark578481	Gm07	23785102	23785305	Mark522605	Gm07	175113	175424	2008	2010	0.0138	1.49%	0.0152	1.57%
	Mark562451	Gm07	5997112	5997385	Mark1185159	Gm20	44749103	44749376	2008	2010	0.0282	2.90%	0.0423	4.42%
	Mark526852	Gm07	5222424	5222708	Mark1185159	Gm20	44749103	44749376	2008	2010	0.0365	3.33%	0.0388	4.02%
	Mark566274	Gm07	5064175	5064476	Mark1185159	Gm20	44749103	44749376	2008	2010	0.037	3.43%	0.0392	4.12%
	Mark555489	Gm07	5260392	5260668	Mark1185159	Gm20	44749103	44749376	2008	2010	0.0323	2.87%	0.0376	3.92%
	Mark555489	Gm07	5260392	5260668	Mark1173999	Gm20	43578962	43579264	2008	2010	0.0434	3.87%	0.062	5.88%
	Mark525636	Gm07	5367141	5367431	Mark1185159	Gm20	44749103	44749376	2008	2010	0.051	3.96%	0.0461	4.85%
	Mark538827	Gm07	2566449	2566734	Mark1153626	Gm20	35400169	35400441	2008	2010	0.0355	2.71%	0.0143	1.81%
	Mark538827	Gm07	2566449	2566734	Mark1168456	Gm20	35258259	35258557	2008	2010	0.0237	1.95%	0.0185	1.85%
	Mark547168	Gm07	2481807	2482103	Mark1168456	Gm20	35258259	35258557	2008	2010	0.0143	1.11%	0.0158	1.62%
	Mark582063	Gm07	2002941	2003236	Mark30320	Gm15	21256364	21256658	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark14537	Gm15	20505752	20506049	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark2782	Gm15	22362772	22363070	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark59416	Gm15	22397177	22397446	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark49678	Gm15	20585924	20586215	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark32716	Gm15	39983463	39983753	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark7373	Gm15	21702438	21702723	2008	2010	0.0131	1.34%	0.0089	0.91%
	Mark582063	Gm07	2002941	2003236	Mark64954	Gm15	32366868	32367152	2008	2010	0.014	1.42%	0.0083	0.84%
	Mark571544	Gm07	1221503	1221792	Mark1337913	Gm08	11864406	11864695	2008	2010	0.0381	3.28%	0.0192	1.61%
	Mark571544	Gm07	1221503	1221792	Mark1158663	Gm20	35433067	35433337	2008	2010	0.0364	3.45%	0.0164	1.80%
	Mark571544	Gm07	1221503	1221792	Mark1131235	Gm20	35378595	35378867	2008	2010	0.0284	2.66%	0.0198	2.11%
	Mark571544	Gm07	1221503	1221792	Mark1168456	Gm20	35258259	35258557	2008	2010	0.0401	3.52%	0.028	3.01%
	Mark554062	Gm07	1743015	1743335	Mark1168456	Gm20	35258259	35258557	2008	2010	0.0379	3.09%	0.0325	3.10%
SW	Mark995411	Gm02	48378623	48378919	Mark1219473	Gm16	8839003	8839303	2008	2011	0.0344	4.04%	0.0285	3.86%
	Mark995411	Gm02	48378623	48378919	Mark1205026	Gm16	8926109	8926397	2008	2011	0.0344	4.04%	0.0285	3.86%
SLW	Mark353845	Gm05	36961723	36962001	Mark570105	Gm07	20535122	20535434	2008	2010	0.0002	0.02%	0.0004	0.04%
	Mark353845	Gm05	36961723	36962001	Mark548065	Gm07	29759163	29759447	2008	2010	0.0002	0.02%	0.0005	0.05%
	Mark353845	Gm05	36961723	36962001	Mark557445	Gm07	27034672	27034944	2008	2010	0.0002	0.02%	0.0005	0.05%
	Mark353845	Gm05	36961723	36962001	Mark561989	Gm07	25686053	25686317	2008	2010	0.0002	0.02%	0.0005	0.05%
	Mark366903	Gm10	152929	153216	Mark1179955	Gm20	45777989	45778289	2008	2010	0.0003	0.03%	0.0008	0.08%
	Mark1389100	Gm17	14609778	14610118	Mark1136879	Gm20	1491796	1492090	2010	2014	0.0004	0.05%	0.0004	0.04%

Note: E^2 and I^2: The epistasis value and contribution rate of significant interaction pairs in year1of interaction year; E^{2a} and I^{2a}: The epistasis value and contribution rate of significant interaction pairs in year2 of interaction year

Mark1185159 (Gm20), Mark995411 (Gm02) and Mark353845 (Gm05) were found in hub sites of subnets B, C, D, E and F, respectively.

