A Confidence Set Inference Method for Identifying SNPs That Regulate Quantitative Phenotypes

Charalampos Papachristou^a Shili Lin^b

^aDepartment of Mathematics, Physics, and Statistics, University of the Sciences, Philadelphia, Pa., and

^bDepartment of Statistics, Ohio State University, Columbus, Ohio, USA

Key Words

Disease gene localization • Linear mixed models • Genome-wide association studies • Family-based association test • Quantitative traits • Confidence set inference • High-density lipoprotein

Abstract

Aims: We introduce a family-based confidence set inference (CSI) method that can be used in preliminary genome-wide association studies to obtain confidence sets of SNPs that contribute a specific percentage to the additive genetic variance of quantitative traits. **Methods:** Developed in the framework of generalized linear mixed models, the method utilizes data from outbred families of arbitrary size and structure. Through our own simulation study and analysis of the Genetics Analysis Workshop 16 simulated data, we study the properties of our method and compare its performance to that of the family association method described by Chen and Abecasis [Am J Hum Genet 2007;81:913–926]. We also analyze the Framingham Heart Study data to identify SNPs regulating high-density lipoprotein levels. **Results:** The simulation studies demonstrated that CSI yields confidence sets with correct coverage and that it can outperform the method introduced by Chen and Abecasis [Am J Hum Genet 2007; 81:913–926]. Furthermore, we identified five SNPs that po-

tentially regulate high-density lipoprotein levels: rs9989419, rs11586238, rs1754415, rs9355648, and rs9356560. **Conclusion:** The CSI method provides confidence sets of SNPs that contribute to the genetic variance of quantitative traits and is a competitive alternative to currently used family association methods. The approach is particularly useful in genome-wide association studies as it significantly reduces the number of SNPs investigated in follow-up studies.

Copyright © 2012 S. Karger AG, Basel

Introduction

As genome-wide association studies (GWAS) become the standard tools for epidemiological and genetic studies, concerns have emerged regarding the correct interpretation of the results of such analyses. This is especially pertinent when such GWAS are used in search of loci that regulate complex phenotypes such as high blood pressure or high-density lipoprotein (HDL) levels. Such complex traits are thought to be regulated by multiple genetic and environmental factors, each with small to at most modest contribution to the overall phenotype [1–3]. The identification of such loci is hindered by the fact that the power of detecting them is very low given their small effect and the sample sizes available in such studies [4]. In

addition, the large number of tests being performed requires multiplicity adjustment to guard against false signals, which further reduces power [4]. GWAS usually rely on SNPs typically classified as common (minor allele frequency, MAF, >0.05). Even though the recent literature suggests that explanations for complex traits may lie on rare variants [2, 3], there is still a wealth of information that can be mined by carefully reanalyzing existing GWAS data [5]. A cost-effective strategy that can help mitigate the hurdles that plague GWAS is to use GWAS as preliminary steps in search of ‘promising’ loci that are then subjected to further more elaborate and perhaps more powerful analyses such as candidate gene methods [6–9]. Thus, there is a need for methods that can be used as preliminary tools to significantly reduce the number of candidate genes identified, while maintaining sufficient power, so that more cost-effective and time-efficient follow-up studies can be designed.

Currently, there are a limited number of methods that can be used as a first step towards reducing the number of candidate genes to be studied in the follow-up analyses. Most of these methods fall under the general umbrella of intermediate fine mapping approaches [6, 7], and they can be used mainly for binary traits [9–13]. In the context of mapping quantitative trait loci (QTLs), reduction of the genomic region to be followed up is usually achieved through the use of the LOD support intervals [14, 15]. Alternatively, parametric or non-parametric bootstrap approaches can be used to obtain confidence intervals for the positions of putative trait loci [16–18]. Nevertheless, intervals obtained by either of these approaches tend to undercover the QTL [19]. This is particularly the case in situations in which either the sample size is small or the locus has minor contribution to the phenotype [16–18].

