OPEN ACCESS Freely available online

PLOS ONE

An Efficient Hierarchical Generalized Linear Mixed Model for Mapping QTL of Ordinal Traits in Crop Cultivars

Jian-Ying Feng, Jin Zhang, Wen-Jie Zhang, Shi-Bo Wang, Shi-Feng Han, Yuan-Ming Zhang*

Section on Statistical Genomics, State Key Laboratory of Crop Genetics and Germplasm Enhancement, Department of Crop Genetics and Breeding, Nanjing Agricultural University, Nanjing, Jiangsu, China

Abstract

Many important phenotypic traits in plants are ordinal. However, relatively little is known about the methodologies for ordinal trait association studies. In this study, we proposed a hierarchical generalized linear mixed model for mapping quantitative trait locus (QTL) of ordinal traits in crop cultivars. In this model, all the main-effect QTL and QTL-by-environment interaction were treated as random, while population mean, environmental effect and population structure were fixed. In the estimation of parameters, the pseudo data normal approximation of likelihood function and empirical Bayes approach were adopted. A series of Monte Carlo simulation experiments were performed to confirm the reliability of new method. The result showed that new method works well with satisfactory statistical power and precision. The new method was also adopted to dissect the genetic basis of soybean alkaline-salt tolerance in 257 soybean cultivars obtained, by stratified random sampling, from 6 geographic ecotypes in China. As a result, 6 main-effect QTL and 3 QTL-by-environment interactions were identified.

Citation: Feng J-Y, Zhang J, Zhang W-J, Wang S-B, Han S-F, et al. (2013) An Efficient Hierarchical Generalized Linear Mixed Model for Mapping QTL of Ordinal Traits in Crop Cultivars. PLoS ONE 8(4): e59541. doi:10.1371/journal.pone.0059541

Editor: Rongling Wu, Pennsylvania State University, United States of America

Received December 19, 2012; **Accepted** February 15, 2013; **Published** April 2, 2013

Funding: This work was supported by the National Basic Research Program of China grant 2011CB109306; the National Natural Science Foundation of China grant 30971848; the Fundamental Research Funds for the Central Universities grants KYT201002 and KJ2011003; a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions and the Specialized Research Fund for the Doctoral Program of Higher Education grants 20100097110035 and 20120097110023. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: soyzhang@njau.edu.cn

Introduction

Many characters of biological interest and economic importance vary in an ordinal form, i.e. disease and tolerance, but are not inherited in a simple Mendelian fashion. More importantly, they cause substantial yield loss. To decrease the loss, developing resistance cultivar is the most economic and effective way. Therefore, there is a critical need for in-depth study of methodology for mining elite alleles for ordinal traits.

During the past several decades, many attempts have been made to mine elite alleles for binary and ordinal traits. The methodologies of mapping quantitative trait loci (QTL) for discrete traits have been well established within the framework of threshold model. On the early stage, almost all the approaches are based on single QTL genetic model [1–7]. Later on, several methods have been proposed to simultaneously identify multiple QTL for ordinal traits [8,9]. Recently, Bayesian methodology has been used to map multi-QTL and epistatic QTL for binary and ordinal traits [10–14]. However, all the above approaches are based on bi-parental segregating populations.

Many commercial inbred lines are available in crops. A large amount of elite alleles have preserved among these lines. Mining these elite alleles is the prerequisite in the integration of genetic analysis with crop breeding. Up to now, some approaches for mining elite alleles in crop cultivars have been developed [15–20]. All kinds of QTL can be effectively identified, elite alleles can be easily mined and novel parental combination can be effectively predicted [18]. However, these approaches in crop cultivars are for quantitative traits but not for discrete traits. As for discrete traits, too much complication comes from seemingly simple descriptions and unknown population structure meanwhile in fact the underlying biological model may be complicated. Accordingly, genetic analyses may be more challenging for discrete traits than for continuous traits. If pedigree information among these lines is known, Bayesian linkage analysis [21] and variance-components approach [22] have been presented. If the pedigree information is not known, relatively little has been known, except for Iwata et al. [23] and Hoggart et al. [24]. Although Iwata et al. [23] have developed Bayesian multilocus association analysis, the method is implemented via Markov chain Monte Carlo, and computing time becomes a major concern. Although Hoggart et al. [24] proposed simultaneous analysis of all SNPs in genome-wide association study, the method is for case-control dataset.

Multi-QTL mapping for discrete and quantitative traits is now the state-of-the-art method [18–20,24,25]. However, it is difficult to implement under the maximum-likelihood framework. At present the Bayesian method implemented via expectation-maximization algorithm [26] is specialized to handle complicated models and thus it is the ideal tool for mapping multiple QTL for ordinal trait in crop cultivars. Accordingly, in this study empirical Bayes approach of Xu [26] and the computational algorithm of Yi et al [27] were incorporated into the hierarchical generalized linear model of Yi et al [12] to map main-effect QTL (M-QTL) and QTL-by-environment (QE) interaction for ordinal traits in

crop cultivars. The new method was validated by a series of Monte Carlo simulation experiments and real data analysis in soybean.

Results

Phenotypic variation for soybean alkaline-salt tolerance

We measured lengths of main root (LR) of 257 soybean cultivars under the cases of control (CK), 100 mM NaCl and 10 mM Na2CO3. These original trait observations might be transferred into alkaline tolerance index (ATI) and salt tolerance index (STI). To measure the degree of salt-alkaline tolerance, these indexes were partitioned into five grades: high tolerance, tolerance, middle tolerance, sensitivity, and high sensitivity. In other words, this data is ordinal. The phenotypic distributions were shown in **Fig. 1** and **Table S1**. All the two discrete indexes almost exhibited skewed distribution, indicating the existence of genetic variation. Results from χ^2 test showed that there is significant relationship between the tolerance and environment ($\chi^2 = 44.83$ and $P<1e\text{-}4$ for ATI, and $\chi^2 = 13.29$ and $P = 0.004$ for STI), indicating the existence of environmental interaction.

Mapping M-QTL and QE interaction for ATI and STI

A total of 6 M-QTL (3 for ATI, and 3 for STI) and 3 QE interactions (one for ATI, and 2 for STI) for soybean alkaline-salt tolerance are detected by new method, and mapped to chromosomes A1, B2, I, L, N and O. Among them, one QTL, associated with marker sat_274, is responsible simultaneously for the above two traits; seven QTL are consistent with those of continuous ATI and STI using enriched compression mixed linear model (ECMLM) [28] and epistatic association mapping (EAM) [18] methods, and the other two were also confirmed by test of independence (χ^2 test); and one M-QTL and one QE interaction are associated simultaneously with marker satt270. A summary of all detected QTL is shown in **Table 1**.

4 ATI QTL, with proportion of phenotypic variance explained by single QTL (PVE) of 3.29–11.04%, are detected and mapped to chromosomes A1, B2 and O. Of these QTL, there are three M-QTL (18.96%) and one QE interaction (11.04%); and three QTL are further identified by ECMLM (or EAM) and χ^2 test. It should be noted that the PVE by *qATI0-2* and *qATI5e*, associated respectively with sat_274 and sat_344, are greater than 10%.

5 STI QTL, with PVE of 4.21–9.17%, are detected and mapped to chromosomes I, L, N and O. Of these QTL, there are three M-QTL (21.06%) and two QE interactions (13.48%); and all the QTL, except for *qSTI10*, are further identified by ECMLM (or EAM) and χ^2 test. It should be noted that the PVE of all the QTL are less than 10%.

Mining elite alleles

The summaries of elite allele and its representative carrier are shown in **Table 1**. As for the *qATI14* associated with Sat_342, there are 12 alleles and one unknown allele. The effects for all these alleles can be estimated by maximum likelihood method. Of these effects, the 260 bp allele has the smallest effect −0.73, being an elite allele, which can be found in soybean cultivar Zunyizongzidou. Similarly, as for the *qSTI3e* associated with satt270, the 223 bp allele shows the smallest effect in 2010, elite allele combination is the 223 bp allele×2010 with an effect of −0.90.

Predicting novel parental combination

In a hypothetical cross of two cultivars, all the recombinant inbred lines (RILs) from the cross may be produced. In these RILs, the trait values can be predicted by the effects of all the detected loci. The best RIL with minimum value would represent the cross. Therefore, the best cross could be selected from all the crosses. It was found that any cultivar-pair does not pyramid all the elite alleles of the detected QTL. However, some four-cultivar combinations might pyramid all the elite alleles of salt-alkaline tolerances in this study, for example, the best two combinations were Zunyizongzidou × Hunanqiudou 1×Ludou 1×Qi 588-8, and Zunyizongzidou×Hunanqiudou 1×Ludou 2×Qi 588-8, which are used to simultaneously improve the two traits.

