

GAI Junyi, WANG Yongjun, WU Xiaolei, CHEN Shouyi

A comparative study on segregation analysis and QTL mapping of quantitative traits in plants—with a case in soybean

© Higher Education Press and Springer-Verlag 2007

Abstract Two approaches of genetic analysis of quantitative traits were compared with a case study on soybean. One approach was the segregation analysis developed by Gai et al. (2003), which utilized information from individuals of one or multiple segregation populations as well as that from parents based on the principles of the major-gene plus polygene inheritance model, mixture distribution, joint maximum-likelihood function, IECM (Iterated Expectation and Conditional Maximization) algorithm, and Akaike's information criterion and goodness of fit tests. Another approach was quantitative trait locus (QTL) mapping with molecular markers. A recombinant inbred line (RIL) population with 201 families derived from (Kefeng No.1 × 1138-2) $F_{2:7:10}$ along with their parents were tested in a randomized block design experiment. The 171 RFLP, 60 SSR, and 79 AFLP molecular markers were used to mark the 201 families. The data of nine traits, i.e., number of days to flowering, number of days to maturity, plant height, number of nodes on main stem, number of pods per node, 100-seed weight, protein content, oil content, and plot yield, were analyzed with the segregation analysis procedure of RIL population with parents (Gai et al., 2003; Zhang and Gai, 2000; Zhang et al., 2001) to detect their genetic system, and those along with the molecular marker data were analyzed with WinQTL Cartographer (Basten et al., 1999; Zeng, 1993, 1994) to detect their QTL system. The results showed that both procedures could detect the main major genes or QTLs, and therefore, could be used as a mutual check and supplement. From the results that

most of the traits were mainly controlled by three or four QTLs, it was impressed that the segregation analysis procedure of four major-gene plus polygene mixed inheritance model should be developed to fit the requirements. The results also showed that the QTLs of the involved traits concentrated on several linkage groups, such as C2, B1, F1, M, and N. Finally, the results showed that the experimental sample was not necessarily coincident with the theoretical population according to equality test, symmetry test, and representation test, and therefore, the sample should be checked, tested and then adjusted to fit the theoretical requirements through deleting the extra-biased families and markers.

Keywords inheritance of quantitative trait, segregation analysis, QTL mapping, soybean

1 Introduction

Fisher (1918) established the earliest genetic model, $p = g + e$, $g = a + d$, based on Nelssen Ehle's multiple factor hypothesis of quantitative traits. Under polygene hypothesis, various approaches for detecting the genetic system (genetic model) of quantitative traits were developed, including generation mean analysis (Mather, 1949; Mather and Jinks, 1989), and genetic variance and covariance analysis (Comstock and Robinson, 1948; Cockerham, 1954; Kempthorne, 1957).

A great number of genetic studies of quantitative traits, especially the studies of quantitative trait locus (QTL) marker analysis, indicated that there existed both major genes and minor genes in a quantitative trait genetic system, not necessarily all being minor genes even with equal effects. Gai and his group indicated that for a QTL system, the major gene plus minor gene model was the general model, while the pure major gene model, or pure minor gene model was the specific case of the general model (Gai et al., 2003; Gai, 2006). Accordingly, Gai and Wang (1998), Wang and Gai (1998), Gai and Zhang (2000), Zhang and Gai (2000a), and

Received October 4, 2006; accepted October 14, 2006

GAI Junyi (✉), WANG Yongjun
Soybean Research Institute of Nanjing Agricultural University, National Center for Soybean Improvement, National Key Laboratory for Crop Genetics and Germplasm Enhancement, Nanjing 210095, China
E-mail: sri@njau.edu.cn

WU Xiaolei, CHEN Shouyi (✉)
Institute of Genetics and Developmental Biology of Chinese Academy of Sciences, Beijing 100101, China
E-mail: sychen@genetics.ac.cn

Gai et al. (2003) established the procedures of segregation analysis of quantitative trait to detect the genetic system, which was firstly established for genetic analysis of qualitative traits by Mendel and could utilize the information from individuals of segregating generations. This procedure includes the following major steps (Gai et al., 2003). First, under the supposition that the segregating population was composed of component distributions controlled by major gene(s) and modified by both polygenes and environments, seven groups and 32 types of genetic models, including a one major-gene, two major-gene, three-major gene, polygene, mixed one major-gene and polygene, mixed two major-gene and polygene, and a mixed three major-gene and polygene models were set up. Second, the joint maximum-likelihood function was constructed from the tested generations, including single generation and multiple generations to estimate the parameters of component distributions through an IECM (Iterated Expectation and Conditional Maximization) algorithm. Third, the best-fitting genetic model was chosen according to Akaike's information criterion, a likelihood-ratio test and tests for goodness of fit. Fourth, the related genetic parameters, including gene effects, as well as the genetic variances of major genes and polygenes and their corresponding heritability values were calculated from the estimates of component distributions. The detailed theoretical bases from which the procedure was derived can be referred to the cited literature (Hasselblad, 1966; Akaike, 1977; Choi, 1969; Elkind and Cahaner, 1986, 1990; Dick and Bowden, 1973; Dempster et al., 1977; Jiang et al., 1994; Liu and Rubin, 1994; Wang and Gai, 1997a, 1997b; Zhang and Gai, 2000b; Zhang et al., 2001a). Segregation analysis of quantitative traits, in fact, is a procedure of genetic experiment and data analysis based on the Mendelian method, or in other words, a procedure to fit possible models with segregating data and then to pick up a best fitted one through a series of criteria and tests.