We introduce a new confidence set inference (CSI) method that is developed in the framework of linear mixed models (LMMs). Our motivation stems from the abundance of SNPs that densely cover the genome in its entirety. As such, many of these SNPs reside within genes or in their promoter regions. Recent studies have identified a number of SNPs associated with complex traits, some of which are believed to play an important role in the genetics of such traits either by being functional variants [20] or by partaking in the regulatory mechanism of the phenotype [2, 21]. Our CSI approach particularly targets such functional SNPs as it is designed to provide a confidence set of QTLs contributing at least a predetermined percentage to the overall genetic variance of a quantitative phenotype. The method is based on family data and can accommodate families of arbitrary size and

structure. As a family-based method, it maintains the simplicity and elegance of population-based association studies; yet, it resolves the issue of confounding caused by population substructure. Finally, the method is very flexible as it allows one to search for loci that have at least a specific level of contribution to the quantitative trait. As such, it enables the researcher to adjust this contribution level to target different loci depending on the available resources.

In the following sections, we first describe the general framework of the CSI method. We then propose a specific version of the approach developed in the context of general LMMs. Through our own extensive simulation study, we investigate the properties of the method in terms of true coverage probability of the resulting confidence sets. Then, we compare the performance of our method to that of the family association method proposed by Chen and Abecasis [22] (as implemented in Merlin [23]) under a complex trait-regulating scenario such as the one used to generate the simulated data for Problem 3 of the Genetics Analysis Workshop (GAW) 16 [24]. Finally, we use our method to analyze a subset of the Framingham Heart Study (FHS) data provided as Problem 2 of the GAW16 in search of QTLs regulating HDL levels.

Methods

The Hypothesis

Let us assume the availability of the genetic map consisting of M markers (SNPs). For each SNP m on the genome, consider testing at level α that it contributes to the quantitative phenotype at least $h_0 \times 100\%$ of the total genetic variance. In other words, consider performing the following α -level hypothesis test

$$H_0: \sigma_{gm}^2 \geq h_0 \sigma_g^2 \text{ vs. } H_a: \sigma_{gm}^2 < h_0 \sigma_g^2, \quad (1)$$

where σ_g^2 and σ_{gm}^2 are the total and the m -th locus-specific genetic variance at the trait, respectively. For a fixed value h ($0 \leq h \leq 1$), one can construct a $1 - \alpha$ confidence set of markers (loci) that contribute to the trait at least $h \times 100\%$ of the total genetic variance by first testing each marker (SNP) on the genetic map individually, and then aggregating all those markers for which the above null hypothesis is not rejected at level α .

In the following sections, we describe a method developed in the context of the LMMs that can be used to test the above hypothesis for SNPs without the need for multiple testing adjustments.

Modeling of the Phenotype

Consider a family of arbitrary structure consisting of n members. Let $Y = (y_1, \dots, y_n)$ be the vector of the family phenotypic values of the trait under study. We assume that for the person i , the expression level of the phenotype is regulated by an underlying

(unobserved) random polygenic effect u_i , some environmental effect e_i , and (potentially) some known covariates X_i (e.g. age, sex, etc.) as follows:

$$y_i = X_i\beta + u_i + e_i, \quad (2)$$

where β is a vector of unknown coefficients. We further assume that there is no interaction between the covariates, the polygenic, and the environmental effects, and that their effects on the phenotype are strictly additive. Finally, to completely specify our model, we assume that the random effects u follow a multivariate normal distribution with mean zero and some covariance matrix Ω , and the e_i 's are independently and identically distributed as normal with mean also zero and some variance σ_e^2 . Note that under the above assumptions, the joint distribution of the family genotypes Y is multivariate normal with mean $X\beta$ and covariance matrix $\Sigma = \Omega + I\sigma_e^2$, where $X^T = (X_1, \dots, X_n)$ and I is the $n \times n$ identity matrix.

Under additivity across the effects of the trait-contributing loci, and assuming that there is very little or no inbreeding in the family, it can be shown [25, 26] that Ω has (approximately) the following form:

$$\Omega = 2\Phi\sigma_a^2 + V_d\sigma_d^2, \quad (3)$$

where σ_a^2 and σ_d^2 are the total additive and dominance genetic variance in the population, with $\sigma_g^2 = \sigma_a^2 + \sigma_d^2$, Φ is a known matrix whose (i, j) -th element is the kinship coefficient between individuals i and j , and the elements in the matrix V_d are the fraternity coefficients, that is, the probabilities that two individuals share two alleles IBD, with neither of them being autozygous. Furthermore, under the assumption of additivity across loci, the overall additive and dominance genetic variance of the trait can be written as

$$\sigma_a^2 = \sum_{t=1}^L \sigma_{at}^2 \quad \text{and} \quad \sigma_d^2 = \sum_{t=1}^L \sigma_{dt}^2, \quad (4)$$

where σ_{at}^2 and σ_{dt}^2 are the additive and dominance effects due to the trait locus t , and L is the unknown total number of loci contributing to the trait.