Based on the physical mark position of two sides of significant SNP interaction of networks, 57 candidate genes were annotated from the database of Glycine max Wm82.a2.v1 (http://phytozome.jgi.doe.gov/pz/portal.html#!info?alias=Org_Gmax) Among them, Glyma07g01840 and Glyma20g36690 are annotated based on the physical mark position of the hub site-Mark571544 and Mark1185159, respectively (Table S1).

Discussion

Soybean seed size traits (SL, SW and SLW) are important quantitative traits under multiple genes controlling. In this study, the MDR method was used to identify stable loci controlling seed traits (SL, SW and SLW) in soybean across multiple years based on a high-density genetic map.

Epistasis is common and can cause cryptic genetic variation for quantitative traits in natural populations (Gibson and Dworkin, 2004; Mackay, 2014). Currently, there are many ways to detect epistatic SNP-SNP interactions, for example, heuristic (Carlborg *et al*., 2000), MDR (Ritchie *et al*., 2001), exhaustive algorithms (Nelson *et al*., 2001), mutual information (Curk *et al*., 2011) and other methods (Su *et al*., 2015). The MDR analysis can reduce genotype predictor dimensions and combined cross-validation–testing/permutation testing method to minimize the rate of false positive findings. Li and Sun (2016) used MDR to analyze SNP–SNP interactions related to essential