Quantitative trait locus mapping is a procedure utilizing some segregation data in linkageship analysis to locate QTLs to the nearest markers on a molecular genetic map. In comparing the segregation analysis with QTL mapping, both procedures, based on a similar set of genetic assumptions, can use the same set of segregation data, but the former can detect some major gene and polygene effects, does not need a genetic linkage map and therefore does not provide the information about the location of QTLs; while the latter can detect and locate the possible QTLs if a linkage map is available. It is obvious that the precision of both approaches depends on the precision of the experimental data, but the latter further depends on the precision of the genetic linkage map. Conventional breeders who have obtained the segregation data already can use the former, but the latter can be used when not only the segregation data but also the molecular marker data have been obtained.

From a breeder's point of view, to reveal the inheritance of a quantitative trait is mainly for recognizing some major

genes or QTLs with large effects so that the plant breeder can operate breeding procedures to converge the major part of the genes into a same individual. The objective of the present paper is to make a comparative study on segregation analysis and QTL mapping to see if both can provide similar results in recognizing major QTLs with the same set of data.

2 Materials and methods

A recombinant inbred line (RIL) population with 201 families derived from (Kefeng No.1 \times 1138-2) $F_{2:7:10}$ (derived family in F_{10} from a F_7 plant randomly obtained from a $F_{2:7}$ line) along with their parents were tested in a randomized block design experiment with 0.7 m \times 0.7 m hill-plots, eight replications at Jiangpu Station, Nanjing Agricultural University, Nanjing, China. The number of days to flowering, number of days to maturity, plant height, number of nodes on main stem, number of pods per node, 100-seed weight, protein content, oil content, and plot yield were measured.

The molecular markers, including 171 RFLP, 60 SSR and 79 AFLP markers, plus two morphological markers in a total of 312 ones, were used to mark the 201 families. The methods and procedures of the molecular marker analysis used in the study are omitted here and can be referred to Wu et al. (2001) and Wang (2001).

A set of tests, including equality test, symmetry test, and representation test, by using χ^2 criterion were designed to examine the coincidence of the practical RIL sample with the theoretical RIL population under the supposition of the distortion of the population was mainly due to the shifting environments during various seasons of generation derivation rather than the viability difference of gametes and zygotes. The detailed procedure will be given later in the context. As the results of the tests, 17 families from 201 ones and five RFLP markers from 171 ones were performed to be extra-biased segregates or outliers. After those outliers were removed, 184 families and 166 RFLP markers were left for next coincidence tests, which showed a good fit to its theoretical population.

The 184 families and 307 markers (including 166 RFLP, 60 SSR, 79 AFLP, and two morphological markers) were utilized to construct a genetic linkage map by using Mapmaker/Exp 3.0 b (Lander and Bostein, 1989; Lander et al., 1987). Among the 307 markers, except two, 305 ones linked in 25 linkage groups with a total length of 3017.9 cM among which 22 groups were corresponding to those of Cregan's integrated map (Cregan, Jarvik, Bush, Shoemaker, Lark, Kahler, Kaya, Van Toai, Lohnes, Chung, and Specht, 1999; Wang, 2001).

The segregation analysis procedure for RIL population with parents was applied to the obtained data according to Zhang et al. (2001b) and Gai et al. (2003). The same set of data was analyzed for QTL mapping by using WinQTL Cartographer (Basten et al., 1999; Zeng, 1993, 1994).

3 Results

3.1 Segregation analysis

The results from segregation analysis are shown in Tables 1 and 2. The segregation analysis procedure for RIL population with parents can detect a major gene up to three ones plus polygene as a whole. Model A means only 1 major gene, B means 2 major genes, C means only polygene, F means 3 major genes, D means 1 major gene plus polygenes, E means 2 major genes plus polygenes, and G means 3 major genes plus polygenes. For Models B and E, after the first dash “1” means without linkage between the two major genes, “2” means with linkage between the two major genes; while after the second dash, “1” means additive-additive \times additive epistasis effect of major gene, “2” means additive effect, “3” means equal additive, “4” means dominance epistasis, “5” means recessive epistasis, and “6” means duplicate epistasis. For the other models, the numbers after the first dash means the same as those after the second dash of Models B and E. At present, the linkageship can be detected only for those models with two major genes, rather than those with more than two major genes due to the very complicated situation in deriving the formulae.