Major Gene Effect Model

The model in equation (2) can be easily extended to include effects of major genes by simply including fixed effects corresponding to each of the genes. Suppose that we are studying a diallelic trait-contributing locus. Let z_{ti} be the number of copies of the minor frequency allele that the i -th member of the family carries at the t -th trait-regulating locus, and consider the following model of the total phenotypic value of this individual

$$y_i = X_i\beta + \gamma_t z_{ti} + u_{-ti} + e_i, \quad (5)$$

where y_i , X_i , e_i , and β are defined as before, γ_t denotes the effect of the trait locus t , and u_{-ti} is the polygenic effect due to all trait-contributing loci but locus t . It can be shown that the covariance matrix Ω_{-t} that corresponds to the joint distribution of the random polygenic effects u_{-ti} 's has exactly the same form as in equation (3) with $\sigma_{-at}^2 = \sum_{l \neq t} \sigma_{al}^2$ and $\sigma_{-dt}^2 = \sum_{l \neq t} \sigma_{dl}^2$ [26].

Under additivity between the alleles for the QTL of interest t , Haseman and Elston [27] showed that

$$\sigma_{gt}^2 = \sigma_{at}^2 = 2\gamma_t^2 p_t(1 - p_t),$$

where σ_{gt}^2 is the total genetic variance due to locus t , and $p_t (< 0.5)$ is the MAF of the trait locus. Now, if we assume that the phenotypic value increases with the number of minor alleles of the major trait locus, that is, $\gamma_t > 0$, then the above equation implies that

$$\gamma_t = \sigma_{at} / \sqrt{2p_t(1 - p_t)}.$$

Hence, the hypothesis test in equation (1) becomes equivalent to

$$\begin{aligned} H_0 : \gamma_m &\geq \sqrt{h_0 \sigma_a^2 / 2p_m(1 - p_m)} \\ \text{vs.} \\ H_a : \gamma_m &< \sqrt{h_0 \sigma_a^2 / 2p_m(1 - p_m)}. \end{aligned} \quad (6)$$

where p_m is the MAF of the marker m .

To be able to test the above H_0 versus H_a , we need to know the total additive genetic variance of the trait σ_a^2 . Even though σ_a^2 is usually unknown, an estimate can be easily obtained from the data themselves by fitting the null model in equation (2), i.e., the model without any major gene effects. We will describe a likelihood method for obtaining such an estimate in the 'Estimation of the Genetic Parameters' section below.

Construction of the Confidence Sets

Let $\hat{\gamma}_m$ and $s(\hat{\gamma}_m)$ be the maximum likelihood estimate (MLE) of γ_m and its estimated standard error, respectively, both obtained by maximizing the full model in equation (5) as discussed in the next section. Then the following statistics can be used to test the hypotheses in equation (6):

$$T_m = \frac{1}{s(\hat{\gamma}_m)} \left[\hat{\gamma}_m - \sqrt{h_0 \sigma_a^2 / 2p_m(1 - p_m)} \right]. \quad (7)$$

Assuming a large enough sample size, the asymptotic distribution of the MLE $\hat{\gamma}_m$ is normal, and thus, under the null hypothesis T_m follows a standard normal. Using this fact, the confidence set is constructed as follows. First, we obtain an estimate, σ_a^2 , of the overall genetic variance by fitting the model in equation (2) without any effects of major genes. Next, for each SNP m on the map individually one at a time, we obtain the MLE of γ_m and its standard deviation, $\hat{\gamma}_m$ and $s(\hat{\gamma}_m)$, respectively, by maximizing the likelihood in equation (5). Then, based on the estimated information and the given value of h_0 , for each SNP m we compute the test statistic T_m . Finally, the confidence set is constructed by aggregating all those SNPs for which $|T_m| \geq z_\alpha$, where z_α is the upper α percentile of the standard normal distribution.