Table S1: Seed size candidate genes

Candidate Gene	GO	Locus tag	Gene description
Glyma06g42040	GO:0005524,GO:0016887,GO:0042626,GO:0006810,GO:0055085,GO:0016021	AT3G28345.1	ABC transporter family protein
Glyma02g43602		AT1G05010.1	ethylene-forming enzyme
Glyma02g43610		AT3G47810.1	Calcineurin-like metallo-phosphoesterase superfamily protein
Glyma02g43620		AT2G04900.1	
Glyma02g43630	GO:0005515,GO:0043531,GO:0007165,GO:0005622,GO:0045087,GO:0031224,GO:0004888,GO:0006915,GO:0005524	AT5G17680.1	disease resistance protein (TIR-NBS-LRR class), putative
Glyma05g31920		AT5G22510.1	alkaline/neutral invertase
Glyma05g31930		AT3G52860.1	
Glyma05g31940		AT4G39390.1	nucleotide sugar transporter-KT 1
Glyma06g41461		AT3G07160.1	glucan synthase-like 10
Glyma06g42061	GO:0003743,GO:0006413	AT4G27130.1	Translation initiation factor SUI1 family protein
Glyma06g42071	GO:0003743,GO:0006413	AT4G27130.1	Translation initiation factor SUI1 family protein
Glyma06g42750	GO:0008234,GO:0006508	AT5G45890.1	senescence-associated gene 12
Glyma06g42770	GO:0008234,GO:0006508	AT5G50260.1	Cysteine proteinases superfamily protein
Glyma06g42780	GO:0008234,GO:0006508	AT5G45890.1	senescence-associated gene 12
Glyma06g43630		AT2G42570.1	TRICHOME BIREFRINGENCE-LIKE 39
Glyma06g43641		AT2G42560.1	late embryogenesis abundant domain-containing protein/LEA domain-containing protein
Glyma06g43741		AT3G58100.1	plasmodesmatacallose-binding protein 5
Glyma06g43750		AT3G58110.1	
Glyma06g43970	GO:0008171,GO:0008168,GO:0046983	AT4G35160.1	O-methyltransferase family protein
Glyma06g44740		AT4G03600.1	
Glyma06g44780			
Glyma06g44790	GO:0016020	AT2G20725.1	CAAX amino terminal protease family protein
Glyma06g44800	GO:0008080,GO:0016747,GO:0008152	AT1G03650.1	Acyl-CoA N-acyltransferases (NAT) superfamily protein
Glyma06g44821		AT1G12800.1	Nucleic acid-binding, OB-fold-like protein
Glyma07g00380		AT3G20240.1	Mitochondrial substrate carrier family protein
Glyma07g00391	GO:0005783	AT1G78895.1	Reticulon family protein
Glyma07g00400	GO:0006412,GO:0005840,GO:0005622,GO:0003735	AT3G20260.1	Protein of unknown function (DUF1666)
Glyma07g00410	GO:0005198,GO:0009507	AT2G46910.1	Plastid-lipid associated protein PAP / fibrillin family protein
Glyma07g00920	GO:0016702,GO:0046872,GO:0055114,GO:0005515	AT1G55020.1	lipoxygenase 1
Glyma07g01820		AT4G12540.1	
Glyma07g01830		AT1G79730.1	hydroxyproline-rich glycoprotein family protein
Glyma07g01840	GO:0004871,GO:0000160	AT3G16360.2	HPT phosphotransmitter 4
Glyma07g02571		AT1G73060.1	Low PSII Accumulation 3
Glyma07g02580		AT1G16880.1	uridylyltransferase-related
Glyma07g02590	GO:0008080,GO:0008152	AT4G37580.1	Acyl-CoA N-acyltransferases (NAT) superfamily protein
Glyma07g02930	GO:0003700,GO:0006355	AT5G25190.1	Integrase-type DNA-binding superfamily protein
Glyma07g03540		AT1G52630.1	O-fucosyltransferase family protein
Glyma07g03550	GO:0003676	AT2G34160.1	Alba DNA/RNA-binding protein
Glyma07g03560		AT1G80160.1	Lactoylglutathionelyase/glyoxalase I family protein
Glyma07g06331			
Glyma07g06340		AT1G01430.1	TRICHOME BIREFRINGENCE-LIKE 25
Candidate Gene	GO	Locus tag	Gene description
Glyma07g06480	GO:0009001,GO:0006535,GO:0005737	AT5G56760.1	serine acetyltransferase 1;1
Glyma07g06520	GO:0003723,GO:0033897	AT2G39780.1	ribonuclease 2
Glyma07g06700	GO:0005515	AT3G61600.1	POZ/BTB containin G-protein 1
Candidate Gene	GO	Locus tag	Gene description
Glyma07g07290	GO:0004650,GO:0005975	AT3G61490.1	Pectin lyase-like superfamily protein
Glyma08g16240		AT5G40410.1	Tetratricopeptide repeat (TPR)-like superfamily protein
Glyma08g16251		AT2G41905.1	
Glyma15g21980			
Glyma15g23270	GO:0003735,GO:0006412,GO:0005622,GO:0005840	AT4G18100.1	Ribosomal protein L32e
Glyma15g35351		AT5G54130.2	Calcium-binding endonuclease/exonuclease/phosphatase family
Glyma20g25590		AT1G15060.1	Uncharacterised conserved protein UCP031088, alpha/beta hydrolase
Glyma20g25600	GO:0005515	AT1G49540.1	elongator protein 2
Glyma20g25750			
Glyma20g25790	GO:0004332,GO:0006096	AT2G01140.1	Aldolase superfamily protein
Glyma20g35300		AT1G04230.1	Protein of unknown function (DUF2361)
Glyma20g36690	GO:0004672,GO:0005524,GO:0006468	AT3G04810.1	NIMA-related kinase 2
Glyma20g36700		AT4G14746.1	

hypertension in the Chinese Han population. Rai *et al.* (2015) performed MDR to investigate the gene-gene interactions involved in gallbladder cancer pre-disposition. de Guia *et al.* (2015) used this technique to reveal the interactions of important gene variants involved in allergies.

In this research, the MDR applied to analyze soybean quantitative traits. SNP interaction pairs of SL, LW and SLW were detected for all 7 years, which were 204,063, 91,973 and 263,338 pairs, respectively. Stable interaction pairs were obtained by merger and de-emphasis of

interaction pairs ($p<0.001$) have 1,962 SL pairs, 465 SW pairs, 1,480 SLW pairs, respectively were in two different years. In the stable interaction pairs, some SNPs and other SNP markers are alternately classified as hot regions. There are 18 hot regions. Very few of these hot regions were matched with QTLs previously detected. Then, we identified 37 SL, 2 SW, 6 SLW significant SNP interaction pairs by the two ways ANOVA test ($p<0.01$) based on epistatic interaction effects and their contribution to genetic values. The highest epistasis value and highest contribution rate of significant SNP interaction pairs were 0.0620 and 5.8756% ($p<0.01$) in seed size, respectively. The minimum epistasis value and contribution rate of significant SNP interaction pairs were 0.0002 and0.0176% ($p<0.01$) in seed size, respectively. One-way of some significant SNP interactions has been detected in previous studies, but there is no fully matched epistemic interaction. These significant SNP-SNP interactions pairs are new discoveries.