For days to flowering, three major genes were detected with additive and additive \times additive epistasis, pretty high major gene heritability and pretty low polygene heritability. A similar situation was for days to maturity and number of nodes, except with larger polygene heritability. For plant height, only two major genes were detected without add.

\times add. effect, but with high major gene heritability and low polygene heritability. A similar case occurred for 100-seed weight except with add. \times add. For plot yield, pods per node and oil content, only two major genes were detected with add. \times add., medium major gene heritability, and different amount of polygene heritability. Large part of variation was due to environment for yield and number of pods per node. Anyway, no major gene was detected for protein content, but polygene accounted for a major part of genetic variation.

3.2 Quantitative trait locus mapping

The results from QTL mapping through QTL Cartographer are shown in Table 3. For days to flowering, seven QTLs were identified. The most important ones were *fd3* and *fd4* located on the C2 linkage group and the next important ones were *fd7* and *fd6* located on the F1 linkage group according to their LOD, r^2 , and additive effect values. The four QTLs accounted for about most of the total genetic variation. Thirteen QTLs were identified for days to maturity. The most important ones were *md1*, *md2*, *md3* located on linkage group B1 and *md9* located on linkage group G. They accounted for most of the genetic variation. It needs to be explained that the major QTLs of both growth period traits were not on the same linkage groups even though days to flowering being a part of days to maturity.

Twelve QTLs were detected for plant height. The most important ones were *ht6*, *ht4*, *ht5* and *ht7*, all located on C2 and accounted for most of the genetic variation. Thirteen QTLs were identified for number of nodes on the main stem.

Table 1 Genetic models of soybean traits detected from segregation analysis

Trait	Best model	Major gene	Polygene	AIC value	Alternative model
Flowering	G-0	3	Yes	1118.06	G-1
Maturity	C-0	3	Yes	1395.22	G-1
Plant height	E-2-5	2 linked	Yes (additive)	1645.98	E-1-4, E-1-4
No. nodes	G-0	3	Yes	942.28	G-1
100-seed weight	E-1-1	2	Yes (additive)	902.68	E-1-4, E-1-5, E-1-6
Plot yield	E-2-0	2 linked	Yes	1903.58	E-2-4, E-2-5
Pods per node	E-2-0	2 linked	Yes	499.91	E-2-6
Protein content	G-0	No	Yes	739.36	No
Oil content	E-2-0	2 linked	Yes	690.38	E-1-0

Table 2 Estimates of genetic parameters in segregation analysis

Trait	Additive			Additive \times Additive				Major gene		Polygene	
	d_a	d_b	d_c	i_{ab}	i_{ac}	i_{bc}	i_{abc}	σ_{mg}^2	h_{mg}^2	σ_{pg}^2	h_{pg}^2
Flowering /d	-0.89	-0.95	0.28	1.97	2.14	2.18	1.54	17.35	89.43	0.28	1.46
Maturity /d	-3.66	-7.00	0.78	4.54	1.68	1.58	5.59	120.05	73.97	30.81	18.98
Plant height /cm	0.99	-16.11	-	-	-	-	-	130.75	76.85	7.93	4.66
No. nodes	-0.58	-1.05	-0.50	1.04	0.51	0.96	1.08	5.13	76.84	0.66	9.90
100-seed weight /g	-1.68	-0.32	-	0.90	-	-	-	3.72	70.03	0.39	7.41
Plot yield /g	12.20	12.14	-	12.14	-	-	-	221.89	41.48	6.19	1.16
Pods per node	0.43	0.43	-	0.43	-	-	-	0.28	41.79	0.40	11.76
Protein content /%	-	-	-	-	-	-	-	-	-	1.83	60.60
Oil content /%	-1.62	-0.89	-	0.21	-	-	-	1.73	56.35	1.34	26.06