Prediction for potential candidate genes

The summary of potential candidate genes for alkaline-salt tolerance in soybean is shown in **Table 2**. A total of 7 soybean genes homologous to *Arabidopsis* are linked to 7 markers detected in this study, with physical distances of 206.21–129132.42 kb; and one gene (Glyma03g38040) is closely linked to the associated markers (satt022) in this study, within 210 kb in physical distance.

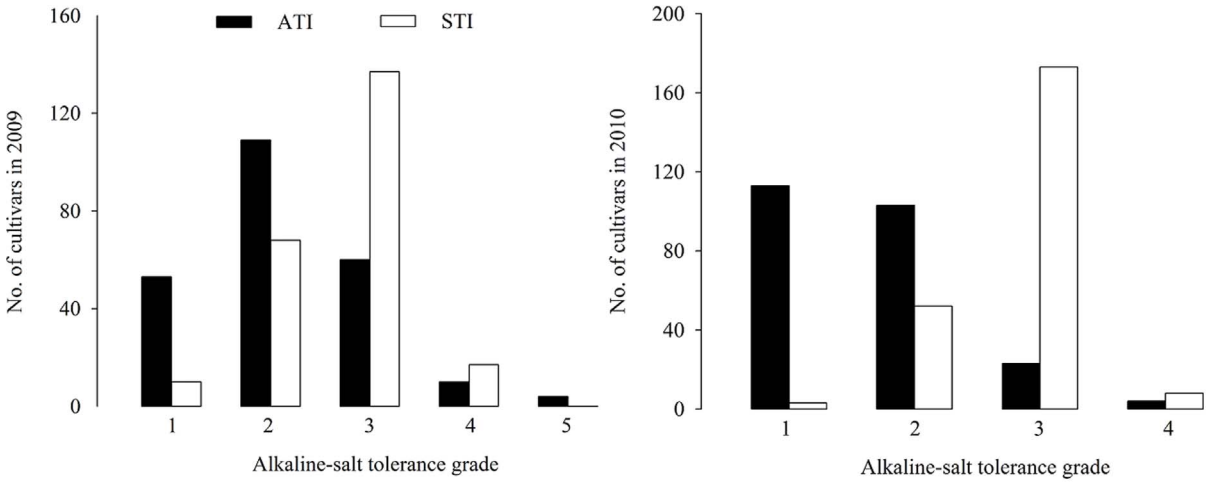

Figure 1. Frequency distribution for soybean alkaline-salt tolerance grade in 2009 (left) and 2010 (right).
doi:10.1371/journal.pone.0059541.g001

Table 1. Association mapping for ordinal alkaline-salt tolerance in 257 soybean cultivars.

Trait	New method								Elite allele of detected QTL			Similar result*		
	QTL	Type	Chr.	Marker	Position(cM)	Variance	LOD	r^2(%)	bp	Effect	Carrier	ECMLM	EAM	$P(H_0)^{\$}$
ATI	*qati14*	MQ	14(B2)	sat_342	20.30	0.0798	6.96	5.58	260	−0.73	Zunyizongzidou	√ (MQ)		1e-4
	qati10-1	MQ	10(O)	satt348	15.29	0.0469	3.12	3.29	232	−0.36	Jindou 3	√ (MQ)		0.0044
	qati10-2	MQ	10(O)	sat_274	107.58	0.1441	3.14	10.09	412	−3.39	Ludou 1			0.0027
	qati5e	QE	5(A1)	sat_344	19.37	0.1578	5.07	11.04	433×2010†	−5.24	Baiqiu 1	√ (MQ)	√ (QE)	<1e-4
STI	*qsti20*	MQ	20(I)	satt270	50.11	0.1173	3.26	7.68	243	−1.06	Jiangechengguanbayuehuang	√ (MQ)		0.5842
	qsti19	MQ	19(L)	satt652	30.87	0.1401	3.23	9.17	241	−0.81	Hunanqiudou 1	√ (MQ)		0.0520
	qsti10	MQ	10(O)	sat_274	107.58	0.0643	4.01	4.21	394	−0.65	Baiqiu 1			0.0027
	qsti20e	QE	20(I)	satt270	50.11	0.1311	3.80	8.58	252×2009	−0.9771	Shuichengzongzidou		√ (QE)	0.5842
	qsti3e	QE	3(N)	satt022	102.05	0.0748	5.04	4.90	223×2010	−0.90	Daheiqi		√ (QE)	0.0216

MQ: main-effect QTL; QE: QTL-by-environment interaction. *similar results for continuous ATI and STI were derived from **Zhang [33]** by enriched compression mixed linear model (ECMLM). and epistatic association mapping (EAM) approaches. †Year, i.e., 2009 and 2010. §Probability of null hypothesis in the test of independence between the tolerance and marker.

doi:10.1371/journal.pone.0059541.t001

Monte Carlo simulation studies

Comparison of new method with both single-QTL method and test of independence. In the first simulation experiment, each simulated sample was analyzed by three methods. One is multi-QTL-based method in this study (new method), one is to use the new method under the condition of single-QTL model and one is test of independence. All the results are shown in **Fig. 2**. Among the three methods, the statistical power of the new method is the maximum, and the false positive rate (FPR) of the new method is the minimum. The estimates of QTL effects and threshold values from the new method are closer to the corresponding true values than those from single-QTL method, although all the estimates were slightly biased. Relatively small variations were observed in the new method for the estimates of position and effects of QTL as well as the threshold values. Therefore, the new method works relatively well.

Effect of phenotypic distribution on QTL mapping. In the second simulation experiment, the effect of the shape of phenotypic distribution on the new method was assessed by letting the phenotypic distribution of five ordinal categories be set as 1:1:1:1:1 (uniform distribution), 1:2:4:2:1 (symmetrical distribution) and 8:5:3:1:1 (skewed distribution). Other parameters were the same as those in the first simulation experiment. The results are given in **Fig. 3**. We found that skewed distribution has decreased the statistical power. The optimal power occurred in the situation where the phenotypic distribution is bell-shaped. Relatively small variations were also observed in the three situations for the estimates of position and effects of QTL as well as the threshold values.

Effect of the number of categories on QTL mapping. In the third simulation experiment, we evaluated the effect of the number of categories on the new method. The design of the simulation was similar to that described in the first simulation experiment, except for the number of phenotypic categories. We simulated three levels for the number of categories: 2, 6 and 9. The corresponding phenotypic distributions were 1:1, 1:3:6:6:3:1 and 1:2:4:6:9:6:4:2:1, respectively. The results are given in **Fig. 4**, which shows that the estimate of QTL position is very close to its true value in the three cases, and the power for QTL detection increases as the number of categories increases. The reason is that increasing the number of categories has increased the information of predicting the liability from the observed categorical phenotype. In addition, relatively small variations were also observed in the three situations for the estimates of QTL effects and the threshold values.

Effect of sample size on QTL mapping. In the fourth simulation experiment, we assumed the pedigree to have the numbers of non-founders of 100, 200, 300 and 500, and the number of founders of 50. One hundred and one equally spaced markers, each with three alleles, were placed on each of three 1000 cM chromosome segments; and eighteen QTL, each with three alleles, were simulated with heritabilities of 0.01–0.15. Other parameters were given in **Table 3**. The results of five QTL are shown in **Fig. 5**. As expected, the QTL detection power increases and the variations for the estimates of QTL parameters and the threshold values decreases as sample size or QTL heritability increases.

Effect of the number of founders on QTL mapping. In the last simulation experiment, we assumed the pedigree to have the number of non-founders of 200, and the numbers of founders of 25, 50 and 75. Other parameters were the same as those in the fourth simulation experiment. The results of five QTL are shown in **Fig. 6**. As expected, the QTL detection power increases as the founder number increases and relatively small variations and

Table 2. Prediction for potential candidate genes that are homologous to alkaline-salt tolerance genes in *Arabidopsis thaliana*.