Li *et al*. (2013) and Lezon *et al*. (2006) found a lot of important information in network. In this research, six interaction networks were constructed based on stable and significant SNP interaction pairs. By the basis SNP-SNP network annotation, obtained 57 candidate genes. Mark571544 and Mark1185159 were located on the hubs in the SNP -SNP interaction network.G*lyma07g01840* was annotated as HPT phosphotransmitter4 (AHP4), which the homologous gene is *At3g16360* in Arabidopsis, Jung *et al*. (2008) Hutchison *et al*. (2006) suggest that *At3g16360* affects the seed size and some cytokinin responses. *Glyma20g36690* was annotated as Never in Mitosis gene A (NIMA)-related kinase2 (NEK2), which the homologous genes are*OsNek3* (NEK3) in rice and *At3g44200* (NEK6) in Arabidopsis. Fujii *et al*. (2009) research finding *OsNek3*-overexpressing lines showed indirectly affects seed length in rice. Zhang *et al*. (2011) found the *NEK6* gene may reduce seed size in Arabidopsis. Therefore, inferencing Mark571544 and Mark1185159 play an important role in controlling seed size traits. We speculated that *Glyma07g01840* and *Glyma20g36690* play an important role in seed size development.

Conclusion

This research found 18 hot regions, 45 significant SNP interaction pairs, 6 interaction networks, and 2 candidate genes controlling seed size traits significantly. This will be beneficial to the studied with focus on seed size traits.Mark538827, Mark54716 and Mark582063 can be developed for molecular assisted breeding. Six interaction networks were constituted significant SNP interaction pairs with the higher epistasis value and higher contribution rate. SNP -SNP interaction network A and D contained the larger number of significant interaction pairs, where their hub site is Mark571544 (Gm20) and Mark1185159 (Gm20), respectively. Furthermore, 2 candidate genes, *Glyma07g01840* and *Glyma20g366690*, were predicted on the hubs. The function of their homologous genes had been verified and associated with seed size on rice or Arabidopsis (Hutchison *et al*., 2006; Jung *et al*., 2008; Fujii *et al*., 2009; Zhang *et al*., 2011), thus validation of genes function should be conducted in soybean for next step.

Acknowledgements

This study was conducted in the Key Laboratory of Soybean Biology in Chinese Ministry of Education in Heilongjiang Province and financially supported by the Research Fund for the "soybean germplasm innovation" of the "seven main crop projects" of Ministry of Science and Technology of China (grant 2016YFD0100300), the national natural science foundation of China (31471516, 31271747, 31401465, 31400074, 31501332), the China Post Doctoral Project (2015M581419), "Dongnongxuezhe project (To Chen QS)" "Qingniancaijun project (To Qi ZM, 518062)" of Northeast agriculture university, "Ministry of Science and Technology - soybean quality improvement (grant 2016YFD0100500)" and SIPT Project of Northeast Agricultural University (201610224146).

References

Carlborg, Ö., L. Andersson and B. Kinghorn, 2000. The use of a genetic algorithm for simultaneous mapping of multiple interacting quantitative trait loci. *Gene*, 155: 2003−2010

Chen, Q., X. Mao, Z. Zhang, R. Zhu, Z. Yin, Y. Leng, H. Yu, H. Jia, S. Jiang, Z. Ni, H. Jiang, X. Han, C. Liu, Z. Hu, X. Wu, G. Hu, D. Xin and Z. Qi, 2016. SNP-SNP interaction analysis on soybean oil content under multi-environments. *PloS One*, 11: e0163692

Cheverud, J.M. and E.J. Routman, 1995. Epistasis and its contribution to genetic variance components. *Gene*, 139: 1455−1461

Curk, T., G. Rot and B. Zupan, 2011. SNPsyn: detection and exploration of SNP-SNP interactions. *Nucl. Acids Res.*, 39: 444−449

Dargahi, H., P. Tanya, P. Somta, J. Abe and P. Srinives, 2014. Mapping quantitative trait loci for yield-related traits in soybean (*Glycine max* L.). *Breed. Sci*., 64: 282−290

de Guia, R.M., M.D.J. Echavez, E.L.C. Gaw, M.R.R. Gomez, K.A.J. Lopez, R.C.M. Mendoza, J.M.C. Rapsing, D.P. Retreta, C.M.B. Tubog, M.H. Ventolero, C.L. Yao and J.D.A. Ramos, 2015. Multifactor-dimensionality reduction reveals interaction of important gene variants involved in allergy. *Int. J. Immunogenet.*, 42: 182−189