Table 3 QTL mapping of agronomic traits of soybeans

Trait	Linkage group ^{a)}	Locus	Marker region	cM	LOD	r^2	Additive effect
1 Flowering	N3-B1	<i>fd1</i>	Satt197—A118T	14.0–6.7	2.37	0.073	-1.667
		<i>fd2</i>	A520T	0.0	3.35	0.058	-1.492
	N6-C2	<i>fd3</i>	A397I—B131V	8.0–3.7	10.01	0.262	-3.149
		<i>fd4</i>	AGCCAC10	0.0	12.94	0.322	-3.531
		<i>fd5</i>	AGCCAC11	0.0	2.17	0.090	-1.836
	N12-F1	<i>fd6</i>	Satt586	0.0	3.50	0.090	-1.842
		<i>fd7</i>	ACGCCAC01— <i>W</i>	6.0–11.3	3.36	0.101	-1.948
2 Maturity	N3-B1	<i>md1</i>	Satt509—Satt197	18.0–4.4	4.59	0.251	-5.234
		<i>md2</i>	Satt197—A118T	14.0–6.7	12.88	0.497	-7.250
		<i>md3</i>	A118T—A520T	2.0–4.7	11.46	0.282	-5.535
	N6-C2	<i>md4</i>	A397I—B131V	2.0–9.7	2.55	0.078	-2.196
		<i>md5</i>	AGCCAC10	0.0	2.30	0.077	-2.938
	N12-F1	<i>md6</i>	Satt586	0.0	2.60	0.095	-3.210
		<i>md7</i>	ACGCAC01— <i>W</i>	6.0–11.3	2.22	0.094	-3.201
	N14-G	<i>md8</i>	AGCCAC17—Satt472	10.0–3.5	2.19	0.130	3.816
		<i>md9</i>	Satt472—K69T	14.0–17.3	3.23	0.221	4.923
		<i>md10</i>	AACCAA04	0.00	2.81	0.091	3.139
	N21-N	<i>md11</i>	LBC—ABAB	2.0–9.9	2.69	0.129	-3.737
		<i>md12</i>	AGGCTA03	0.0	2.05	0.139	-3.883
		<i>md13</i>	AAGCAT12—AAGCAT10	10.0–3.1	2.00	0.187	-4.504
3 Plant height	N3-B1	<i>ht1</i>	Satt197—A118T	10.0–10.7	2.13	0.121	-5.207
		<i>ht2</i>	A520T	0.0	2.25	0.058	-3.613
	N6-C2	<i>ht3</i>	STAS8_14T	0.0	2.32	0.059	3.802
		<i>ht4</i>	A748V—A397I	16.0–6.3	10.77	0.371	-9.069
		<i>ht5</i>	A397I—B131V	4.0–7.7	10.52	0.318	-8.393
		<i>ht6</i>	AGCCAC10	0.00	12.54	0.360	-9.070
		<i>ht7</i>	LI26T—AGCCAC02	30.0–17.7	4.42	0.486	-10.379
		<i>ht8</i>	AGCCAC11	0.0	3.04	0.186	-6.422
	N12-F1	<i>ht9</i>	Satt586	0.0	2.14	0.083	-4.278
		<i>ht10</i>	ACGCAC01— <i>W</i>	12.0–5.3	2.19	0.097	-4.638
	N13-F2	<i>ht11</i>	B174I	0.0	2.29	0.065	3.848
		<i>ht12</i>	B174I—Satt335	4.0–3.9	2.48	0.084	4.368
4 Stem nodes	N3-B1	<i>sn1</i>	A520T	0.0	2.13	0.055	-0.699
		<i>sn2</i>	A148I—AACCAT02	2.0–7.6	2.50	0.099	0.104
	N4-B2	<i>sn3</i>	A748V—A397I	16.0–6.3	9.19	0.322	-1.679
		<i>sn4</i>	A397I—B131V	6.0–5.7	9.77	0.299	-1.620
		<i>sn5</i>	AGCCAC10	0.0	10.93	0.315	-1.686
	N6-C2	<i>sn6</i>	LI26T—AGCCAC02	30.0–17.7	4.53	0.568	-2.231
		<i>sn7</i>	AGCCAC11	0.0	2.74	0.143	-1.118
		<i>sn8</i>	gmrpbp—Satt269	12.0–7.7	2.52	0.157	-1.174
		<i>sn9</i>	Satt586	0.0	3.45	0.123	0.128
		<i>sn10</i>	ACGCAC01— <i>W</i>	6.0–11.3	3.49	0.143	-1.118
		<i>sn11</i>	L37_2I—Sat036	2.0–16.5	2.31	0.090	0.890
	N12-F1	<i>sn12</i>	B174T—B174I	6.0–1.2	3.53	0.099	0.945
		<i>sn13</i>	B174I—Satt335	4.0–3.9	4.23	0.149	1.159
<i>sw1</i>		Satt509	0.0	2.25	0.082	-0.468	
5 100-seed weight	N3-B1	<i>sw2</i>	B146H—A611D	2.0–8.8	4.26	0.128	-0.581
		<i>sw3</i>	A199H—A64_3I	14.0–5.2	2.02	0.100	-0.514
	N9-D2a	<i>sw3</i>	A199H—A64_3I	14.0–5.2	2.02	0.100	-0.514
6 Plot yield	N18-K	<i>yd1</i>	STAS8_14T	0.0	2.81	0.073	4.959
		<i>yd2</i>	A748V—A397I	20.0–2.3	4.63	0.149	-6.717
		<i>yd3</i>	A397I—B131V	6.0–5.7	4.91	0.154	-6.829
	N6-C2	<i>yd4</i>	AGCCAC10	0.0	6.22	0.197	-7.834
		<i>yd5</i>	Satt586	0.0	2.36	0.085	-5.064
		<i>yd6</i>	ACGCAC01— <i>W</i>	12.0–5.3	2.44	0.081	-4.966
	N14-G	<i>yd7</i>	AACCAA04	0.0	2.04	0.065	4.451
		<i>yd8</i>	AAGCAT12—AAGCAT10	4.0–9.1	3.45	0.312	-9.730
	N21-N	<i>yd9</i>	AAGCAT05	0.0	2.14	0.178	-7.369
		<i>pn1</i>	B30T—K418_2V	6.0–1.9	2.09	0.060	0.144
		<i>pn2</i>	A748V—A397I	14.0–8.3	3.19	0.141	0.221
	7 Pods per node	N6-C2	<i>pn3</i>	A397I—B131V	8.0–3.7	4.05	0.129
<i>pn4</i>			AGCCAC10	0.0	5.19	0.181	0.254
<i>pn5</i>			K11_3T—A636_1T	4.0–7.7	3.76	0.141	0.223
N6-C2		<i>pn6</i>	LI26T—AGCCAC02	16.0–31.7	2.10	0.201	0.264
		<i>pn7</i>	Satt463	0.0	2.10	0.088	0.175
8 Protein content		N20-M	<i>pt1</i>	Satt197—A118T	12.0–8.7	2.43	0.108
	<i>pt2</i>		A118T—A520T	2.0–4.7	2.18	0.058	0.413
	N3-B1	<i>pt3</i>	A481V—A725_2V	4.0–5.9	2.21	0.067	0.441
9 Oil content	N8-D1b+W	<i>ol1</i>	A953_1H—B221T	2.0–24.8	2.20	0.091	-0.398
		<i>ol2</i>	A60V—AACCAA08	16.0–0.8	2.38	0.133	-0.482
	N4-B2	<i>ol3</i>	AACCAA08—AACCAA09	2.0–6.8	2.43	0.139	-0.491
		<i>ol4</i>	AACCAA09—AAGCAT11	4.0–12.6	2.27	0.139	-0.491

