Theor Appl Genet (2006) 113: 39–48
DOI 10.1007/s00122-006-0268-4

ORIGINAL PAPER

B. Guo · D. A. Sleper · J. Sun · H. T. Nguyen
P. R. Arelli · J. G. Shannon

Pooled analysis of data from multiple quantitative trait locus mapping populations

Received: 22 September 2005 / Accepted: 17 March 2006 / Published online: 20 April 2006

Abstract Quantitative trait locus (QTL) analysis on pooled data from multiple populations (pooled analysis) provides a means for evaluating, as a whole, evidence for existence of a QTL from different studies and examining differences in gene effect of a QTL among different populations. Objectives of this study were to: (1) develop a method for pooled analysis and (2) conduct pooled analysis on data from two soybean mapping populations. Least square interval mapping was extended for pooled analysis by inclusion of populations and cofactor markers as indicator variables and covariate variables separately in the multiple linear models. The general linear test approach was applied for detecting a QTL. Single population-based and pooled analyses were conducted on data from two $F_{2:3}$ mapping populations, Hamilton (susceptible) × PI 90763 (resistant) and Magellan (susceptible) × PI 404198A (resistant), for resistance to soybean cyst nematode (SCN) in soybean. It was demonstrated that where a QTL was shared among populations, pooled analysis showed increased LOD values on the QTL candidate region over single population analyses. Where a QTL was not shared among populations, however, the pooled analysis showed decreased LOD values on the QTL candidate region over single population analyses. Pooled analysis on data from genetically similar populations may have higher power of QTL detection than single population-based analyses. QTLs were identified by pooled analysis on linkage groups (LGs) G, B1 and J for resistance to SCN race 2 whereas QTLs on LGs G, B1 and E for resistance to SCN race 5 in soybean PI 90763 and PI 404198A. QTLs on LG G and B1 were identified in both PI 90763 and PI 404198A whereas QTLs on LG E and J were identified in PI 90763 only. QTLs on LGs G and B1 for resistance to race 2 may be the same or closely linked with QTLs on LG G and B1 for resistance to race 5, respectively. It was further demonstrated that QTLs on G and B1 carried by PI 90763 were not significantly different in gene effect from QTLs on LGs G and B1 in PI 404198A, respectively.

Communicated by S. J. Knapp

B. Guo · D. A. Sleper (✉) · H. T. Nguyen · J. G. Shannon
Division of Plant Sciences and National Center for Soybean Biotechnology, 271-F Life Sciences Center,
University of Missouri-Columbia, Columbia
MO 65211-7310, USA
E-mail: SleperD@missouri.edu
Tel.: +1-573-8827320

J. Sun
Department of Statistics, University of Missouri-Columbia, Columbia, MO 65211, USA

P. R. Arelli
USDA-ARS-MSA, 605 Airways Blvd, Jackson, TN 38301, USA

Introduction

Very often, more than two mapping populations are studied for the same traits or related traits. Analysis on pooled data from multiple mapping populations (pooled analysis) was suggested by Lander and Kruglyak (1995). Pooled analysis is a good method for evaluating the overall evidence for existence of a quantitative trait locus (QTL) on a region or a linkage group (LG) from different studies where the results may be conflicting (Lander and Kruglyak 1995). It can also be used for statistically examining differences in gene effect of a QTL among different lines (populations) (Walling et al. 2000; Li et al. 2005) whereas single population-based QTL analyses do not provide direct comparisons of a QTL among different populations. Information on differences of a QTL among different lines are useful for selection of parents in breeding and will aid for a novel strategy (multiple cross mapping) of QTL cloning (see Discussions). In addition, pooled analysis is expected to increase the power and precision of QTL detection (Walling et al. 2000; Heo et al. 2001).

Single population-based statistical methods are well developed for QTL analysis (Lander and Bostein 1989; Zeng 1994; Haley and Knott 1992; Sen and Churchill 2001). Interval mapping (IM) (Lander and Bostein 1989) and composite interval mapping (CIM) (Zeng 1994) are the commonly used methods. Both methods use flanking markers for delimiting one putative QTL and the maximum likelihood for estimating mapping parameters but CIM also uses cofactor markers to reduce the interfering effects on QTL analysis of QTLs located elsewhere on the genome. All QTL mapping data analyses were based on single populations, with few exceptions from animal species where pooled analysis was conducted (Walling et al. 2000; Li et al. 2005). Walling et al. (2000) extended least square interval mapping (LSIM) (Haley et al. 1994) for analysis on combined data from seven porcine populations. Li et al. (2005) extended the Bayesian QTL analysis method (Sen and Churchill 2001) for analysis on combined data from four mouse populations. The former one (Walling et al. 2000) is simple in computation and general statistical software such as SAS is applicable. The latter one (Li et al. 2005) adopted a new QTL analysis method and requires special software. Some earlier studies (Rebai and Goffinet 1993; Xu 1998; Liu and Zeng 2000) developed QTL analysis methods for data which may be produced from several populations. These studies, however, did not handle key issues that pooled analysis faces such as: (1) different sets of molecular markers used in different populations, with some markers in common, and (2) phenotypic data collected under different conditions. All these studies analyzed computer-simulated data using their methods.

Least squares interval mapping was originally developed for analyses on single populations from crosses between inbred lines (Haley and Knott 1992) and from crosses between outbred lines (Haley et al. 1994) for increased efficiency in computation. It also uses flanking markers for delimiting one putative QTL but uses the least square method instead of the maximum likelihood method for estimating mapping parameters. Walling et al. (2000) extended LSIM (Haley et al. 1994) for pooled analysis. But they did not discuss the genetic basis of pooled analysis and did not clearly give the formula of the test statistics. The power and precision of QTL detection may be compromised and results may be biased due to ignorance of the interfering effects of QTLs located elsewhere on the genome (Jansen 1993; Zeng 1993).

Objectives of this study were to: (1) extend LSIM by Haley and Knott (1992) for analysis on pooled data from multiple populations by inclusion of populations and cofactor markers as indicator variables and covariate variables separately in multiple linear regression models, and (2) conduct pooled analysis, as a demonstration example, on data from two soybean mapping populations for resistance to SCN.