Genes in Arabidopsis thaliana	Homologous genes in soybean			Associated marker in this study is around soybean homologous gene (SHG)				Marker closely linked to SHG			
	Gene	Chr.	Position (bp)	Marker 1	Position (bp)	Distance to SHG (kb)	Associated QTL	Marker 2	Distance to SHG (kb)	Distance between markers 1 & 2 (cM)
AT2G47190 (MYB2)	Glyma03g38040	3	44477360–44476292	satt022	44682505–44682712	206.21	qSTI3e	satt022	206.21	0.0
AT5G63310 (NDPK2)	Glyma05g03010	5	2300110–2296337	sat_344	3691696–3691992	1,395.36	qATI5e	Sat_368	587.14	5.01
AT5G27150 (NHX1)	Glyma10g30020	10	38712235–38706503	sat_274	43209577–43209848	4,503.07	qATI10-2, qSTI10	Sat_242	132.468	33.53
AT1G69270 (RPK1)	Glyma13g06210	10	6487011–6490433	satt348	5491146–5491362	995.87	qATI10-1	Satt269	93.41	3.92
AT3G55990 (ESK1)	Glyma14g06370	14	4626260–4623046	sat_342	2954747–2954981	1,668.30	qATI14	Satt126	309.53	7.32
AT2G40950 (BZIP17)	Glyma19g30680	19	38336451–38334669	satt652	9202248–9202462	29,132.42	qSTI19	AW508247	834.98	7.95
AT3G05880 (RCI2A)	Glyma20g22290	20	32331094–32331431	satt270	34223110–34223331	1,892.02	qSTI20, qSTI20e	Satt354	1098.67	3.89

doi:10.1371/journal.pone.0059541.t002

biasedness for the estimates of QTL parameters and the threshold values were observed.

Discussion

In this study the probability of y_j, $\Pr(y_j|\bullet)$, is viewed as an approximate normal distribution so that empirical Bayes approach could be adopted to estimate genetic effects in the hierarchical generalized linear model for ordinal trait association studies. As a result, M-QTL and QE interaction for ordinal traits in crop cultivars can be identified, elite alleles can be mined and novel parental combinations can be predicted. Clearly, it integrates genetic analyses with crop breeding design. More importantly, the mapping results in this study are reliable because they have been validated in four aspects. First, seven QTL detected by new method are consistent with those by at least one of three approaches: ECMLM, EAM and single marker analysis (**Table 1**). Second, a total of 7 potential candidate genes homologous to *Arabidopsis* are found to be around 7 associated markers (**Table 2**). Third, some QTL were simultaneously identified among alkaline-salt tolerance index, original and ordinal traits, for example, Sat_342 and Satt348 were associated with alkaline tolerance, and Satt270 was associated with salt tolerance. Finally, the results from Monte Carlo simulation studies show that new method improves statistical power and precision, and reduces FPR.

The major contribution of this study is the pseudo data normal approximation of the likelihood function for ordinal trait association studies. The normal likelihood approximation was first developed by Wolfinger and O'Connell [29] and continued by Gelman et al [30]. McGilchrist [31] used a different approach for the same problem, but much easier to understand. Although the method has been explored for binary and binomial traits in linkage studies [32], this study is the first report of the pseudo data approximation for ordinal trait association studies.

We compared the new method with that of Lü et al. [18]. There are some commons between the two approaches. For example, the similar effects of phenotypic distribution (the number of categories, sample size and heritability) on QTL mapping in homozygous cultivars are observed. However, the differences exist as well. For example, the trait is quantitative in Lü et al. [18] and ordinal in this study; and the power for the detection of QTL is lower for this study than for Lü et al. [18], because limited information is observed for ordinal traits. As the number of categories increases, it is better to use the normal trait hierarchical linear mixed model. Note that the main benefit of this study comes from small number of categories. Although Iwata et al. [23] and Hoggart et al. [24] are for ordinal traits, in this study main-QTL, environmental effect and QTL-by-environment interactions were simultaneously considered in our full genetic model, improving the statistical power and estimation precision.

As compared with genome-wide association studies in Yu et al. [16] and Zhang et al. [17], kinship matrix was not considered in this study. In fact, this term is related to background control, which is similar to co-variable markers in composite interval mapping. Note that all the main-effect QTL and QTL-by-environment interactions are included in the full genetic model of this study. Thus, it is unnecessary to consider this term in the current study. In addition, in real data analysis we also consider the effect of population structure on association studies. As a result, a slightly different result is observed while Q matrix is deleted from the above full model.

Epistasis, the interaction between QTL, plays an important role in the dissection of genetic architecture for complex traits. To date,

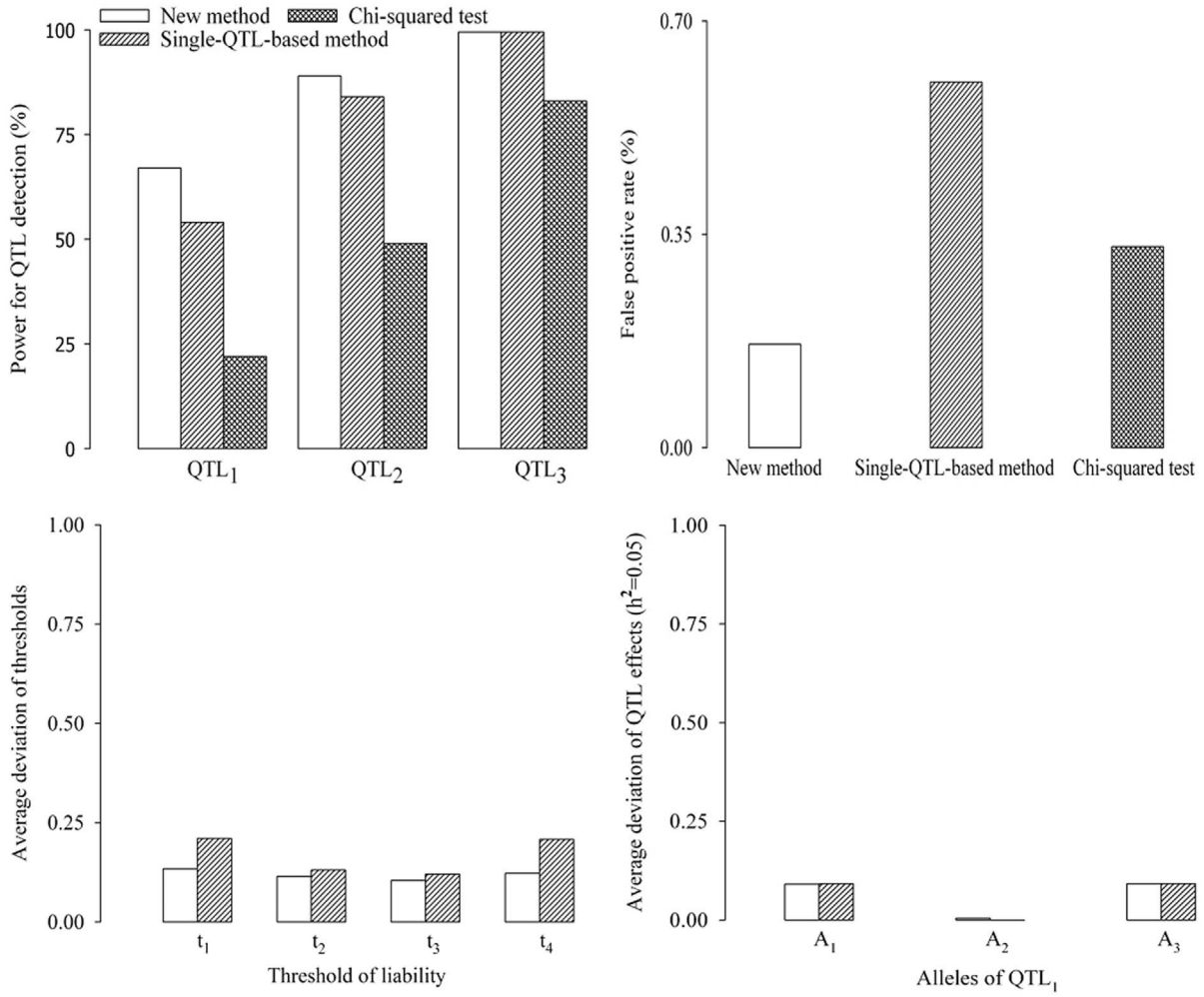

Figure 2. Comparison of new method with single-QTL-based method and Chi-squared test.
doi:10.1371/journal.pone.0059541.g002

several approaches have been developed, including multiple interval mapping, Bayesian approach, and penalized maximum likelihood method. Most of these methods are for quantitative traits in bi-parental segregating populations. In homozygous cultivars, it is relatively difficult. Because of its complexity, it will be investigated separately in a future project.