Alternatively, we can construct the confidence set as follows. First, for a given α and for each marker m on the map, one can solve the equation $T_m \geq z_\alpha$ with respect to h_0 to obtain an upper confidence bound, $h_m^{\max}$, for the contribution level of the SNP to the variance of the trait as follows:

$$h_m^{\max} = \frac{2}{\sigma_a^2} p_m(1 - p_m) [\hat{\gamma}_m - z_\alpha s(\hat{\gamma}_m)]^2. \quad (8)$$

Then, a $(1 - \alpha)100\%$ confidence set of SNPs contributing at least $100 \times h_0\%$ to the phenotype of interest will consist of all markers for which $h_m^{\max} \geq h_0$.

In practice, a level h_0 deemed reasonable by the researcher, say 2 or 3%, may result in a confidence set that includes a substantial number of SNPs. However, the limited resources available may prohibit further investigation of all identified SNPs. In such cases, a more practical strategy for prioritizing SNPs would be as fol-

lows. For each SNP, the researchers can obtain a 95% upper confidence bound for its contribution to the total additive variance of the trait, as described above. They can then sort the SNPs in decreasing order according to these bounds and choose the top loci for further investigation. Even though the selection of SNPs can be done solely based on their rank in the list, biological relevance and/or information from prior studies can also be used to identify the best subset of SNPs for further investigation. The exact number of such SNPs to be selected will usually be dictated by the available resources, which will set limitations as to how many SNPs can be followed up. Note that such a practical approach is very similar to that often used in GWAS where all markers are first ranked based on the significance level of their corresponding test statistic, and then the top markers on the list are selected for follow-up analysis [28].

We further note that equation (8) implies that the estimated contribution of a SNP to the phenotype depends on the frequency of its alleles. Different MAFs can result in significantly different h_m^{max} , thereby helping reduce the effect of high linkage disequilibrium (LD) on the resulting confidence sets [29]. This is particularly useful as it can help us distinguish the true QTL from nearby SNPs in high LD with it, especially if the latter have significantly different MAFs than the QTL.

Estimation of the Genetic Parameters – Likelihood Function

Let us assume that we have available information on F families of arbitrary size and structure. Let Φ_f and V_{df} be the matrix of the kinship and fraternity coefficients for the f -th family, $f = 1, \dots, F$. Then, the log-likelihood function that corresponds to the model in equation (2) is given by

$$\begin{aligned} \log L(\beta, \sigma_a, \sigma_d, \sigma_e) = & \sum_{f=1}^F \left\{ -\frac{n_f}{2} \log 2\pi - \frac{1}{2} \log |\Omega_f| - \frac{1}{2} (Y_f - X_f \beta)^T \Omega_f^{-1} (Y_f - X_f \beta) \right\} \\ & \propto \sum_{f=1}^F \left\{ -\frac{1}{2} \log |2\Phi_f \sigma_a^2 + V_{df} \sigma_d^2 + I_{n_f} \sigma_e^2| - \frac{1}{2} (Y_f - X_f \beta)^T \right. \\ & \quad \left. (2\Phi_f \sigma_a^2 + V_{df} \sigma_d^2 + I_{n_f} \sigma_e^2)^{-1} (Y_f - X_f \beta) \right\}, \end{aligned} \quad (9)$$

where n_f is the size of the f -th family, Y_f is the vector of the family phenotypes, X_f is a matrix whose rows are the values of the covariates for each family member and $I\{n_f\}$ is the n_f dimensional identity matrix. Maximization of this function can be easily achieved through the use of existing software packages designed to fit LMMs. Using a similar procedure, one can also obtain the log-likelihood function for the model that includes the major gene locus effect (model given in equation (5)).

A Simulation Study: Effect of Model Misspecification

Two important assumptions are required for the CSI method to work: (a) the availability of the total genetic variance components, and (b) the putative QTL contributing only to the additive variance of the phenotype. In practice, the values of the genetic variance components

are estimated from the data themselves, while the QTL may also contribute to the dominance variance of the trait. Through a small simulation under several simple trait-regulating scenarios, we studied the coverage probability of the CSI confidence sets under violations of these two assumptions.