Development of methods

Development of linear regression models and least square interval mapping pooled analysis: two populations

In this section, we consider two mapping populations (each of them are F_2s from crosses between two inbred lines) and assume that a quantitative trait is controlled by one gene. In the following sections, we will extend to more than two populations and assume that a quantitative trait is controlled by polygenes. Suppose that we want to test for a QTL at one position flanked by two markers on a composite linkage map (see below). Let the alternative alleles of QTL be Q and q in the first population and genetic values of QQ, Qq and qq be $\alpha + a$, $\alpha + d$ and be $\alpha - a$, respectively, where α is the mid-parent value, a is the additive gene effect and d is the dominant effect. Let the frequencies of QQ, Qq and qq be separately p_1, p_2 and p_3 given the genotype of flanking markers in the F_2 generation. The expected genotypic value (mean) of individuals given the genotype of flanking markers can be written as $g = \alpha + aX_a + dX_d$, where $X_a = p_1 - p_3$ and $X_d = p_2$. X_a and X_d are known for a given position of a putative QTL and can be obtained using the formulae described in column 1 and column 2 of Table 1 in Haley and Knott's (1992) paper. Now assume that one QTL also exists at the same position on the composite linkage map in the second population (but likely it is flanked by different markers) and their alternative alleles are Q′ and q′. Let genetic values of Q′ Q′, Q′q′ and q′q′ in the second population be $(\alpha + \beta) + (a + \delta_a)$, $(\alpha + \beta) + (d + \delta_d)$ and $(\alpha + \beta) - (a + \delta_a)$, respectively, where β, δ_a and δ_d are differences for mid-parent value, additive and dominant effects between the first population and the second population. The expected genetic value in the second population can be written as $g = (\alpha + \beta) + (a + \delta_a)X_a + (d + \delta_d)X_d$. We can write the statistical model as

$$Y_{ji} = \alpha + aX_a + dX_d + \beta X_1 + \delta_a X_a X_1 + \delta_d X_d X_1 + \varepsilon_{ji} \quad \text{(F)}$$

where Y_{ji} is the phenotype of the ith individual in the jth population. Here, $j = 1, 2$. X_a, X_d, α, β, a, d, δ_a and δ_d are the same as described above. It is noted that X_a is constant over generations and X_d will be decreased by half every generation. This gives a desirable characteristic: the following test statistic F* (T1) is not affected by generation in which phenotyping is conducted if phenotypic data is collected from the same generation for different populations. This characteristic is useful, because in practical QTL mapping F_2s are often used for molecular marker genotyping and their $F_{2:3}$ or later generations for phenotyping. X_1 is an indicator variable, taking $X_1 = 1$ if the second population and 0 otherwise (here, the first population). It is assumed that ε_{ji} is an identically and independently distributed normal variable with mean zero and variance σ^2. No epistasis is assumed.

Table 1 Single population analyses using least square composite interval mapping

SCN races	Population	Linkage groups	Position[a]	1 LOD CI[b]	LOD	R^2 (%)[c]
II	Hamilton × PI 90763	G	0.0	0.0–6.0	6.5[g]	12.0
		B1	118.5	106.5–122.5[d]	3.2[f]	5.6
		J	77.5	63.5–89.5[d]	5.1[g]	9.2
	Magellan × PI 404198A	G	2.0	0.0–12.0	5.4[g]	9.8
		B1	116.5	104.5–122.5[d]	5.5[g]	9.9
V	Hamilton × PI 90763	G	2.0	0.0–10.0	4.9[g]	9.0
		B1	110.5	92.5–120.5	5.8[g]	10.3
		E[e]	39.3	25.3–55.3	5.6[g]	10.3
	Magellan × PI 404198A	G	0.0	0.0–30.0	2.5	4.8
		B1	112.5	104.5–122.5[d]	6.5[g]	12.4
		N	52.3	38.3–60.3	3.0[f]	5.5

[a]Distance from the start of linkage group on the soybean composite linkage map
[b]1-LOD confidence interval
[c]A proportion of total variation explained by a QTL given the other QTLs (cofactor markers). It is defined as SSE(F)−SSE(R)/SSTO, where SSE(F) is error sum of square associated with the full model (QTL exist at the putative position), SSE(R) error sum of square associated with reduced model (no QTL exist at the putative position), and SSTO total sum of square
[d]The end of the part of a linkage group that was searched for a QTL in this study
[e]Two peaks occurred on this chromosome but their 1-LOD confidence intervals overlapped substantially (Fig. 2). One QTL was declared for the peak with the highest LOD on this chromosome. The 1-LOD confidence interval covered the 1-LOD confidence intervals of both peaks. In order to exclude or confirm that closely linked QTLs may exist, fine mapping is needed
[f]Suggestive QTL (LOD ≥ 3.0, genome-wise type I error = 0.63)
[g]Significant QTL (LOD ≥ 4.0, genome-wise type I error = 0.05)

Assuming no QTL exists at the putative position in any population, i.e., $a = d = \delta_a = \delta_d = 0$, model (F) will be reduced to:

$$Y_{ji} = \alpha + \beta X_1 + \varepsilon_{ji} \quad \text{(R1)}$$

Assuming no QTL exists at the putative position in the first population, i.e., $a = d = 0$, the model (F) will be reduced to:

$$Y_{ji} = \alpha + \beta X_1 + \delta_a X_a X_1 + \delta_d X_d X_1 + \varepsilon_{ji} \quad \text{(R2)}$$

Assuming no QTL exists at the putative position in the second population, i.e., $a + \delta_a = d + \delta_d = 0$, the model (F) will be reduced to:

$$Y_{ji} = \alpha + \beta X_2 + a X_a X_2 + d X_d X_2 + \varepsilon_{ji} \quad \text{(R3)}$$

where $X_2 = 1$ if the first population and $X_2 = 0$ otherwise (here, the second population).

Assuming QTL has the same effect in both populations, i.e., $\delta_a = \delta_d = 0$, the model (F) will be reduced to:

$$Y_{ji} = \alpha + \beta X_1 + a X_a + d X_d + \varepsilon_{ji} \quad \text{(R4)}$$

Models (R1), (R2), (R3) and (R4) are called reduced models (R) whereas model (F) is called a full model.