Materials and Methods

Soybean samples

257 soybean cultivars used in this study were mainly provided by the National Center for Soybean Improvement, China. All the cultivars were obtained by stratified random sampling from six geographic ecotypes in China, planted in three-row plots in a completely randomized design and evaluated at the Jiangpu experimental station at Nanjing Agricultural University in 2009 and 2010. The plots were 1.5 m wide and 2 m long. Twelve seeds for each cultivar were sown in a 30×20×15 cm plastic container with the 3.5 cm height sand and then treated with control (CK), 100 mM NaCl and 10 mM Na_2CO_3, and each with two replications. They were grown in a growth chamber under white fluorescent light (600 μmol m^{-2} s^{-1}; 14 h light/10 h dark) at 25±1°C. Length of main root (LR, centimetre) for healthy seedlings were measured from 5 plants 7 days after sowing. To measure the degree of salt-alkaline tolerance, original trait observations might be transferred into salt-alkaline tolerance index for each trait using the below equations

$$\text{salt tolerance index (STI)} = (x_{CK} - x_{NaCl})/x_{CK} \times 100\%$$

$$\text{alkaline tolerance index (ATI)} = \left(x_{CK} - x_{Na_2CO_3}\right) \Big/ x_{CK} \times 100\%$$

where x_{CK},x_{NaCl} and $x_{Na_2CO_3}$ stand for phenotypic values in control, saline and alkaline treatments, respectively [33]. The tolerance grades 1 to 5, used in this study, were indicated by 0–20%, 20–40%, 40–60%, 60–80% and 80–100%, respectively.

Approximately 0.3 g of fresh leaves obtained in 2008 from each cultivar was used to extract genomic DNA using the cetyltrimethylammonium bromide method as described by Lipp et al. [34]. To screen for polymorphisms among all the cultivars, polymerase chain reaction (PCR) was performed with 135 simple sequence repeat (SSR) primer pairs. The primer sequences were obtained from the soybean database Soybase (http://www.ncbi.nlm.nih.gov). PCR was performed as described by Xu et al. [35] and Wei et al. [36].

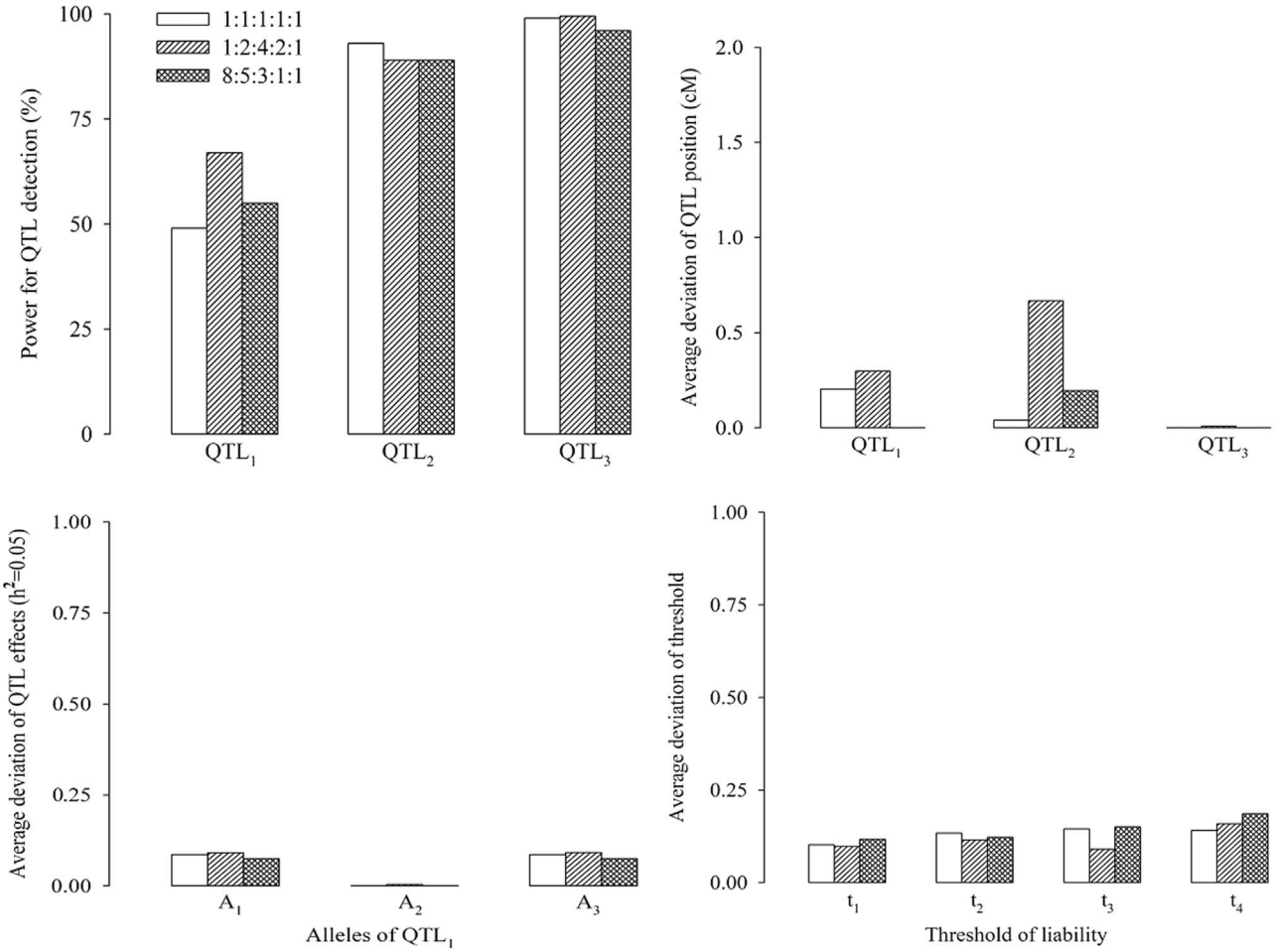

Figure 3. Effect of phenotypic distribution on association mapping for ordinal traits.
doi:10.1371/journal.pone.0059541.g003

Population structure

For the soybean sample, the STRUCTURE software [37] was used to investigate the population structures of all selected cultivars. The number of subpopulations (K) was set from 2 to 10. In the Markov chain Monte Carlo (MCMC) Bayesian analysis for each K, the length of a Markov chain consisted of 110,000 sweeps. The first 10,000 sweeps (the burn-in period) were deleted, and thereafter, the chain was used to calculate the mean of log-likelihood. This process was repeated 20 times, and the total average for mean log-likelihood at fixed K was used. STRUCTURE analysis with 135 SSR molecular markers showed that the log-likelihood increased with the increase of the model parameter K, so a suitable number of K could not be determined. In this situation, using the ad hoc statistic ΔK, based on the rate of change in the log-probability of data between successive K values [38], STRUCTURE accurately detected the uppermost hierarchical level of structure. Here, the ΔK value was much higher for the model parameter $K=4$ than for other values of K [33]. By combining this high ΔK value with knowledge of the breeding history of these cultivars, we chose a value of 4 for K. The Q matrix was calculated based on SSR markers and incorporated into the hierarchical generalized linear mixed model in this study.

Generalized linear mixed Model

Let l_j $(j=1,2,\cdots,n)$ be the vector of underlying latent variable or liability of cultivar j. For the jth cultivar, it is postulated that

$$l_j=\sum_{m=1}^{q}x_{jm}\beta_m+\sum_{k=1}^{v}u_{jk}\gamma_k+\sum_{m=2}^{q}\sum_{k=1}^{v}x_{jm}u_{jk}\gamma_{(m-1)v+k}+\varepsilon_j \quad (1)$$

where $\beta=(\beta_1,\cdots,\beta_q)'$ is non-genetic effects, i.e., population mean (β_1) and environmental effect $(\beta_2,\cdots,\beta_q)'$; γ_k is allelic effect for $k=1,\cdots,v$ and allele-by-environment interaction effect for $k=v+1,\cdots,vq$, $v=\sum_{i=1}^{p}(g_i-1)$, and g_i is the number of alleles for locus i $(i=1,\cdots,p)$; x_{jm} and u_{jk} are dummy variables of β_m and γ_k for cultivar j, respectively; and ε_j is the random residual error with an$N(0,\sigma^2)$ distribution. $\sigma^2=1$ will be adopted here because the liabilities are unobservable.