We considered a genetic map consisting of two chromosomes densely covered by 400 SNPs with an average intermarker distance of 0.1 cM. The MAF for the markers ranged from 0.10 to 0.48. Marker genotypes were generated using the software package SIMPED [30], which allows for LD between blocks of markers on the same chromosome. In our simulations, markers were arranged in LD blocks of 20 SNPs. Markers within a block were in high LD with each other (D' between 0.6 and 1), while the strength of the LD reduced with the increase in the distance between markers. We considered several two-locus trait-regulating models, all devised to follow the observed variance components of the cholesterol level phenotype estimated from the FHS data: heritability of 33% (32% attributed to the total additive and 1% to the total dominance genetic variance). The total variance (combined genetic and environmental) of the trait was 1,309. Under all simulation scenarios, the phenotype was regulated by two independently segregating major loci, a polygenic effect with additive and (potentially) dominance components, and some environmental factors manifested as a random environmental component. The two major loci were diallelic, in Hardy-Weinberg and linkage equilibrium with each other, and interacted in an additive manner. Of the two QTLs, one had a MAF of 0.10 and the other of 0.20. The first locus (L1) contributed only additive variance to the phenotype, while the second locus (L2) contributed both additive and dominance variance. Each locus contributed 1% of the total additive genetic variance of the trait, or equivalently 0.3% of the total variance (genetic and environmental). For the second locus that contributed also to the total dominance genetic variance, we considered varying degrees of contribution: 17.8, 50, and 100%. Finally, the two trait loci L1 and L2 were set to be the 210th markers on chromosomes 1 and 2, respectively.

We generated phenotypes and genotypes for 500 replicates for three sample sizes: 500, 1,000, and 1,500 nuclear families with two children, for a total of 1,000, 4,000, and 6,000 individuals, respectively. We then used our CSI method to obtain 95% confidence sets for loci that contribute 1% to the total additive variance, that is, $h_0 = 0.01$, which corresponds to the actual contribution of each of the loci under the simulating model. For each replicate,

the overall additive variance of the trait needed for performing the hypothesis tests in equation (6) was estimated from the data themselves as described earlier.

In table 1, we summarize the simulation results for the model for which L2 contributed roughly 17.8% of the total dominance genetic variance. The results for the other models were very similar and thus they are omitted. For each sample size, we report the locus-specific true discovery rate (TDR) for each major QTL and the corresponding observed average ($\bar{N}_S$) and standard deviation (s_{N_S}) of the number of SNPs included in the confidence set of markers. The locus-specific TDR is defined as the proportion of replicates, out of the 500, that yielded a confidence set that included the specific trait locus. From the table, we can see that the observed coverage probabilities of the confidence sets for both loci (TDR columns in the table) match very well with the nominal level of 95%. This was expected as, under the simulation model, both loci contribute exactly 1% to the total additive variance at the trait. It seems that estimating the overall variance components from the data themselves has minimal effect on the coverage probability of the resulting confidence sets. Furthermore, the observed coverage probabilities for the second loci (TDR column under chromosome 2) suggest that violation of additivity within the trait locus may not significantly impact the performance of the CSI method, as long as the threshold h_0 correctly reflects the contribution of the locus on the additive variance of the trait.

GAW16 Simulated Data

The GAW16 Problem 3 simulated data [24] provide an excellent opportunity to investigate the performance of our method in preliminary genome-wide association scans, and also to objectively compare it to the widely used family-based association method proposed by Chen and Abecasis [22] as implemented in Merlin [23]. The simulated data are based on a subset of the FHS data set. More specifically, using the genotypes on 500K SNPs for 6,848 individuals from the FHS data, Kraja et al. [24] generated 200 replicates of various simulated phenotypes that contribute to coronary heart disease. Of the 6,848 individuals, 6,621 were members of 1 of 766 pedigrees, and the remaining 227 were unrelated individuals. Each replicate imitated a longitudinal study that recorded information on the phenotypes of interest at three different time points, 10 years apart. Note that the genotypes for all individuals were the same for all replicates. From all the available phenotypes, we focused on the simulated

Table 1. Simulation results using as threshold of the genetic contribution $h_0 = 0.01$, the threshold that reflects the actual contribution of each of the trait loci to the additive genetic variance under the simulation models