a) In N3-B1, the left part N3 designates the linkage group No. 3 of the present study, while the right part B1 is the equivalent linkage group by Cregan et al. (1999). The similar designations are for the other linkage groups.

The most important ones were *sn5*, *sn4*, *sn3*, and *sn6* located also on C2. It seems that QTLs for plant height and related traits mainly involved with C2 linkage group.

Three QTLs were detected for 100-seed weight as *sw2*, *sw1*, and *sw3* located on D2a, B1, and K linkage group, respectively. They accounted for only a small part of total genetic variation. That means most of the genetic variation might be due to polygenes. Nine QTLs were identified for plot yield. The most important ones were *yd8* on N linkage group and *yd4*, *yd3*, *yd2* on C2 linkage group, which accounted for the most part of total genetic variation. Seven QTLs were detected for number of pods per node. The most important ones were *pn4*, *pn3*, *pn5*, *pn2* and *pn6*, all located on C2. It seems that C2 is also a major linkage group for yield and related traits.

Three QTLs were identified for protein content, *pt1* and *pt2* on B1 and *pt3* on D1b+W. Four QTLs were detected for oil content, *ol2*, *ol3*, and *ol4* on M and *ol1* on B2. The detected QTLs of both traits accounted for only a relatively small part of their total genetic variation. That means most of the genetic variation for both traits might be accounted for by polygenes.

From the above results, linkage groups C2, B1, F1, M, N are more likely involved with the nine agronomic and quality traits.

3.3 Comparisons between the results from segregation analysis and QTL mapping

The comparisons are summarized in Table 4. The number of major genes from segregation analysis was two to three. It might be more than three, but the capacity of the procedure was limited to three or less since the developed models have a capacity of only up to three major genes. Assuming each QTL from mapping analysis could be equivalent to a major gene, the number of main major genes for each of the nine traits was about four. Therefore, the segregation analysis can detect most of the main major genes or QTLs and leaves the left minor effect QTLs as polygenes.

Segregation analysis provided an overall concept about the genetic system of a trait, including the major gene plus

polygene model, all kinds of genetic effects of individual major genes (additive, dominance, epistasis), all kinds of genetic effects of polygene as a whole, heritability values of individual major genes and that of entire polygenes, while QTL mapping could locate the QTLs on linkage groups but could not give the epistasis effects for the present Cartographer version of Composite Interval Mapping (CIM).

The accuracy and precision of the results from segregation analysis depended on those of the experiment, but those from QTL mapping depended not only on it, but also on those of the linkage map.