The general linear test approach (Neter et al. 1996) is used to detect a QTL through comparison of the full model (F) with the reduced models (R). Comparison of the full model (F) with the reduced model (R1) (F vs. R1) is to test if a QTL exists at a putative position. Comparison of the full model with the reduced model (R2) or (R3) (F vs. R2, F vs. R3) is to provide a test for existence of a QTL in a specific population. Comparison of the full model with the reduced model (R4) (F vs. R4) is to determine if QTLs carried by different populations have a difference in gene effect. Test statistic for comparison of any two models (F vs. R) is given by the below formula:

$$\mathrm{F}^* = \frac{\mathrm{SSE(R)} - \mathrm{SSE(F)}}{df_R - df_F} \Big/ \frac{\mathrm{SSE(F)}}{df_F} \quad \text{(T1)}$$

where SSE(F) and SSE(R) are error sum of squares associated with the full model and reduced model separately and df_{F} and df_{R} are the degrees of freedom associated with SSE(F) and SSE(R), respectively. F* follows the F distribution when the reduced model holds, with the numerator degrees of freedom df_{R}-df_{F} and denominator degrees of freedom df_{F}. We recommend that the *P*-value corresponding to the observed F* value be used for reporting statistical evidence of existence of a QTL. The below equivalent means instead of directly reporting *P*-values can be used:

$$\mathrm{LR} = -2\log(P) \quad \text{(T2)}$$
$$\mathrm{LOD} = -2\log(P)/4.6 \quad \text{(T3)}$$

where log is the natural logarithm and *P* is the probability value corresponding to the observed F* value. LR follows the Chi square distribution with degrees of freedom of 2. The above LR and LOD are comparable separately to the LR (likelihood ratio, 2 *df*) and LOD (logarithm of odds ratio, 2 *df*) that are usually used in IM and CIM, because LR (4.6 LOD) obtained using IM and CIM also follows the Chi square distribution with degrees of freedom of 2.

Like IM and CIM, the above analysis (referred to as LSIM pooled analysis) is conducted in a series of positions along the genome. The position with the largest LR or LOD on a region or a whole chromosome is used for giving the most likely position of a QTL. If LOD or LR is greater than or equal to a pre-determined threshold, a QTL is declared.

Extension for more than two populations

In case of more than two populations, the above multiple linear models can easily be extended and the same analysis can be conducted as above. k populations are represented by k indicator variables $X_1,...X_j,...X_k$, X_j taking on the value 1 if observations (progeny individuals) come from the jth population and 0 otherwise. The full model can be written as:

$$\begin{aligned} Y_{ji} = {} & \alpha + \beta_2 X_2 \cdots + \beta_j X_j + \cdots + \beta_k X_k + aX_a + dX_d \\ & + \delta_{2,a} X_a X_2 + \delta_{2,d} X_d X_2 + \cdots \delta_{j,a} X_a X_j + \delta_{j,d} X_d X_j \\ & + \cdots + \delta_{k,a} X_a X_k + \delta_{k,d} X_d X_k + \varepsilon_{ji} \end{aligned} \quad (\mathrm{F}')$$

where the first and second subscripts of δ are the population and additive effect (a) or dominant effect (d), respectively.

Assuming that no populations have a QTL at the putative position, i.e., $a = d = \delta_{2,a} = \delta_{2,d} = \cdots = \delta_{j,a} = \delta_{j,d} = \delta_{k,a} = \delta_{k,d} = 0$, the model (F′) will be reduced to:

$$Y_{ji} = \alpha + \beta_2 X_2 \cdots + \beta_j X_j + \cdots + \beta_k X_k + \varepsilon_{ji}$$

We can also write model (F′) as

$$\begin{aligned} Y_{ji} = {} & \alpha + \beta_2 X_2 \cdots + \beta_j X_j \cdots + \beta_k X_k + a_1 X_a X_1 + d_1 X_d X_1 \\ & + a_2 X_a X_2 + d_2 X_d X_2 + \cdots a_j X_a X_j + d_j X_d X_j \\ & + \cdots + a_k X_a X_k + d_k X_d X_k + \varepsilon_{ji} \end{aligned} \quad (\mathrm{F}'')$$

where a and d are additive and dominant effects, respectively, with the subscripts of a and d associated with the corresponding population.

We can write models where it is assumed that some populations have a QTL but other populations do not, just by taking off from the model (F″) the terms $a_j X_a X_j + d_j X_d X_j$ corresponding to the populations which are assumed to have no QTL.

We can also write models where it is assumed that the QTL has the same effect in some populations (say, $j,...,k$). Let $X_c = X_j + \cdots + X_k$. The model (F′) will be reduced to:

$$\begin{aligned} Y_{ji} = {} & \alpha + \beta_2 X_2 \cdots + \beta_j X_j + \beta_k X_k + aX_a + dX_d + \delta_{2a} X_a X_2 \\ & + \delta_{2d} X_d X_2 + \cdots \delta_{j-1,a} X_a X_{j-1} + \delta_{j-1,d} X_d X_{j-1} \\ & + \delta_{c,a} X_a X_c + \delta_{c,d} X_d X_c + \varepsilon_{ji} \end{aligned}$$

We can write the other models we would expect to test using the same approach.

Multiple QTLs: least square composite interval mapping pooled analysis

A quantitative trait is controlled by polygenes. Detection of a putative QTL at one position is affected by QTLs located elsewhere on the genome (Haley and Knott 1992). It was demonstrated that the gene effect of one QTL can be absorbed by its linked markers (Jansen 1993; Zeng 1993). Cofactor markers have been used to reduce the interfering effects of QTLs located elsewhere on the genome in CIM (Zeng 1994). Similarly, cofactor marker terms can be added to the multiple linear models of LSIM pooled analysis described above. LSIM pooled analysis with inclusion of cofactor marker terms is referred to as least square composite interval mapping (LSCIM) pooled analysis. Principles and methods for selection of cofactor markers used in CIM (Zeng 1994; Basten et al. 2002) can be applied. However, we recommend that QTL-linked molecular markers detected separately for individual populations using IM or CIM be used as cofactor markers in the models, because too many cofactor markers will reduce the degrees of freedom associated with error sums of squares and may offset reduction of error sum of squares due to introduction of cofactor markers. Different cofactor markers may be used in different populations, because different QTLs may exist in different populations. The cofactor marker terms can be defined as $\sum_j^k \sum_l^m b_{jl} M_{jli} X_j$, where b_{jl} is the partial regression coefficient of phenotype Y_{ji} on the lth marker in the the jth population; M_{jli} is a known coefficient of the lth cofactor marker in the ith individual of the jth population, taking 1 for one homozygous genotype, 0 for the heterozygous genotype and −1 for the other homozygous genotype for F_2 populations; X_j is population indicator variable, as described above. The terms are added to the full models and reduced models of LSIM. Like CIM, a region of a length (window size) around the putative QTL position being tested can be set so that no markers in this region can be included as cofactor markers in the multiple linear models (Basten et al. 2002).