Methods of estimating allelic effects and allele-by-environment interaction effects are the same. For the sake of clarity of notation, we redefine the design matrix and the regression coefficients as follows. Let $z_{jk}=u_{jk}(k=1,\cdots,v)$ and $z_{jk}=x_{jm}u_{jr}(k=v+1,\cdots,vq;m=2,\cdots,q;r=1,\cdots,v)$. The above model is now rewritten as

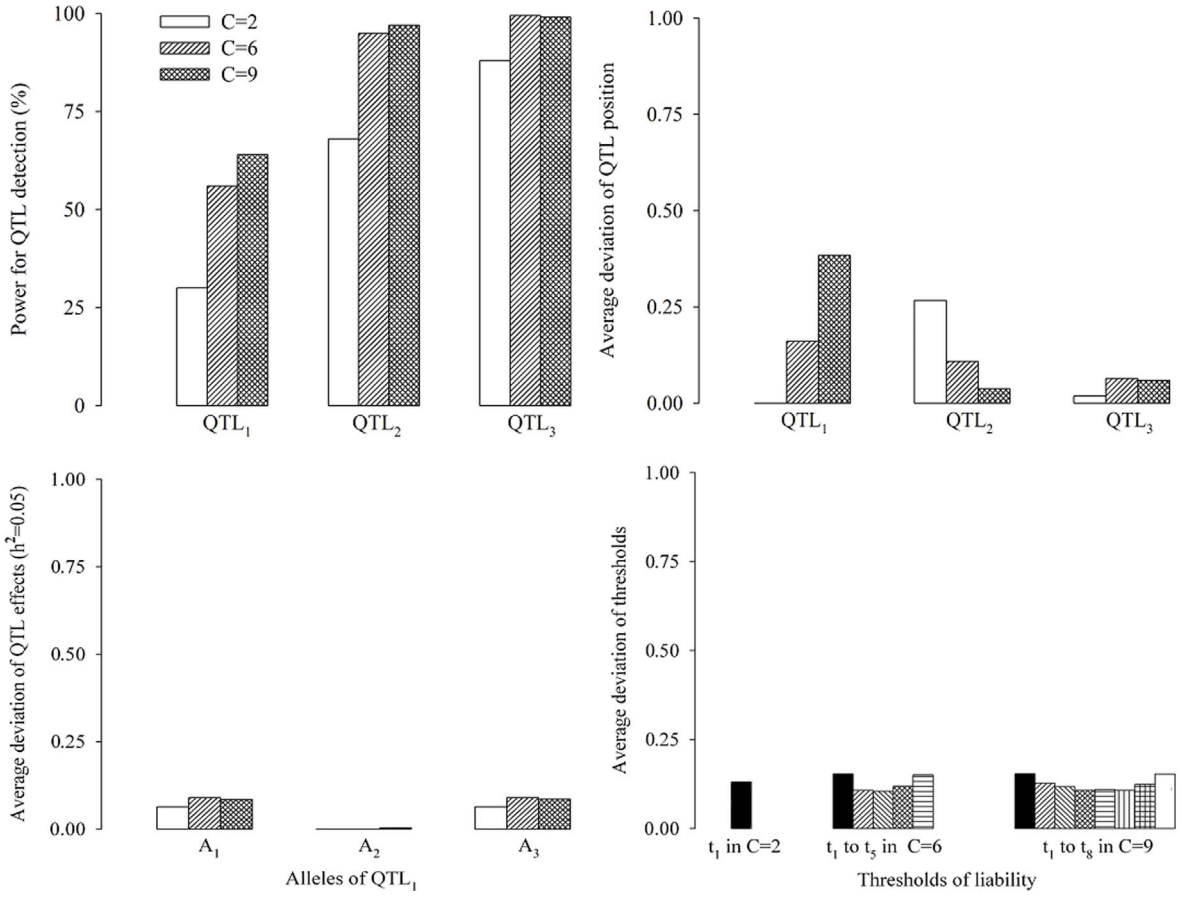

Figure 4. Effect of the number of categories on association mapping for ordinal traits.
doi:10.1371/journal.pone.0059541.g004

$$l_j = X_j\beta + Z_j\gamma + \varepsilon_j \quad (j = 1, \cdots, n) \tag{2}$$

where $\gamma = (\gamma_1, \cdots, \gamma_{vq})'$.

Let $\mathbf{Y} = \{y_j\}_{j=1}^{n}$ denote the vector of observed ordinal data. Here each y_j represents an assignment into C ordinal categories. These classes result from the hypothetical existence of $C+1$ thresholds $(t_0 = -\infty < t_1 < \cdots < t_C = +\infty)$ in the latent scale. The relationship between y_j and l_j is indicated by

$$t_{c-1} \le l_j < t_c \Leftrightarrow y_j = c \quad (c = 1, \cdots, C) \tag{3}$$

The conditional probability that y_j falls in category c, given β, γ and $t = (t_0, t_1, \cdots, t_C)$, is given by

$$\begin{aligned} \Pr(y_j = c|X_j, Z_j, t, \beta, \gamma) &= \Pr(t_{c-1} < l_j < t_c|X_j, Z_j, t, \beta, \gamma) \\ &= \Phi(t_c - X_j\beta - Z_j\gamma) - \Phi(t_{c-1} - X_j\beta - Z_j\gamma) \end{aligned} \tag{4}$$

where $\Phi(\bullet)$ is the cumulative distribution function of standard normal distribution. The data are conditionally independent, given β, γ and t. Therefore, log-likelihood function can be written as

$$\begin{aligned} &L(\theta|\mathbf{Y}) \\ &= \sum_{j=1}^{n}\sum_{c=1}^{C} \mathrm{I}_{y_j=c} \log \Pr(y_j = c|X_j, Z_j, t, \beta, \gamma) \\ &= \sum_{j=1}^{n}\sum_{c=1}^{C} \mathrm{I}_{y_j=c} \log\{\Phi(t_c - X_j\beta - Z_j\gamma) - \Phi(t_{c-1} - X_j\beta - Z_j\gamma)\} \end{aligned} \tag{5}$$

where $\theta = \{t_1, \cdots, t_C, \beta_1, \cdots, \beta_q, \gamma_1, \cdots, \gamma_{vq}\}$; and $\mathrm{I}_{y_j=c}$ is an indicator function taking value of 1 if $y_j = c$ and 0 otherwise.

Prior distribution and joint posterior density

The parameters β and γ are treated as fixed and random effects, respectively. The number of random effects in the above genetic model is very large so that the model is oversaturated. Therefore, the hierarchical generalized linear mixed model is adopted in this study. It is assumed that each genetic effect γ_k has a different variance σ_k^2. The following prior distributions are chosen for building the hierarchical model

$$\begin{aligned} &\gamma_k|\sigma_k^2 \sim N(0, \sigma_k^2), \sigma_k^2 \sim \chi^{-2}(\tau, \omega_k), \omega_k \sim \\ &Gamma(a, b_k), P(b_k) \propto 1 \quad (k = 1, \cdots, vq) \end{aligned}$$