Segregation analysis is simple, needs only precise data and a corresponding computer program, and can provide plant breeders with information about the genetic system of the major breeding target traits at only a little resource consumption, while QTL mapping needs some additional conditions in molecular technological equipment and financial resources. Therefore, segregation analysis can be used independently or as a preliminary analysis of the data set before QTL mapping. Both segregation analysis and QTL mapping can be used as a mutual supplement and check.

At present, the segregation analysis procedure has been developed for up to three major-gene plus polygene mixed inheritance models. As indicated in Table 4, the analytical procedure of four major-gene plus polygene mixed inheritance models is expected to be developed, but to do so is very complicated. Unfortunately, it is really tedious for developing analytical procedures of models with four major genes and polygenes since even a four Mendelian gene analysis alone is complicated enough.

3.4 Test and correction for coincidence of experimental sample with theoretical population

It was indicated above that the experiment sample was adjusted from 201 to 184 families. The results of segregation analysis, map construction and QTL mapping from the adjusted data set and unadjusted data set were quite different, which indicated that a test and correction for coincidence of the practical sample with the theoretical population was really necessary.

Table 4 Comparisons between the results from segregation analysis and QTL mapping

Trait	Segregation analysis			QTL mapping		
	Major gene	h_{mg}^2	h_{pg}^2	QTL	Variation explained ^{b)}	Linkage group
Flowering	3	89.4	1.5	<i>fd4</i> , <i>fd3</i> (<i>fd7</i> , <i>fd6</i>) ^{a)}	58.4%	C2(1st two) F1(2nd two)
Maturity	3	74.0	19.0	<i>md2</i> , <i>md3</i> , <i>md1</i> , <i>md9</i>	125.1%	B1(1st three)
Plant height	2, linked	76.9	4.7	<i>ht6</i> , <i>ht4</i> , <i>ht5</i> , <i>ht7</i>	153.5%	C2(all four)
Stem nodes	3	76.9	9.9	<i>sn5</i> , <i>sn4</i> , <i>sn3</i> , <i>sn6</i>	150.4%	C2(all four)
100-seed weight	2	70.0	7.4	<i>sw2</i> , (<i>sw1</i> , <i>sw3</i>)	12.8%	
Plot yield	2, linked	41.5	1.2	<i>yd4</i> , <i>yd3</i> , <i>yd2</i> , <i>yd8</i>	81.2%	C2(1st three)
Nodes per pod	2, linked	41.8	11.8	<i>pn4</i> , <i>pn3</i> , <i>pn5</i> , <i>pn2</i>	59.2%	C2(all four)
Protein content	No	–	60.6	(<i>pt1</i> , <i>pt3</i> , <i>pt2</i>)		
Oil content	2, linked	56.4	26.1	<i>ol3</i> , <i>ol2</i> , <i>ol4</i>	41.1%	M(all three)

a) Genes with relatively small effect are in the parentheses.

b) Variation explained is r^2 of the genes not included in parentheses; those of genes with tight linkages and /or with interaction might be larger than 100%.

The coincidence test was designed as including three sets of χ^2 tests: (1) equality test, to test whether the germplasm (markers) from both parents are equal, or $p(P_1) : p(P_2) = 1 : 1$; (2) symmetry test, to test whether the families with $p(P_1) : p(P_2) > 1 : 1$ and the families with $p(P_1) : p(P_2) < 1 : 1$ are equal (in a symmetry distribution); (3) representation test, to test whether each family as well as the whole experiment sample is a random sample from the corresponding theoretical population. To do the representation test, two steps were taken. The first step was to test each family to see if it was an extra-biased family ($\chi^2 < \chi^2_{0.05}$). The second step was to look at the rate of extra-biased family to see if it was less than the expected rate obtained from a procedure of sampling the simulated population, called Simulated Population Sampling Criteria (SPSC). If the rate was larger than the SPSC rate, the extra-biased families should be checked and deleted one by one until the adjusted sample fitting the SPSC rate. The simulation procedure was referred to Tanksley and Nelson (1996). The developed software for SPSC was named GenoSim. The markers were checked, tested, and adjusted in a similar way until the rate of extra-ordinary biased marker fitted the SPSC rate.

Table 5 shows the results of coincidence test of the NJRIKY population before and after adjustment. The equality test of NJRIKY population before adjustment showed a very large χ^2_c value (30.59), indicating not equal genetic contribution from both parents; and that after adjustment showed a very small χ^2_c value (0.10), indicating a good fit after adjustment.

The symmetry test of the unadjusted NJRIKY fitted basically a 1:1 ratio ($\chi^2_c = 3.16$, less than 3.86), but after adjustment the χ^2 tests showed a better fit to 1:1 ratio ($\chi^2_c = 0.92$).