Merging data from multiple populations

Different sets of molecular markers may be used in different populations, with some molecular markers in common. This gives two challenges to pooled analysis: (1) different linkage maps will be produced in different populations, and (2) molecular marker information from multiple populations cannot be directly merged for pooled analysis. Pooled analysis is based on a composite linkage map which is created using data from several mapping populations (Stam 1993). Correct relative ordering of markers is crucial for pooled analysis (Li et al. 2005). In order to merge data from multiple populations, the coefficients (X_a and X_d) of the additive and dominant effects (see above) should be computed at the same series of positions along the composite linkage map for each population.

Stabilization of residual variances among populations

One important assumption for pooled analysis is that the error term (ε_{ji}) is identical among populations. The pooled analysis is sensitive to violation of this assumption (see below). Two factors may contribute to the

violation: (1) phenotypic data may be collected from different laboratories or conditions, and (2) different populations may have different segregating QTLs. If the error term is not equal, transformations (square root, log and other transformations) should be applied (Li et al. 2005). If transformed data fails to equalize the error terms; or if the original data are obtained using different methods or procedures; or if different populations are evaluated under different conditions; phenotypic data can be standardized to residual standard deviation units (Walling et al. 2000).

Soybean mapping data analysis

Soybean mapping data

Data were collected from $F_{2:3}$ families of crosses between 'Hamilton' and plant introduction (PI) 90763 (Guo et al. 2005) and between 'Magellan' and PI 404198A (Guo et al. 2006a) for resistance to soybean cyst nematode (*Heterodera glycines Ichinohe*) (SCN) in soybean (*Glycine max (L.) Merr*). PI 90763 and PI 404198A are resistant to SCN races 2 and 5. Hamilton and Magellan are susceptible to all known SCN races. Molecular marker data from Hamilton × PI 90763 and Magellan × PI 404198A populations consisted of 176 and 182 co-dominant SSR loci, respectively. Both populations had 74 molecular markers in common, ranging from 1 to 7 for each LG. Phenotypic data for reaction to SCN races 2 (HG type 1.2.5.7, PA 2) and 5 (HG type 2.5.7, PA 5) were collected under controlled conditions in the greenhouse at the University of Missouri-Columbia using the procedure described by Arelli et al. (1997). The female index (Schmitt and Shannon 1992) was used to measure reaction of soybean plants to SCN races 2 and 5.

Data analysis

We checked the residual variances of two mapping populations through regression of phenotypes on QTL-linked markers detected in individual populations (Table 5). We failed to stabilize the residual variances between the two soybean populations using square root and log transformations. Phenotypic values were standardized for pooled analysis in residual standard deviation units for each population according to Walling et al. (2000).

The below models were considered for pooled analysis: (1) a QTL exists in both populations (F), (2) no QTL exists in any populations (R1), (3) a QTL exists in Hamilton × PI90763 population only (R2), (4) a QTL exists in Magellan × PI404198A only (R3), and (5) a QTL has the same effect in both populations (R4). F vs. R1 analysis, F vs. R2 analysis, F vs. R3 analysis and F vs. R4 analysis were conducted using LSCIM and LSIM. QTL-linked molecular markers detected in individual populations using CIM (Table 5) were used for cofactors in the LSCIM pooled analysis. In addition, two populations were individually analyzed using LSCIM.

The soybean composite linkage map (Song et al. 2004) was used for the analysis. The coefficients (X_a and X_d) were computed at the same sets of positions (every 2 cM) along the map for Hamilton × PI 90763 and Magellan × PI 404198A populations. All LGs except for D1a were searched, with a total genome length of 1676 cM (66% of the whole genome). LG D1a was not searched because of few SSRs.

The formula (T3), i.e., LOD, was used for reporting the statistical evidence for a QTL. LODs = 3.0 and 4.0 were used for declaring suggestive QTLs and significant QTLs, respectively. These threshold values had been used for declaring a QTL in our previous studies (Guo et al. 2005, 2006a) where single population analyses on Hamilton × PI 90763 and Magellan × PI 404198A populations were conducted using CIM. They were approximate to genome-wide type I error = 0.63 (the suggestive level) and 0.05 (the significant level), respectively. A suggestive level often gives false positive QTL but is worth reporting if accompanied with an appropriate warning label, so that discovery of a QTL may not be delayed. A QTL is usually declared at genome-wide type I error = 0.05 (Lander and Kruglyak 1995).

Results

Soybean cyst nematode is the most important pest of soybean in the world (Wrather et al. 1995, 2001). QTL analysis have been extensively studied for resistance to SCN in soybean (see summaries by Concibido et al. 2004; Guo et al. 2006b). But previous studies used single population-based methods. In this section, we demonstrated pooled analysis using the two available populations. SCN populations are described in two ways. One is the race determination test (Schmitt and Shannon 1992). The other is the HG type classification system recently published by Niblack et al. (2002). For convenience of comparison to earlier studies, the former one was used below.

Single population analyses using LSCIM

In order to be compared with the pooled analysis below, the analyses on Hamilton × PI 90763 and Magellan × PI 404198A populations were individually conducted using LSCIM. QTLs were found on LGs G, B1 and J in soybean PI 90763 and on LGs G and B1 in soybean PI 404198A for resistance to race 2 (Table 1). QTLs were identified on G, B1 and E in PI 90763 and on LGs B1 and N in PI 404198A for resistance to race 5 (Table 1). Statistical significance for QTL on LG G for resistance to race 5 in PI 404198A (LOD = 2.5) did not reach but was close to the suggestive level.