where τ and a are the constants given in advance. When

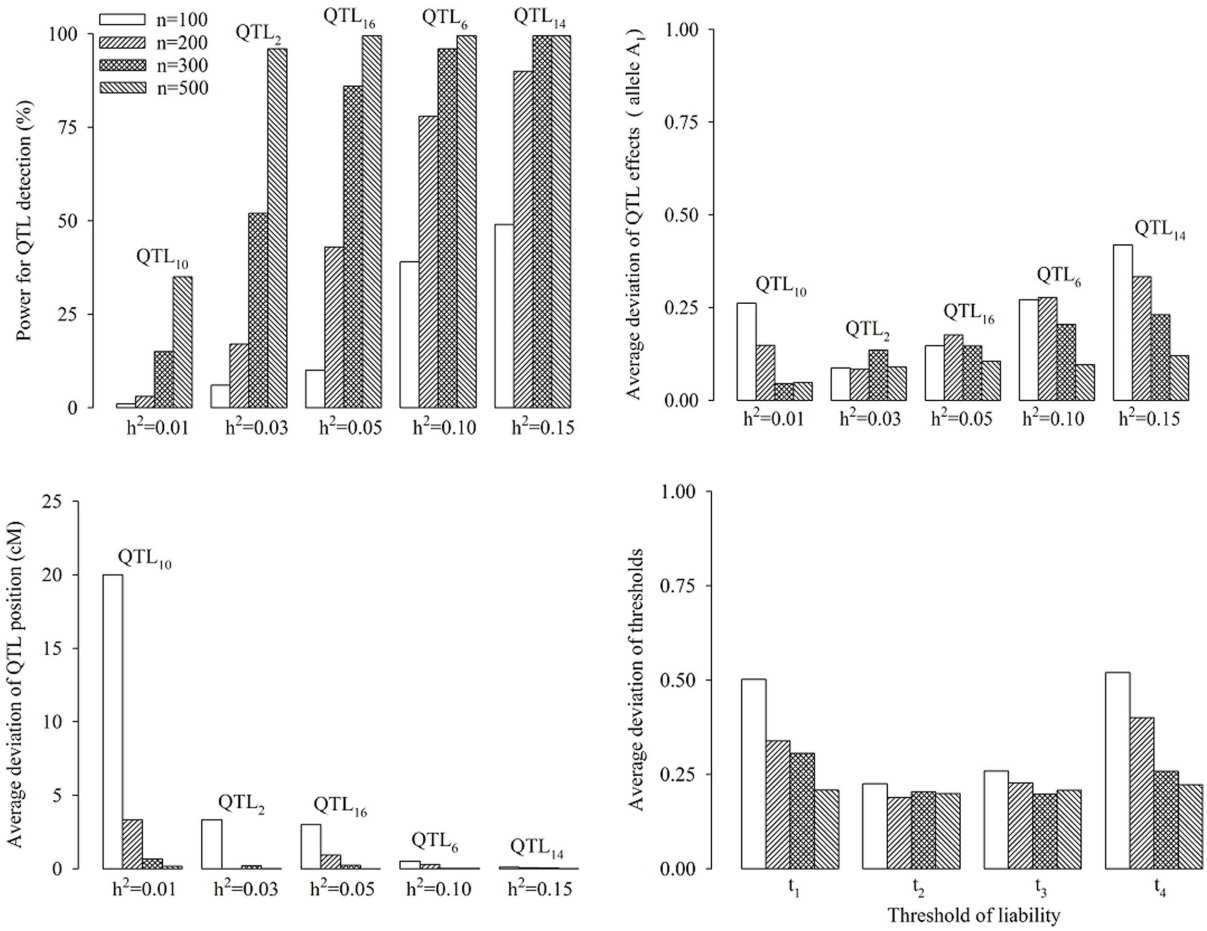

Figure 5. Effect of sample size on association mapping for ordinal traits.
doi:10.1371/journal.pone.0059541.g005

$(\tau,a)=(1,1.5)$, the method works well. The joint posterior distribution has a form of

$$
\begin{aligned}
&P(\phi|\mathrm{Y})\\
&=\prod_{i=1}^{n}\sum_{c=1}^{C}\left\{\mathrm{I}_{y_j=c}\Pr\left(y_j=c|X_j,Z_j,t,\beta,\gamma\right)\right\}\prod_{k=1}^{vq}N\left(\gamma_k|0,\sigma_k^2\right)\\
&\prod_{k=1}^{vq}\chi^{-2}\left(\sigma_k^2|\tau,\omega_k\right)\prod_{k=1}^{vq}Gamma(\omega_k|a,b_k)\prod_{k=1}^{vq}P(b_k)\\
&=\prod_{j=1}^{n}\sum_{c=1}^{C}\mathrm{I}_{y_j=c}\left\{\Phi\left(t_c-X_j\beta-Z_j\gamma\right)-\Phi\left(t_{c-1}-X_j\beta-Z_j\gamma\right)\right\} \qquad (6)\\
&\times\prod_{k=1}^{vq}\left\{\left(2\pi\sigma_k^2\right)^{-\frac{1}{2}}\exp\left(-\frac{\gamma_k^2}{2\sigma_k^2}\right)\right\}\\
&\times\prod_{k=1}^{vq}\left\{\left(\sigma_k^2\right)^{-\frac{\tau+2}{2}}\exp\left(-\frac{\tau\omega_k}{2\sigma_k^2}\right)\frac{(\tau\omega_k/2)^{\tau/2}}{\Gamma(\tau/2)}\right\}\\
&\times\prod_{k=1}^{vq}\left\{\frac{\left(b_k^a\right)}{\Gamma(a)}\omega_k^{a-1}\exp(-b_k\omega_k)\right\}\times\prod_{k=1}^{vq}P(b_k)
\end{aligned}
$$

where $\phi=\left\{\theta,\sigma_1^2,\cdots,\sigma_{vq}^2,\omega_1,\cdots,\omega_{vq},b_1,\cdots,b_{vq}\right\}$

Parameter estimation

Genetic effect. As shown in Wolfinger and O'Connell [29], $\Pr\left(y_j|X_j,Z_j,\theta\right)$ is an approximate normal distribution $N\left(w_j|\mu_j',\sigma_j'^2,t\right)$, where pseudo-data $w_j=\eta_j-L_j'\big/L_j''$, $\eta_j=X_j\beta+Z_j\gamma$; pseudo-mean $\mu_j'=X_j\beta+Z_j\gamma$; pseudo-variance $\sigma_j'^2=-1\big/L_j''$; $L_j=\log\{\Phi(u_c)-\Phi(u_{c-1})\}$, $u_c=t_c-X_j\beta-Z_j\gamma$; $L'_j=\dfrac{dL}{d\eta_j}=\dfrac{\varphi(u_{c-1})-\varphi(u_c)}{\Phi(u_c)-\Phi(u_{c-1})}$;

$$
L''_j=\frac{d^2L}{d\eta_j^2}=
$$

$$
\frac{[\Phi(u_c)-\Phi(u_{c-1})][u_{c-1}\varphi(u_{c-1})-u_c\varphi(u_c)]-[\varphi(u_{c-1})-\varphi(u_c)]^2}{[\Phi(u_c)-\Phi(u_{c-1})]^2}
$$

where $\phi(\bullet)$ is the probability density function of standard normal distribution. The conditional log-posterior distribution, related to γ, is indicated by

Table 3. Simulated parameters in all simulated experiments (3 alleles for marker and QTL, and 3 chromosomes).

Case	Maize pedigree		Marker Density (cM)	Genome length (cM)	Phenotype		QTL	
	No. of founders	No. of Non-founders		No. of categories	Distribution	Position (cM)		Heritability (%)
1	100	200	Equal, 10	100×3	5	1:2:4:2:1	50, 50, 50	5, 10, 15
2	100	200	Equal, 10	100×3	5	1:1:1:1:1; 1:2:4:2:1; 8:5:3:1:1	50, 50, 50	5, 10, 15
3	100	200	Equal, 10	100×3	2,6,9	1:1; 1:3:6:6:3:1;1:2:4:6:9:6:4:2:1	50, 50, 50	5, 10, 15
4	100	200	Equal, 10	100×3	5	1:2:4:2:1	50, 50, 50	5, 10, 15
5	50	300,200,100	Equal, 10	1000×3	5	1:2:4:2:1	90,240,390,540,690,840;80,230,380,530,680,830;120,270,420,570,720,870	1×5,3×5,5×6,10, 15
6	25, 50, 75	200	Equal, 10	1000×3	5	1:2:4:2:1	90,240,390,540,690,840;80,230,380,530,680,830;120,270,420,570,720,870	1×5,3×5,5×6,10,15

doi:10.1371/journal.pone.0059541.t003

$$\sum_{j=1}^{n}\log\left(\sum_{c=1}^{C}\mathrm{I}_{y_j=c}N(w_j|\mu'_j,\sigma'^2_j,t)\right)+\sum_{k=1}^{vq}\log\Pr(\gamma_k|0,\sigma^2_k)$$

Using expectation-maximization empirical Bayes approach of Xu [26], the expectation of the quadratic term required in the maximization step is expressed as

$$E(\gamma_k^2)=[E(\gamma_k)]^2+Var(\gamma_k) \tag{7}$$

where $E(\gamma_k|\mathbf{Y})=\sigma_k^2Z_k^T\mathbf{V}^{-1}(\mathbf{W}-\mathbf{X}\beta)$, $Var(\gamma_k|y)=\sigma_k^2-\sigma_k^2Z_k^T\mathbf{V}^{-1}Z_k\sigma_k^2$, $\mathbf{W}=(\,w_1\quad w_2\quad\cdots\quad w_n\,)'$, $\mathbf{X}=(\,X_1\quad X_2\quad\cdots\quad X_n\,)'$, and $\mathbf{V}=\sum_{k=1}^{qv}Z_kZ'_k\sigma_k^2+diag(\,\sigma'^2_1\quad\sigma'^2_2\quad\cdots\quad\sigma'^2_n\,)$. Once a certain criterion of convergence is satisfied, the converged $E(\gamma_k|\mathbf{Y})$ is the estimate for γ_k.