$N_f(N_p)^b$	Chromosome 1 ^a			Chromosome 2 ^a		
	TDR ^c	N_S^d	$s_{N_S}^d$	TDR ^c	N_S^d	$s_{N_S}^d$
500 (2,000)	93.8	196.7	41.0	95.0	197.1	40.1
1,000 (4,000)	96.6	40.8	15.4	97.0	43.1	14.8
1,500 (6,000)	96.2	9.0	4.7	94.0	10.4	5.0

^a QTL L1 located on chromosome 1 contributes 1% to the total additive genetic variance, while QTL L2 located on chromosome 2 contributes 1% of the total additive genetic variance and 17.8% to the total dominance genetic variance.

^b Number of families in the study (N_f). In parentheses, the total number of individuals in the study (N_p).

^c Percentage of replicates, out of 500, that the 95% confidence set included the trait locus.

^d Observed mean ($\bar{N}_S$) and standard deviation (s_{N_S}) of the number of SNPs included in the 95% confidence sets.

trait of HDL levels. Under the simulation model, there were five major SNP QTLs. One of these QTLs contributed 1.7% of the total additive genetic variance of the trait (or 1% of the total variance), three contributed 0.5% (or 0.3% of the total), and one contributed 0.33% (or 0.2% of the total). In addition, there were 1,000 SNPs with smaller individual effects scattered all over the genome, and they jointly contributed 58% of the total variability of the trait, with individual contributions ranging from 0.03 to 0.15%.

We preprocessed the data to remove individuals who were either not genotyped or had no phenotypic values. Extended families were appropriately split and trimmed to smaller units so that all individuals in the resulting subfamilies were genotyped. This resulted in a sample of 441 families of 1,906 individuals in total. To avoid issues with ill-distributed markers, SNPs with rare minor alleles (MAF <1%) were excluded from the analyses, leaving 125,735 SNPs spanning the entire genome. As the data were of longitudinal nature, for each of the individuals in the sample, the phenotypic values for the trait and all covariates were taken to be those from the last recorded visit of the person. For the analyses, we included in the model as fixed covariates the age of the person and whether the person was medicated for coronary heart disease at the last recorded visit. For each replicate, the overall additive variance needed for the construction of the confi-

Fig. 1. GAW16 simulated data: observed TDRs of the CSI method and the method introduced by Chen and Abecasis [22] (CA) as functions of the corresponding observed FDRs.

dence set by the CSI method was estimated from the data themselves, as described in the Methods section.

Under the simulation model, all chromosomes harbor trait-contributing loci. As such, there is no chromosome that can be used to gauge the false discovery rate (FDR) of the methods. Instead, we devised a definition of what constitutes a true or false signal based on the distance between the inferred ‘linked’ SNP and the closest trait-contributing locus. According to the simulation model, the median distance between consecutive QTLs is 1.5 cM. Thus, we defined any SNP that is located within (more than) 0.5 cM from a QTL as a true (false) positive signal. Likewise, a functional variant (SNP) was considered ‘identified’ if it lies within 0.5 cM from at least one ‘linked’ SNP. Note that for the CSI approach, a SNP is inferred as ‘linked’ if it is included in the 95% confidence set of SNPs contributing at least $100 \times h\%$ to the additive variance of the trait, while for the approach of Chen and Abecasis [22], a SNP is inferred as ‘linked’ if its computed p value is less than the chosen threshold α .

Table 2 and figure 1 summarize the results of the analyses using the CSI method and the method introduced by Chen and Abecasis [22]. The ‘TDR’ column in the table gives the proportion of replicates for which the method ‘identified’ at least one SNP that contributes to the quantitative phenotype. Similarly, the ‘FDR’ column displays the proportion of replicates that yielded ‘signal’, but none of the inferred linked SNPs were within 0.5 cM from any QTL. The two columns under the heading ‘QTLs’ display the mean and standard deviation of the number of functional SNPs identified by the method. Finally, the ‘Un-