Pooled analysis using LSCIM

In order to test if a QTL exists at a putative position, the F vs. R1 pooled analysis on data from Hamilton × PI 90763 and Magellan × PI 404198A populations was conducted using LSCIM. QTLs on LGs G, B1 and J were found for resistance to race 2 whereas QTLs on G, B1 and E were detected for resistance to race 5 (Table 2). Statistical significance for QTL on LG N for resistance to race 5 identified by the above single population analysis in PI 404198A did not reach the suggestive level in the pooled analysis, indicating that there was no strong evidence, as a whole, from two populations to support the existence of a QTL on LG N. Compared with single population analyses above, no additional QTL was detected by the pooled analysis. The 1-LOD confidence intervals of the QTLs on LGs G, B1 and J for resistance to race 2 identified by the pooled analysis overlapped substantially with the ones of the QTLs on LGs G, B1, and J for resistance to race 2 identified by single population analyses in one or both PI 90763 and PI 404198A (Table 1 vs. Table 2). The 1-LOD confidence intervals of the QTLs on LGs G, B1 and E for resistance to race 5 identified by the pooled analysis also overlapped substantially with the confidence intervals of the QTLs on LGs G, B1 and E for resistance to race 5 identified by single population analyses in one or both PI 90763 and PI 404198A (Table 1 vs. Table 2). Therefore, the F vs. R1 pooled analysis was consistent with the above single population analyses. Compared with single population analyses, however, the F vs. R1 pooled analysis showed significantly increased LOD values on the QTL candidate regions (the 1-LOD confidence interval regions in Table 2) of LGs G and B1 for resistance to races 2 and 5 where a QTL was detected by single population analyses in both PI 90763 and PI 404198A (Fig. 1). But, the F vs. R1 analysis showed slightly decreased LOD values on the QTL candidate regions of LG J for resistance to race 2 and of LG E for resistance to race 5 where a QTL was identified in only one of the two populations (Fig. 1).

If the QTL carried by PI 90763 were located away from the one by PI 404198A on a chromosome, bi-peaks should be expected on the plot of LOD against a linkage group map in the F vs. R1 analysis. No obvious bi-peaks appeared on LGs G and B1 for resistance to races 2 and 5 (Fig. 1), indicating that QTLs on LG G and B1 identified in PI 90763 may be located on the same locus or closely linked with those on LGs G and B1 in PI 404198A for resistance to races 2 and 5.

The 1-LOD confidence intervals of QTLs on LGs G and B1 for resistance to race 2 overlapped substantially with the ones on LGs G and B1 for resistance to race 5 in the pooled analysis (Table 2), respectively, indicating that QTLs on LGs G and B1 for resistance to race 2 may be the same or closely linked with the ones on LGs G and B1 for resistance to race 5, respectively.

The F vs. R1 pooled analysis did not provide information on existence of a QTL in specific populations. In order to know this information, the F vs. R2 and F vs. R3 pooled analysis or single population analyses on individual populations need to be conducted. In this study, phenotypes standardized in residual values were used for the pooled analysis. Nearly the same curves in the plot of LOD against a linkage group map were produced by the pooled analysis (F vs. R2, F vs. R3) and single population analyses on corresponding populations, as expected (data not shown). However, if residual variances were not equal among populations (the original data of this study gave significant differences for residual variance between the two populations), more QTLs were declared by the pooled analysis (F vs. R2, F vs. R3) in the population with the larger residual variance and fewer QTLs were declared in the population with the smaller residual variance (data not shown).

In order to determine if the QTLs on LGs G and B1 carried by PI 90763 are the same or different in gene effect from the ones on G and B1 in PI 404198A for resistance to races 2 and 5, the F vs. R4 pooled analysis was conducted on LGs G and B1 using LSCIM. It showed that QTLs on G and B1 carried by PI 90763 were not significantly different in gene effect from QTLs on LGs G and B1 in PI 404198A (LOD > 3.0 in F vs. R4 analysis), respectively, for resistance to races 2 and 5 (Table 3, Fig. 1).

The most significant distinction of LSCIM pooled analysis described by this study from the LSIM pooled analysis by Walling et al. (2000) is that cofactor markers

Table 2 F versus R1 pooled analysis on combined data from multiple populations using least square composite interval mapping

SCN races	Linkage groups	Position[a]	1 LOD CI[b]	LOD	R^2 (%)[c]
II	G	2.0	0–6.0	10.3[g]	10.8
	B1	118.5	108.5–122.5[d]	7.3[g]	7.8
	J	75.5	59.5–89.5[d]	4.3[g]	4.8
V	G	2.0	0–8	6.2[g]	6.9
	B1	112.5	104.5–118.5	10.8[g]	11.2
	E[e]	39.3	25.3–55.3	4.9[g]	5.6
	N	50.3	36.3–76.3	2.6	3.2

F versus R1 pooled analysis: F: the full model, R1: the reduced model (R1). This analysis was used to test if a QTL exists at a putative position

[a]Distance from the start of a linkage group on the soybean composite linkage map

[b]1-LOD confidence interval

[c]A proportion of the total variation explained by a QTL given the other QTLs (cofactor markers). It is defined as SSE(F)−SSE(R)/SSE (X), where SSE(F) is error sum of square associated with the full model (F) (QTL exist in both populations), SSE(R) error sum of square associated with reduced model (R1) (no QTL exist in any populations), and SSE(X) total sum of square excluding the sum of square due to differences among populations

[d]The end of the part of a linkage group that was searched for a QTL in this study

[e]Two peaks occurred on this chromosome but their 1-LOD confidence intervals overlapped substantially (Fig. 2). One QTL was declared for the peak with the highest LOD on this chromosome. The 1-LOD confidence interval covered the 1-LOD confidence intervals of both peaks. In order to exclude or confirm that closely linked QTLs may exist, fine mapping is needed

[f]Suggestive QTL (LOD ≥ 3.0, genome-wise type I error = 0.63)

[g]Significant QTL (LOD ≥ 4.0, genome-wide type I error = 0.05)

were used for reducing the interfering effects of QTLs located elsewhere on the genome. In order to examine the effect of cofactor markers on the pooled analysis, the F vs. R1 pooled analysis was also conducted using LSIM, which is different from LSCIM by the fact that no cofactor markers were included in the models. The same QTLs were declared by LSIM pooled analysis and LSCIM pooled analysis (Table 2 vs. Table 4). However, LSCIM pooled analysis showed significantly increased LOD values on the QTL candidate regions of LGs B1 and J for resistance to race 2 and of LGs B1 and E for resistance to race 5 (Fig. 2b is presented as an example). LSCIM and LSIM pooled analyses showed similar LOD values on LG G for resistance to races 2 and 5 (Fig. 2a is presented as an example). In conclusion, inclusion of cofactor markers in the pooled analysis may increase the LOD values and therefore it may be favorable to pooled analysis.