Genetic effect variances and related hyperparameters. According to joint posterior density in equation (6), conditional posterior distribution is $\sigma_k^2|\gamma_k,\omega_k\sim\chi^{-2}(\tau+1,(\tau\omega_k+\gamma_k^2)/(\tau+1))$ for σ_k^2, $\omega_k|\sigma_k^2,b_k\tilde{}$ $Gamma(\tau/2+a,b_k+\tau/(2\sigma_k^2))$ for ω_k and $b_k|\omega_k\sim Gamma(a,\omega_k)$ for b_k. Here the mode is used to estimate the corresponding parameter, such as,

$$\sigma_k^2=\frac{\tau\omega_k+\gamma_k^2}{\tau+1+2}\quad\omega_k=\frac{\tau/2+a-1}{b_k+\tau/(2\sigma_k^2)}\quad b_k=\frac{a-1}{\omega_k} \tag{8}$$

Non-genetic effect β. Formula for the fixed effect follows the standard procedure of mixed model methodology, we have

$$\beta=\left(\mathrm{X}'\mathrm{V}^{-1}\mathrm{X}\right)^{-1}\left(\mathrm{X}'\mathrm{V}^{-1}\mathrm{W}\right) \tag{9}$$

Thresholds. Using the Newton–Raphson method, the threshold$t$ are estimated by

$$t_c^{(s+1)}=t_c^{(s)}-L'_{t_c}(\theta)\Big/L''_{t_c}(\theta) \tag{10}$$

where $t_c^{(s)}$ is the estimate of parametert_c at the sth iteration,

$$L'_{t_c}(\theta)=\sum_{j=1}^{n}\left\{\frac{\varphi(u_c)}{\Phi(u_c)-\Phi(u_{c-1})}\mathrm{I}_{y_j=c}+\frac{-\varphi(u_c)}{\Phi(u_{c+1})-\Phi(u_c)}\mathrm{I}_{y_j=c+1}\right\}$$

$$L''_{t_c}(\theta)=\sum_{j=1}^{n}\left\{\frac{-u_c\varphi(u_c)[\Phi(u_c)-\Phi(u_{c-1})]-\varphi^2(u_c)}{[\Phi(u_c)-\Phi(u_{c-1})]^2}\mathrm{I}_{y_j=c}+\frac{u_c\varphi(u_c)[\Phi(u_{c+1})-\Phi(u_c)]-\varphi^2(u_c)}{[\Phi(u_{c+1})-\Phi(u_c)]^2}\mathrm{I}_{y_j=c+1}\right\}$$

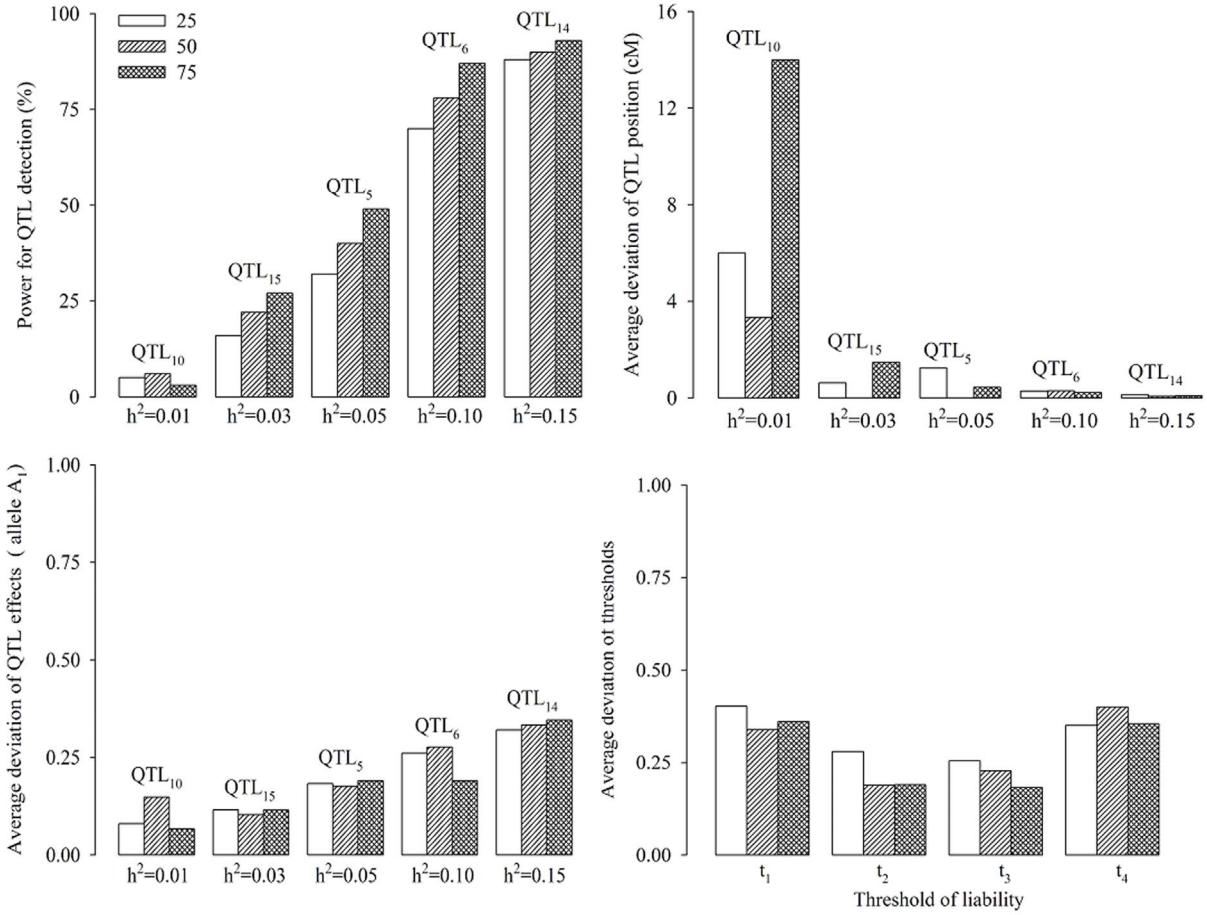

Figure 6. Effect of the number of founders on association mapping for ordinal traits.
doi:10.1371/journal.pone.0059541.g006

Summary of iterations

1. Let $\tau=1$ and $a=1.5$, and provide initial values for φ, for example, let $\gamma_k^{(0)}$ be an uniform random number, $Var\left(\gamma_k^{(0)}\right)=0$, $t_c^{(0)}$ be the quantile of the standard normal distribution based on the phenotypic distribution of y, $\omega_k^{(0)}$ be a gamma random number. $\sigma_k^{2(0)}$ and $b_k^{(0)}$ can be obtained by equation (8).

2. Update σ_k^2, ω_k and b_k using equation (8);

3. Update γ_k using the estimate of $E(\gamma_k|\mathbf{Y})$;

4. Update t_c using equation (10);

5. Update β using equation (9);

6. Repeat step 2 to step 5 until predetermined criterion of convergence is satisfied.

Statistical test

A two-stage selection process in Lü et al. [18] was used to conduct likelihood ratio test (LRT) for all the QTL. In the first stage, all the markers were included in the model. If the estimate of an absolutely allelic effect (environmental interaction effect) at the kth locus ($k=1,\cdots,p$) is greater than 10^{-4}, the kth locus is picked up. In the second stage, we modified the full model only to contain the effects passing the first round of selection. If doing so, we can use the maximum likelihood method to perform the LRT.

The overall null hypothesis is no effect of the oth QTL (or interacted QTL), denoted by $H_0:\gamma_1=\gamma_2=\cdots=\gamma_T=0$, where γ_a is the ath effect for the QTL. If we solve the maximum likelihood estimation of the parameters under the restriction of the H_0 and calculate the log-likelihood value using the solutions with this restriction, we obtain $L(\hat{\theta}'|H_0)$. We can also evaluate the log-likelihood value of the solutions without restrictions and obtain $L(\hat{\theta}')$. Therefore, the LRT statistic is $LOD=\left[L\left(\hat{\theta}\right)-L\left(\hat{\theta}|H_0\right)\right]/2.30$ and the significance threshold of the LOD score was set at 2.0.