Ellis, R.H., 1992. Seed and seedling vigour in relation to crop growth and yield. *Plant Growth Regul*., 11: 249−255

Fan, S.X., B. Li, F.K. Yu, F.X. Han, S.R. Yan, L.Z. Wang and J.M. Sun, 2015. Analysis of additive and epistatic quantitative trait loci underlying fatty acid concentrations in soybean seeds across multiple environments. *Euphytica*, 206: 689−700

Fujii, S., M. Yamada and K. Toriyama, 2009. Cytoplasmic male sterility-related protein kinase, *OsNek3*, is regulated downstream of mitochondrial protein phosphatase 2C, DCW11. *Plant Cell Physiol*., 50: 828−837

Gibson, G. and I. Dworkin, 2004. Uncovering cryptic genetic variation. *Nat. Rev. Genet.*, 5: 681−690

Han, W., K.Y. Kim, S.J. Yang, D.Y. Noh, D. Kang and K. Kwack, 2012. SNP-SNP interactions between DNA repair genes were associated with breast cancer risk in a Korean population. *Cancer*, 118: 594−602

Hu, Z.B., H.R. Zhang, G.Z. Kan, D.Y. Ma, D. Zhang, G.X. Shi, D.L. Hong, G.Z. Zhang and D.Y. Yu, 2013. Determination of the genetic architecture of seed size and shape via linkage and association analysis in soybean (*Glycine max* L. Merr.). *Genetics*, 141: 247–254

Hutchison, C.E., J. Li, C. Argueso, M. Gonzalez, E. Lee, M.W. Lewis, B.B. Maxwell, T.D. Perdue, G.E. Schaller, J.M. Alonso, J.R. Ecker and J.J. Kieber, 2006. The *Arabidopsis* histidine phosphotransfer proteins are redundant positive regulators of cytokinin signaling. *Plant Cell*, 18: 3073−3087

Jiang, Y.S., R.J. Zhang, G.Y. Liu, Z. Wang, Z.Q. Chen, P. Sun, C. Huang and X.H. Zhang, 2009. Multifactor dimensionality reduction for detecting haplotype- haplotype interaction. *In: Proceedings of the 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery Discovery*, Volume 01. pp: 241−245. IEEE Computer Society Washington, DC, USA

Jung, K.W., S.I. Oh, Y.Y. Kim, K.S. Yoo, M.H. Cui and J.S. Shin, 2008. Arabidopsis histidine-containing phosphotransfer factor 4 (AHP4) negatively regulates secondary wall thickening of the anther endothecium during flowering. *Mol. Cells*, 25: 294−300

Kuo, H.C., J.C. Chang, M.M. Guo, K.S. Hsieh, D. Yeter, S.C. Li and K.D. Yang, 2015. Gene-gene associations with the susceptibility of Kawasaki disease and coronary artery lesions. *PloS One*, 10: e0143056

Lezon, T.R., J.R. Banavar, M. Cieplak, A. Maritan and N.V. Fedoroff, 2006. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proc. Nat. Acad. Sci*., 103: 19033−19038

Li, F.G., G. Hu, H. Zhang, S.Z. Wang, Z.P. Wang and H. Li, 2013. Epistaticeffects on abdominal fat content in chickens: results from a genome-wide snp-snp interaction analysis. *PloS One*, 8: e81520

Li, Y.B. and G.Q. Sun, 2016. Case-control study on association of peroxisome proliferator-activated receptor-δ and SNP-SNP interactions with essential hypertension in Chinese Han population. *Funct. Integr. Genom.*, 16: 95−100

Lin, H.Y., E.K. Amankwah, T.S. Tseng, X. Qu, D.T. Chen and J.Y. Park, 2013. SNP-SNP interaction network in angiogenesis genes associated with prostate cancer aggressiveness. *PloS One*, 8: e59688

Mackay, T.F.C.,2014.Epistasis and quantitative traits: using model organisms to study gene-gene interactions. *Nat. Rev. Genet.*, 15: 22−23

Moore, J.H., 2014. Computational analysis of gene-gene interactions using multifactor dimensionality reduction. *Expert Rev. Mol. Diagnost.*, 4: 795−803

Nelson, M.R., S.L. Kardia, R.E. Ferrell and C.F. Sing, 2001. A combinatorial partitioning method to identify multilocus genotypic partitions that predict quantitative trait variation. *Genom. Res.*, 11: 458−470