Discussion

Comparison of LSCIM pooled analysis with CIM single population analyses

Composite interval mapping (CIM) (Zeng 1994) is the most commonly used method for QTL analysis at this time. In our previous studies (Guo et al. 2005, 2006a), we analyzed the data from Hamilton × PI 90763 and Magellan × PI 404198A separately using CIM and LG maps constructed in the respective experiments. Results are summarized in Table 4, where declared QTLs were projected on the composite linkage map used by this study based on the relative positions of QTLs between their two flanking markers. LSCIM and CIM single population analyses gave similar results (Table 1 vs. Table 5). The same QTLs were detected by LSCIM and CIM. The differences of QTL positions between LSCIM and CIM were between 0 and 6.3 cM (three test positions only) (Table 1 vs. Table 5). In most cases, CIM showed larger LOD than LSCIM. These differences may be partly due to the use of different LG maps used by the LSCIM and the CIM.

Compared with CIM single population analyses, LSCIM F vs. R1 pooled analysis showed a large increase for the highest LOD value on LGs G and B1 for resistance to race 2 and on LG B1 for resistance to race 5 where a QTL was detected in both populations (Table 2 vs. Table 5). LSCIM F vs. R1 pooled analysis, however, showed a decrease for the highest LOD value on LG G for resistance to race 5. This may be due to a large difference for LOD value between the two populations in single population analyses. LSCIM F vs. R1 pooled analysis showed a decrease for the highest LOD value on

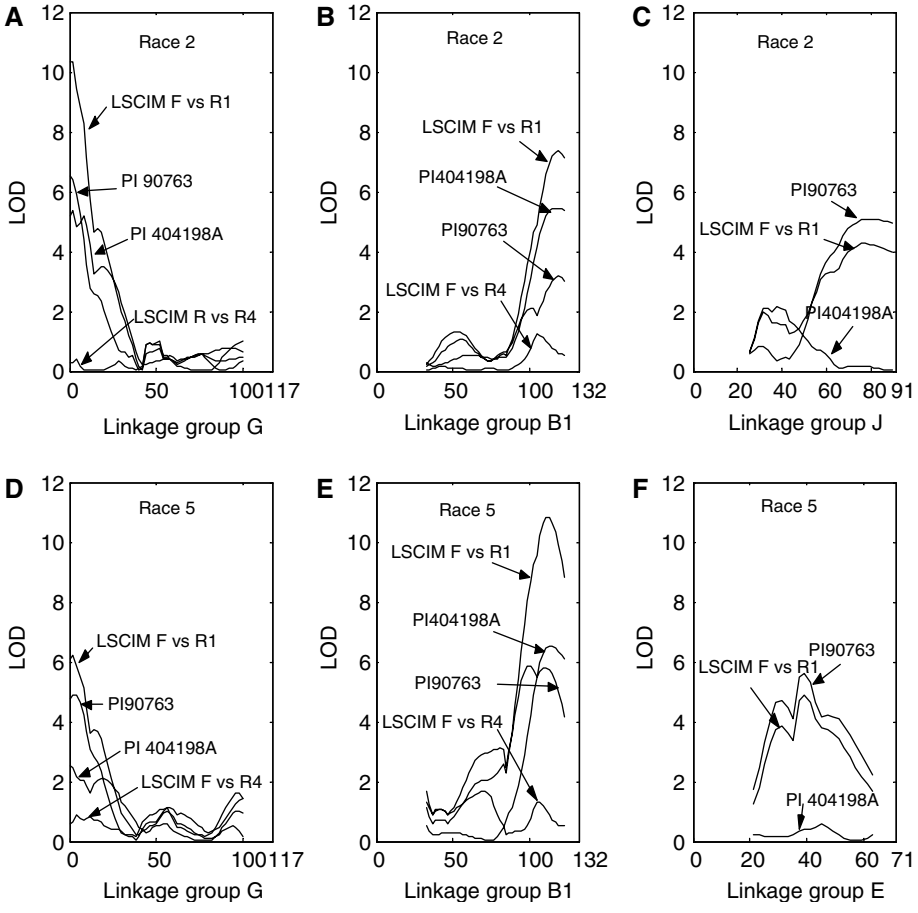

Fig. 1 Least square composite interval mapping (*LSCIM*) pooled analysis and single population analyses on soybean populations Hamilton × PI 90763 and Magellan × PI 404198A for resistance to soybean cyst nematode races 2 and 5. LSCIM F versus R1: pooled analysis for testing if a QTL exists at a putative position. LSCIM F versus R4: pooled analysis for testing if QTLs carried by different populations have a significant difference in gene effect. PI 90763: single population analysis on Hamilton × PI 90763 population using LSCIM. PI 404198A: single population analysis on Magellan × PI 404198A population using LSCIM

Table 3 F versus R4 pooled analysis using least square composite interval mapping

SCN races	Linkage groups	1-LOD region[a]	LOD[b]
II	G	0–6.0	0.4
	B1	108.5–122.5[c]	1.1
V	G	0–8	0.9
	B1	104.5–118.5	1.3

F versus R4 pooled analysis: F: the full model, R4: the reduced model (R4). This analysis was used to test if QTLs carried by different populations have a significant difference in gene effect
[a]1-LOD confidence interval in Table 2
[b]The largest LOD on the 1-LOD region. If LOD < 3.0, it was concluded QTLs carried by different populations may have no significant difference in gene effect
[c]The end of the part of a linkage group that was searched for a QTL in this study

LG J for resistance to race 2 and on LGs E and N for resistance to race 5 where a QTL was detected in only one of the two populations.