Simulation design

We performed six simulation experiments in this study. In the first, the simulated pedigree was similar to the maize pedigree described by Zhang et al. [15]. In current pedigree, the numbers of founders and non-founders were 100 and 200, respectively. Of these, founder lines were in linkage equilibrium so that the genotypes for markers and QTL with three alleles could be simulated. In other words, three alleles for each locus were assigned in equal proportions to each founder. Non-founders were bred via repeated self-pollination of a hybrid between two inbred lines. Thus, each non-founder line represents a RIL with respect to a known pair of parents. The genotypes of all the non-founders could be generated from the genotypes of their parents, analogous to simulating the genotypes of RIL from their parents. All of the non-founder lines could be used to detect QTL. Thirty-three equally spaced markers were simulated on three-chromosome segments 300 cM long. A total of 3 QTL, all of which overlapped with the markers, were placed at 50 cM of each chromosome; the

QTL size, being the proportion of total phenotypic variance explained by the QTL, is 0.05, 0.10 and 0.15, respectively. The allelic effects were calculated by relating the genetic variance of the QTL to the allelic frequencies and effects. The phenotypic value of each line was the sum of the corresponding QTL genotypic values and the residual error, with an assumed normal distribution. These phenotypic values could be transferred into five ordinal categories with four threshold values: −1.2816, −0.5244, 0.5244 and 1.2816. Therefore, the frequencies of the five ordinal categories occurring in all the inbred lines have a ratio of 1:2:4:2:1. Each simulation run consisted of 100 replicates. For each simulated QTL, we counted the samples in which the LOD statistic surpassed 2.0. The ratio of the number of such samples (m) to the total number of replicates (100) represented the empirical power of this QTL. The FPR was calculated as the ratio of the number of false positive effects to the total number of zero effects considered in the full model. The other simulation experiments were performed similarly. All simulated parameters are given in **Table 3**.

A SAS program is available from the authors on request.

Supporting Information

Table S1 Phenotypic values of ATI and STI in 257 soybean cultivars under study.
(DOC)

Author Contributions

Conceived and designed the experiments: YMZ. Performed the experiments: JYF JZ WJZ SBW SFH. Analyzed the data: JYF JZ. Contributed reagents/materials/analysis tools: JYF. Wrote the paper: YMZ JYF.

References

1. Hackett CA, Weller JI (1995) Genetic mapping of quantitative trait loci for traits with ordinal distributions. Biometrics 51: 1252–1263
2. Xu S, Atchley WR (1996) Mapping quantitative trait loci for complex binary diseases using line crosses. Genetics 143: 1417–1424.
3. Rao SQ, Xu S (1998) Mapping quantitative trait loci for categorical traits in four-way crosses. Heredity 81:214–224.
4. Rao SQ, Li X (2000) Strategies for genetic mapping of categorical traits. Genetica 109: 183–197.
5. Xu S, Yi N, Burke D, Galecki A, Miller RA (2003) An EM algorithm for mapping binary disease loci: application to fibrosarcoma in a four-way cross mouse family. Genet Res 82: 127–138.
6. Xu C, Zhang YM, Xu S (2005) An EM algorithm for mapping quantitative resistance loci. Heredity 94: 119–128.
7. Ramalingam J, Sevi A (2010) Mapping and tagging of qualitative traits in crop plants. In: Singh RK, Singh R, Ye GY, et al. Molecular Plant Breeding: Principle, Method and Application. Houston: Studium Press LLC. pp135–159.
8. Coffman CJ, Doerge RW, Simonsen KL, Nichols KM, Duarte CK, et al. (2005) Model selection in binary trait locus mapping. Genetics 170: 1281–1297.
9. Li J, Wang S, Zeng Z-B (2006) Multiple interval mapping for ordinal traits. Genetics 173: 1649–1663.
10. Yi N, Xu S (2000) Bayesian mapping of quantitative trait loci for complex binary traits. Genetics 155: 1391–1403.
11. Yi N, Xu S, George V, Allison DB (2004) Mapping multiple quantitative trait loci for complex ordinal traits. Behav Genet 34: 3–15.
12. Yi N, Banerjee S, Pomp D, Yandell BS (2007) Bayesian mapping of genomewide interacting quantitative trait loci for ordinal traits. Genetics 176: 1855–1864.
13. Wu XL, Gianola D, Weigel K (2009) Bayesian joint mapping of quantitative trait loci for Gaussian and categorical characters in line crosses. Genetica 135: 367–377.
14. Gonzalez-Recio O, Forni S (2011) Genome-wide prediction of discrete traits using Bayesian regressions and machine learning. Genetics Selection Evolution 43(1): 7.
15. Zhang Y-M, Mao Y C, Xie C Q, Smith H, Luo L, et al. (2005) Mapping QTL using naturally occurring genetic variance among commercial inbred lines of maize (*Zea mays* L.). Genetics 169: 2267–2275.
16. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38: 203–208.
17. Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, et al. (2010) Mixed linear model approach adapted for genome-wide association studies. Nat Genet 42: 355–360.
18. Lü HY, Liu XF, Wei SP, Zhang YM (2011) Epistatic association mapping in homozygous crop cultivars. PLoS ONE 6(3): e17773
19. Zhou X, Stephens M (2012) Genome-wide efficient mixed-model analysis for association studies. Nat Genet 44: 821–824.
20. Segura V, Vilhjálmsson BJ, Platt A, Korte A, Seren Ü, et al. (2012) An efficient multi-locus mixed-model approach for genome-wide association studies in structured populations. Nat Genet 44: 825–830.
21. Brisbin A, Weissman MM, Fyer AJ, Hamilton SP, Knowles JA, et al. (2010) Bayesian linkage analysis of categorical traits for arbitrary pedigree designs. PLoS ONE 5: e12307.
22. Diao G, Lin DY (2010) Variance-components methods for linkage and association analysis of ordinal traits in general pedigrees. Genetic Epidemiology 34: 232–237.
23. Iwata H, Ebana K, Fukuoka S, Jannink JL, Hayashi T (2009) Bayesian multilocus association mapping on ordinal and censored traits and its application to the analysis of genetic variation among *Oryza sativa* L. germplasms. Theor Appl Genet 118: 865–880.
24. Hoggart CJ, Whittaker JC, De Iorio M, Balding DJ (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. PLoS Genet 4: e1000130.
25. Zhang YM, Xu S (2005) A penalized maximum likelihood method for estimating epistatic effects of QTL. Heredity 95: 96–104.
26. Xu S (2010) An expectation–maximization algorithm for the Lasso estimation of quantitative trait locus effects. Heredity 105: 483–494.
27. Yi N, Liu N, Zhi D, Li J (2011) Hierarchical generalized linear models for multiple groups of rare and common variants: jointly estimating group and individual-variant effects. PLoS Genet 7: e1002382.
28. Li M (2011) Methodologies for functional mapping of quantitative trait loci and genome-wide association study (Ph D dissertation). Nanjing Agricultural University.
29. Wolfinger R, O'Connell M (1993) Generalized linear mixed models a pseudo-likelihood approach. Journal of Statistical Computation and Simulation 48: 233–243.
30. Gelman A, Carlin JB, Stern HS, Rubin DB (2004) Bayesian data analysis. London: Chapman and Hall/CRC, New York.
31. Mcgilchrist CA (1994) Estimation in generalized mixed models. Journal of Royal Statistical Society, Series B 56: 61–69.
32. Che X, Xu S (2012) Generalized linear mixed models for mapping multiple quantitative trait loci. Heredity 109: 41–49.
33. Zhang WJ (2012) Evaluation and association mapping for soybean salt-alkaline tolerance at seeding stage (Master of Science Dissertation). Nanjing Agricultural University.
34. Lipp M, Brodmann P, Pietsch K, Pauwels J, Anklam E, et al. (1999) IUPAC collaborative trail study of a method to detect genetically modified soybeans and maize in dried powder. Journal of AOAC International 82: 923–928.
35. Xu Y, Li HN, Li GJ, Wang X, Cheng LG, et al. (2011) Mapping quantitative trait loci for seed size traits in soybean (*Glycine max* L. Merr.). Theor Appl Genet 122: 581–594.
36. Wei S-P, Liu X-F, Yang S-X, Lü HY, Niu Y, et al. (2011) Comparison of various clustering methods for population structure in Chinese cultivated soybean (*Glycine max* L. Merr.). Journal of Nanjing Agricultural University 34(2): 13–17.
37. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics, 155(2): 945–959.
38. Evanno G, Regnaut S, Goudet J (2005) Detecting the number of clusters of individuals using the software structure: a simulation study. Molecular Ecology 14:2611–2620.