Onay, V.Ü., L. Briollais, J.A. Knight, E. Shi, Y.Y. Wang, S. Wells, H. Li, I. Rajendram, L.I. Andrulis and H. Ozcelik, 2006. SNP-SNP interactions in breast cancer susceptibility. *BMC Cancer*, 6: 114

Qi, Z.M., J.B. Pan, X. Han, H.D. Qi, D.W. Xin, W. Li, X.R. Mao, Z.Y. Wang, H.W. Jiang, C.Y. Liu, Z.B. Hu, G.H. Hu, R.S. Zhu and Q.S. Chen, 2016. Identification of major QTLs and epistatic interactions for seed protein concentration in soybean under multiple environments based on a high-density map. *Mol. Breed.*, 36: 55

Qi, Z.M., L. Huang, R.S. Zhu, D.W. Xin, C.Y. Liu, X. Han, H.W. Jiang, W.G. Hong, G.H. Hu, H.K. Zheng and Q.S. Chen, 2014. A high-density genetic map for soybean based on specific length amplified fragment sequencing. *PloS One*, 9: e104871

Qin, M., X.Q. Zhao, J. Ru, G.Q. Zhang and G.Y. Ye, 2015. Bigenic epistasis between QTLs for heading date in rice analyzed using single segment substitution lines. *Field Crops Res.*, 178: 16−25

Qiu, L.J. and R.Z. Chang, 2006. Descriptors and data standard for soybean (*Glycine* spp.). Agricultural Press, Beijing, Chinese

Rai, R., J.J. Kim, S. Misra, A. Kumar and B. Mittal, 2015. A multiple interaction analysis reveals ADRB3 as a potential candidate for gallbladder cancer predisposition via a complex interaction with other candidate gene variations. *Int. J. Mol. Sci*., 16: 28038−28049

Ritchie, M.D., L.W. Hahn, N. Roodi, L.R. Bailey, W.D. Dupont, F.F. Parl and J.H. Moore, 2001. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cance. *Amer. J. Human Genet*., 69: 138−147

Salas, P., J.C. Oyarzo-Llaipen, D. Wang, K. Chase and L. Mansur, 2006. Genetic mapping of seed shape in three populations of recombinant inbred lines of soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet*., 113: 1459−1466

Su, L.T., G.X. Liu, H. Wang, Y. Tian, Z.H. Zhou, L. Han and L. Yan, 2015. Research on single nucleotide polymorphisms interaction detection from network perspective. *PloS One*, 10: e0119146

Timmerman-Vaughan, G.M., L. Moya, T.J. Frew, S.R. Murray and R. Crowhurst, 2016. Ascochyta blight disease of pea (*Pisum sativum* L.): defence-related candidate genes associated with QTL regions and identification of epistatic QTL. *Theor. Appl. Genet*., 129: 879−896

Vazquez, M.D., R. Zemetra, C.J. Peterson, X.M. Chen, A. Heesacker and C.C. Mundt, 2015. Multi-location wheat stripe rust QTL analysis: genetic background and epistatic interactions. *Theor. Appl. Genet*., 128: 1307−1318

Wilson, D.J., 1995. Storage of orthodox seeds. *In*: *Seed Quality: Basic Mechanisms, Agricultural Implications*, pp: 173−208. A.S. Basra (ed.). Food Products Press, New York, USA

Xin, D., Z. Qi, H. Jiang, Z. Hu, R. Zhu, J. Hu, H. Han, G. Hu, C. Liu and Q. Chen, 2016. QTL location and epistatic effect analysis of 100-seed weight using wild soybean (*Glycine soja* Sieb. & Zucc.) chromosome segment substitution lines. *PloS One*, 11: e0149380

Xu, Y., H.N. Li, G.J. Li, X. Wang, L.G. Cheng and Y.M. Zhang, 2011. Mapping quantitative trait loci for seed size traits in soybean (*Glycine max* L. Merr.). *Theor. Appl. Genet*., 122: 581−594

Zhang, B., H.W. Chen, R.L. Mu, W.K. Zhang, M.Y. Zhao, W. Wei, F. Wang, H. Yu, G. Lei, H.F. Zou, B. Ma, S.Y. Chen and J.S. Zhang, 2011. NIMA-related kinase NEK6 affects plant growth and stress response in *Arabidopsis*. *Plant J*., 68: 830−843

(Received 23 June 2017; Accepted 18 December 2017)