Uses of pooled analysis

No additional QTL was declared by the pooled analysis compared with single-population-based analysis. But this study and Li et al. (2005) demonstrates that where a QTL was shared among different populations, pooled analysis showed a significantly increased statistical evidence (LOD) for QTL over single population-based analyses. It is expected that a QTL will be detected by the pooled analysis but it may not be detected by single population analyses if a QTL with small gene effect is shared by different populations. Pooled analysis on data from genetically similar populations, therefore, may have higher power of QTL detection than single population-based analyses. In this study, pooled analysis did not show significantly increased precision of QTL detection (narrowing down the confidence interval of a QTL) (Table 1 vs. Table 2 for 1 LOD CI). But, Li et al. (2005) and Walling et al. (2000) showed that pooled analysis increased the precision of QTL detection. In order to get a comprehensive understanding of the power and precision of pooled analysis, it would be helpful to conduct an extensive simulation study.

A novel method has been developed for identifying the gene underlying a QTL which combines multiple cross mapping with molecular marker haplotype analysis (Hitzemann et al. 2003; Park et al. 2003; Wang et al. 2004). This method is based on the fact that the gene and its causative polymorphism(s) underlying a QTL should be the ones which are shared among inbred lines with the same alleles but differ among inbred lines with different alleles. Therefore, comparison in gene effect of a QTL among inbred lines is the key for this method. As stated above, pooled analysis can be used for examining differences of a QTL among different mapping populations (lines). Pooled analysis would make an important contribution in identifying the genes underlying the QTLs.

Table 4 F versus R1 pooled analysis on combined data from multiple populations using least square interval mapping (LSIM)

SCN races	Linkage groups	Position[a]	1 LOD CI[b]	LOD	R^2 (%)[c]
II	G	2.0	0.0–6.0	10.3[g]	11.9
	B1	118.5	106.5–122.5[d]	6.2[g]	7.7
	J	75.5	65.5–89.5[d]	3.2[f]	4.5
V	G	0.0	0.0–8.0	5.8[g]	7.5
	B1	112.5	104.5–122.5[d]	9.6[g]	11.2
	E[e]	39.3	23.3–61.3	3.8[f]	5.4
	N	58.3	36.3–58.3	2.7	4.0

F versus R1 pooled analysis: F: the full model, R1: the reduced model (R1). No cofactor markers were included in LSIM. This analysis was used to test if a QTL exists at a putative position
[a]Distance from the start of a linkage group on the soybean composite linkage map
[b]1-LOD confidence interval
[c]A proportion of total variation explained by a QTL. It is defined as SSE(F)−SSE(R)/SSE (X), where SSE(F) is error sum of square associated with full model (F) (QTL exist in both populations), SSE(R) error sum of squares associated with reduced model (R1) (no QTL exist in any populations), and SSE(X) total sum of squares excluding the sum of square due to differences among populations
[d]The end of the part of a linkage group that was searched for QTL in this study
[e]Two obvious peaks occurred on this chromosome but their 1-LOD confidence intervals overlapped substantially (Fig. 2). One QTL was declared for the peak with the highest LOD on this chromosome. The 1-LOD confidence interval covered the 1-LOD confidence intervals of both peaks. In order to exclude or confirm that closely linked QTLs may exist, fine mapping is needed
[f]Suggestive QTL (LOD ≥3.0, genome-wise type I error = 0.63)
[g]Significant QTL (LOD ≥4.0, genome-wide type I error = 0.05)

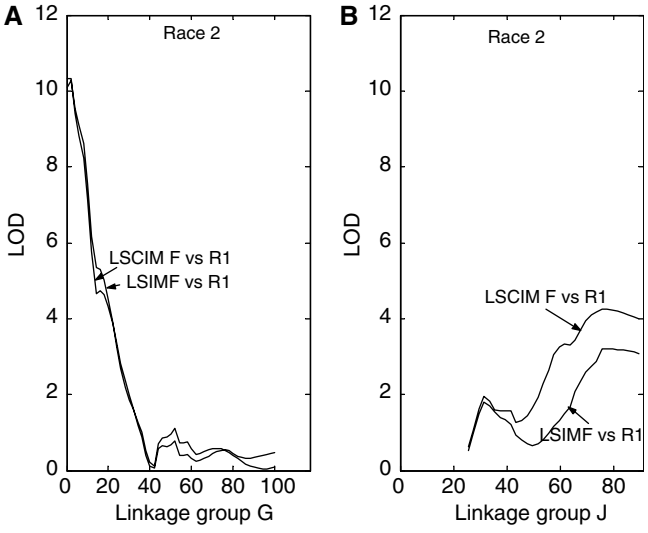

Fig. 2 Comparison of least square composite interval mapping (*LSCIM*) pooled analysis with least square interval mapping (*LSIM*) pooled analysis. LSCIM is different from LSIM by the fact that cofactor markers are included as covariate variables in the multiple linear models. LSCIM or LSIM F versus R1: analysis on pooled data from Hamilton × PI 90763 and Magellan × PI 404198A populations for testing if a QTL exists at a putative position

Table 5 Single population analyses using composite interval mapping (CIM)

SCN races	Population	Linkage groups	Linked markers	QTL positions[a]	LOD	R^2 (%)[b]
II	Hamilton × PI 90763	G	Satt163	1.4	7.9[e]	14.7
		B1	Satt453	116.8	3.0[d]	6.7
		J	Sat_224	72.1	4.6[e]	7.8
	Magellan × PI 404198A	G	Satt163	0.0	7.1[e]	12.5
		B1	Satt453	(124)[c]	5.5[e]	11.1
V	Hamilton × PI 90763	G	Satt163	2.7	7.1[e]	13.0
		B1	Satt453	116.8	6.0[e]	11.2
		E	Satt573	39.3	7.2[e]	12.5
	Magellan × PI 404198A	G	Satt163	0.0	3.3[d]	6.3
		B1	Satt453	(124)[c]	6.7[e]	13.0
		N	Sat_280	48	3.0[d]	9.5

[a]QTLs were detected in our previous studies using linkage group maps constructed in particular mapping populations and CIM. For comparison, the QTLs were projected on the soybean composite linkage map used by this study based on the relative position of a QTL between its flanking markers
[b]A proportion of the total variation explained by a QTL
[c]Position of linked molecular marker Satt453
[d]Suggestive QTL (LOD ≥3.0, genome-wide type I error = 0.63)
[e]Significant QTL (LOD ≥4.0, genome-wide type I error = 0.05)

Acknowledgements The authors thank Dr. Michael McMullen, USDA-ARS, University of Missouri-Columbia for helpful advice on this investigation.