The representation test of the unadjusted NJRIKY did not fit the SPSC requirements. According to the SPSC, the

critical value of the rate of extra-biased marker should be less than 20.36% and that of family should be less than 24.47%. Unfortunately, those of the unadjusted NJRIKY were 21.05% and 29.35%, respectively. After the five most extra-biased markers and 17 most extra-biased families (according to their χ^2_c values) were deleted, the two rates of the adjusted NJRIKY became 19.28% and 24.45%, respectively, less than the critical values, which therefore fitted the SPSC requirements.

4 Discussion

Genetic information about breeding target traits, especially those of quantitative traits, is extremely important to plant breeders in designing their breeding procedures, choosing parents for crossing, progeny selection, gene pyramiding, etc. Segregation analysis can provide genetic information on the number of major genes, their kinds of genetic effects, heritability values as well as genetic information on all kinds of genetic effects and heritability value of whole polygenes without any extra requirements on lab conditions except a precise experiment. Therefore, it is a simple and useful tool in the plant breeder's hands.

Quantitative trait locus mapping is an advanced tool for plant breeders if they have a molecular biological lab or have a molecular geneticist cooperating with them. Based on QTL mapping, marker-assisted selection can be used for effective and efficient selection of quantitative traits. It is suggested to conduct segregation analysis first before QTL mapping so that plant breeders can have a first impression on the genetic system of the involved trait. Both segregation analysis and QTL mapping can be used as a check for each other.

Table 5 Coincidence test of the NJRIKY population before and after adjustment

Test			Before	After	
Equality	Total number of effective loci	Kefenf No. 1	14141	12113	
		1138-2	13225	12164	
		χ^2_c	30.59	0.10	
Symmetry	Family number	$p > 0.5, q < 0.5$	112	98	
		$P < 0.5, q > 0.5$	86	84	
		χ^2_c	3.16	0.93	
		$\chi^2_c < 3.84$	$p > 0.5, q < 0.5$	75	75
			$P < 0.5, q > 0.5$	64	62
	$\chi^2_c > 3.84$	χ^2_c	0.72	1.05	
		$p > 0.5$	37	23	
		$q > 0.5$	22	22	
		χ^2_c	3.32	0	
		$p = q = 0.5$	3	2	
Test			Before	After	Simulated critical value ($\alpha = 0.05$)
Representation	RFLP number		171	166	
	Family number		201	184	
	Marker extra-biased rate		21.05%	19.28%	20.36%
	Marker largest χ^2_c		38.56	23.08	23.75
	Family extra-biased rate		29.35%	24.45%	24.47%
	Family largest χ^2_c		144.06	12.45	28.35

Both the segregation analysis and QTL mapping procedures need to be further improved and completed. Since the accuracy and precision are related with the experiment design, replicated test of lines or families are preferred for both procedures. For segregation analysis, the analytical procedure of the four major-gene plus polygene mixed inheritance model needs to be developed, linkage between more than two genes should be considered, and procedures of more segregating generations to resolve more estimates of genetic parameters should be studied. For QTL mapping, ghost problems and noises among QTLs need to be further resolved and a procedure for the estimation of epistasis effects should be considered.

Acknowledgements The project was supported by the National Natural Science Foundation of China (No. 30490250), the National Key Basic Research Program (No. 2002CB111304, No. 2004CB7206, No. 2006CB101708), the National “863” Program (No. 2002AA211052, 2006AA100104) and the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT).