References

Arelli PR, Wilcox JA, Myers J, Gibson PT (1997) Soybean germplasm resistant to races 1 and 2 of *Herodera glycines*. Crop Sci 37:1367–1369

Basten CJ, Weir BS, Zeng ZB (2002) QTL cartographer, Version 1.16 Department of Statistics. North Carolina State University, Raleigh, NC

Concibido VC, Diers BW, Arelli PR (2004) A decade of QTL mapping for cyst nematode resistance in soybean. Crop Sci 44:1121–1131

Guo B, Sleper DA, Arelli PR, Shannon, Nguyen HT (2005) Identification of QTLs associated with resistance to soybean cyst nematode races 2, 3 and 5 in soybean PI 90763. Theor Appl Genet 111:965–971

Guo B, Sleper DA, Nguyen HT, Arelli PR, Shannon JG (2006a) Quantitative trait loci underlying resistance to three soybean cyst nematode populations in soybean PI 404198A. Crop Sci. 46:224–233

Guo B, Sleper DA, Lu P, Shannon JG, Nguyen HT, Arelli PR (2006b) QTLs associated with resistance to soybean cyst nematode in soybean: meta-analysis of QTL locations. Crop Sci 46:595–602

Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324

Haley CS, Knott SA, Elsen JM (1994) Mapping quantitative trait loci in crosses between outbred lines using least squares. Genetics 136:1195–1207

Heo M, Leibel RL, Boyer BB, Chung WK, Koulu M, Karvonen MK, Pesonen U, Rissanen A, Laakso M, Uusitupa MIJ, Chagnon Y, Bouchard C, Donohoue PA, Burns TL, Shuldiner AR, Silver K, Andersen RE, Pedersen O, Echald S, Sorensen TIA, Behn P, Permutt MA, Jacobs KB, Elston RC, Hoffman DJ, Alllison DB (2001) Pooling analysis of genetic data: the association of leptin receptor (LEPR) polymorphisms with variables related to human adiposity. Genetics 159:1163–1178

Hitzemann R, Malmanger B, Reed C, Lawler M., Hitzemann B, Coulombe S, Buck K, Rademacher B, Walter N, Polyakov Y, Sikela J, Gensler B, Burgers S, Williams RW, Manly K, Flint J, Talbot C (2003) A strategy for the integration of QTL, gene expression, and sequence analyses. Mamm Genome 14:733–747

Jansen RC (1993) Interval mapping of multiple quantitative trait loci. Genetics 135:205–211

Lander ES, Bostein D (1989) Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps. Genetics 121:185–199

Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. Nat Genet 11:241–247

Li R, Lyons MA, Wittenburg H, Paigen B, Churchill GA (2005) Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics 169:1699–1709

Liu Y, Zeng ZB (2000) A general mixture model approach for mapping quantitative trait loci from diverse cross designs involving multiple inbred lines. Genet Res 75:345–355

Neter J, Kutner MH, Nachtsheim CJ, Wasserman W (1996) Applied linear statistical models. The McGraw-Hill Companies, Inc

Niblack TL, Arelli PR, Noel GR, Opperman CH, Orf J, Schmitt DP, Shannon JG, Tylka GL (2002) A revised classification scheme for genetically diverse populations of *Heterodera glycines*. J Nematol 34:279–288

Park YG, Clifford R, Buetow KH, Hunter KW (2003) Multiple cross and inbred strain haplotype mapping of complex-trait candidate genes. Genome Res 13:118–121

Rebai A, Goffinet B (1993) Power of tests for QTL detection using replicated progenies derived from a diallel cross. Theor Appl genet 86:1041–1022

Schmitt DP, Shannon JG (1992) Differentiating soybean responses to *Heterodera glycines* races. Crop Sci. 32:275–277

Sen S, Churchill GA (2001) A statistical framework for quantitative trait mapping. Genetics 159:371–387

Song QJ, Marek LF, Shoemaker RC, Lark KG, Concibido VC, Delannay X, Specht JE, Cregan PB (2004) A new integrated genetic linkage map of the soybean. Theor Appl Genet 109:122–128

Stam P (1993) Construction of integrated genetic linkage maps by means of a new computer package: JoinMap. Plant J 3:739–744

Walling GA, Visscher PM, Andersson L, Rothschild MF, Wang L, Moser G, Groenen AM, Bidanel JP, Cepica S, Archibald AL, Geldermann H, Koning DJ, Milan D, Haley CS (2000) Combined analysis of data from quantitative trait loci mapping studies: chromosome 4 effects on porcine growth and fatness. Genetics 155:1369–1378

Wang X, Korstanje R, Higgins D, Paigen B (2004) Haplotype analysis in multiple crosses to identify a QTL gene. Genome Res 14:1767–1772

Wrather JA, Chambers AY, Fox JA, Moore FW, Sciumbato GL (1995) Soybean disease loss estimates for the southern United States, 1974 to 1994. Plant Dis 79:1076–1079

Wrather JA, Anderson TR, Arsyad DM, Tan Y, Ploper LD, Porta-Puglia A, Ram HH, Yorinori JT (2001) Soybean disease loss estimates for the top ten soybean-producing countries in 1998. Can J Plant Pathol 23:115–121

Xu S (1998) Mapping quantitative trait loci using multiple families of line crosses. Genetics 148:517–524

Zeng ZB (1993) Theoretical basis of separation of multiple linked gene effects on mapping quantitative trait loci. Proc Natl Acad Sci USA 90:10972–10976

Zeng ZB (1994) Precision mapping of quantitative trait loci. Genetics 136:1457–1468