References

- Akaike H (1977). On entropy maximum principle. In: Krishnaiah P R, ed. Applications of Statistics. Amsterdam: North-Holland Publishing Company, 27–41
- Basten C J, Weir B S, Zeng Z B (1999). QTL Cartographer. Version 1.13. Raleigh (NC): Department of Statistics, North Carolina State University
- Choi K (1969). Estimators for the parameters of distributions. *Ann Inst Statist Math*, 21: 107–116
- Cockerham C C (1954). An extension of the concept of partitioning hereditary variance for analysis of covariance among relatives when epistasis is present. *Genetics*, 39: 859–882
- Comstock R C, Robinson H F (1948). The component of quantitative variance in populations. *Biometrics*, 4: 254–266
- Cregan P B, Jarvik T, Bush A L, Shoemaker R C, Lark K G, Kahler A L, Kaya N, Van Toai T T, Lohnes D G, Chung J, Specht A L (1999). An integrated genetic linkage map of the soybean genome. *Crop Sci*, 39: 1464–1490
- Dempster A P, Laird N M, Robin D B (1977). Maximum likelihood from incomplete data via the EM algorithm. *J R Statist Soc B*, 39: 1–38
- Dick N P, Bowden D C (1973). Maximum likelihood estimation for mixtures of two normal distributions. *Biometrics*, 29: 781–790
- Elkind Y, Cahaner A (1986). A mixed model for the effects of single gene, polygenes and their interaction on quantitative traits. 1. The model and experimental design. *Theor Appl Genet*, 72: 377–383
- Elkind Y, Cahaner A (1990). A mixed model for the effects of single gene, polygenes and their interaction on quantitative traits. 2. The effects of the major gene and polygenes on tomato fruit softness. *Heredity*, 64: 205–213
- Fisher R A (1918). The correlations between relatives on the supposition of Mendelian inheritance. *Trans Roy Soc Edin*, 52: 399–433
- Gai J Y, Wang J (1998). Identification and estimation of a QTL model and its effects. *Theor Appl Genet*, 97(7): 1162–1168
- Gai J Y, Zhang Y, Wang J (2000). A joint analysis of multiple generations for QTL models extended to mixed two major genes plus polygene. *Acta Agronomica Sinica*, 26(4): 385–391 (in Chinese)
- Gai J Y, Zhang Y, Wang J (2003). Genetic System of Quantitative Traits in Plants. Beijing: Academic Press (in Chinese)
- Gai J Y (2006). Segregation analysis on genetic system of quantitative traits in plants. *Front Biol China*, 1: 85–92
- Hasselblad V (1966). Estimation of parameters for a mixture of normal distributions. *Technometrics*, 8: 431–444
- Jiang C, Pan X, Gu M (1994). The use of mixture models to detect effects of major genes on quantitative characters in plant breeding experiment. *Genetics*, 136: 383–394
- Kempthorne O (1957). An Introduction to Genetics Statistics. New York: Wiley
- Lander E S, Bostein D R (1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage map. *Genetics*, 121: 185–189
- Lander E S, Green P, Abrahamson J, Barlow A, Daly M J, Lincoln S E, Newburg L (1987). Mapmaker: An interactive computer package for constructing genetic linkage maps of experimental and natural populations. *Genomics*, 1: 174–181
- Liu C, Rubin D R (1994). The ECME algorithm: A simple extension of ECM with faster monotone convergence. *Biometrika*, 81(4): 633–648
- Mather K (1949). Biometrical Genetics. London: Methum
- Mather K, Jinks J L (1989). Biometrical Genetics. 3rd ed. London: Chapman and Hall
- Tanksley S D, Nelson J C (1996). Advanced backcross QTL analysis: A method for the simultaneous discovery and transfer of valuable QTLs from unadapted germplasm into elite breeding lines. *Theor Appl Genet*, 92: 191–203
- Wang J, Gai J Y (1997a). Identification of major-polygene mixed inheritance model of quantitative traits from F_2 population. *Acta Genetica Sinica*, 24(3): 181–190 (in Chinese)
- Wang J, Gai J Y (1997b). EM algorithm in the analysis of major gene and polygene mixed inheritance. *Journal of Biomathematics*, 12(5): 540–548 (in Chinese)
- Wang J, Gai J Y (1998). Identification of major gene and polygene mixed inheritance model of quantitative traits by using joint analysis of P_1 , F_1 , P_2 , F_2 and $F_{2,3}$. *Acta Agronomica Sinica*, 24(6): 651–659 (in Chinese)
- Wang Y (2001). Establishment and adjustment of RIL population and its application to map construction, mapping genes resistant to SMV, and QTL analysis of agronomic and quality traits in soybeans. Dissertation for the Doctoral Degree. Nanjing: Nanjing Agricultural University (in Chinese)
- Zeng Z B (1993). Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. *Proc Natl Sci USA*, 90: 10972–10976
- Zeng Z B (1994). Precision mapping of quantitative trait loci. *Genetics*, 136: 1457–1468
- Wu X L, He C Y, Wang Y J, Chen S Y, Gai J Y, Wang S C (2001). Construction and analysis of a genetic linkage map of soybean. *Acta Genetica Sinica*, 28(11): 1051–1061 (in Chinese)
- Zhang Y, Gai J Y (2000a). Identification of mixed major-gene and polygene inheritance model of quantitative traits by using DH or RIL population. *Acta Genetica Sinica*, 27(7): 634–640 (in Chinese)
- Zhang Y, Gai J Y (2000b). The IECM algorithm for estimation of component distribution parameters in segregating analysis of quantitative traits. *Acta Agronomica Sinica*, 26(6): 699–706 (in Chinese)
- Zhang Y, Gai J Y, Qi C (2001a). The precision of segregating analysis of quantitative trait and its improving methods. *Acta Agronomica Sinica*, 27(6): 787–793 (in Chinese)
- Zhang Y, Gai J Y, Wang Y (2001b). An expansion of joint segregation analysis of quantitative trait for using P_1 , P_2 and DH or RIL populations. *Hereditas (Beijing)*, 23(5): 467–470 (in Chinese)