Table 2. Results of the analysis of the GAW16 simulated phenotype HDL

Method	h or α^a	TDR ^b	FDR ^c	Genes ^d		Unlinked SNPs ^e	
				N_S	s_{N_S}	N_S	s_{N_S}
CSI	0.044	28.0	25.0	1.1	1.5	1.4	1.2
	0.048	14.5	15.0	1.0	1.3	1.0	0.8
	0.50	8.5	10.0	1.0	1.3	0.8	0.7
	0.054	3.0	5.0	0.9	1.3	0.9	0.6
CA	2.5E-6	11.0	25.0	0.7	1.2	1.0	0.8
	1.3E-6	5.0	15.0	0.6	1.1	0.9	0.6
	7.9E-7	4.0	10.0	0.6	1.1	0.8	0.6
	4.6E-7	3.0	5.0	0.8	1.1	0.7	0.6

CA = Method by Chen and Abecasis.

^a CSI: Threshold for proportion of additive variance; CA: threshold for p value for declaring linkage for the method described by Chen and Abecasis [22].

^b Proportion of replicates, out of 200, that each method ‘identified’ at least one trait locus.

^c Proportion of replicates, out of 200, that each method inferred at least one SNP as linked but failed to ‘identified’ any trait locus.

^d Observed mean ($\bar{N}_S$) and SD (s_{N_S}) of the number of trait loci ‘identified’ per replicate.

^e Observed mean ($\bar{N}_S$) and SD (s_{N_S}) of the number of non-trait-contributing SNPs inferred linked per replicate.

linked SNPs’ section summarizes the number of SNPs that were inferred as linked but were more than 0.5 cM away from any trait-contributing loci. For each method, we present the results for four different thresholds chosen such that they resulted in 0, 5, 10, and 15% observed FDR, which corresponded to $h_0 = 0.068, 0.054, 0.050$, and 0.048 for the CSI method and $\alpha = 2.0E-8, 4.6E-7, 7.9E-7$, and $1.3E-6$ for the method described by Chen and Abecasis [22]. Finally, figure 1 displays the observed TDRs of the two methods as functions of the corresponding observed FDRs.

Figure 1 clearly demonstrates that the CSI method displayed a greater aptitude for discovering QTLs than the approach by Chen and Abecasis [22]. The CSI exhibited at least the same TDRs as the approach by Chen and Abecasis and, in many cases, it achieved twice as much TDRs as Chen and Abecasis’s approach, when both methods had identical FDRs. The TDRs for both methods were low though, but this is not surprising given the fact that each simulated QTL had a low contribution to the heritability of the simulated trait.

Table 3. FHS data analysis results: genes contributing to HDL levels

SNP	Chromosome	Location	MAF	In/near gene	Genetic variance	
					$\hat{h}^a$	UCB ^b
rs9989419	16	55,542,640	0.4460	<i>CETP</i>	0.039	0.061
rs11586238	1	117,064,661	0.4883	<i>CD2, CD58</i>	0.024	0.044
rs1754415	6	158,683,813	0.1064	<i>TULP4</i>	0.019	0.036
rs9355648	6	158,656,799	0.1068	<i>TULP4</i>	0.020	0.036
rs9356560	6	158,621,016	0.5973	<i>TULP4</i>	0.018	0.035

^a Estimated contribution of SNP to the genetic variance of the HDL levels.

^b 95% upper confidence bound (UCB) for the contribution of SNP to the genetic variance of the HDL levels.

HDL Levels in the Framingham Heart Study

In addition to the simulated data from the GAW16, we also analyzed the real FHS data. The quantitative phenotype on which we focused was the plasma HDL levels. As with the simulated data, for the analyses we used the phenotypic values for the trait and all covariates as they were recorded at the last visit of the person. The genotypes for the individuals in this data set are the same as those in the simulated data analyzed in the previous section. Hence, we followed the same preprocessing procedure as in the previous section to clean the data, which resulted in the same number of individuals and the same number of SNPs scanned (see previous section). Again, the genetic variance components were estimated from the data themselves. In a preliminary analysis, we fitted a model that included two genetic variance components (additive and dominance) and four fixed covariates (gender, age, BMI, and whether the person was receiving cholesterol-related medication at the time of the visit). The preliminary analyses had shown that the dominance genetic component was not statistically significant and so it was removed from the model used in the final analysis.

Implementing the practical approach described in the Methods section, we scanned all the SNPs and estimated their contributions to the additive variance and also the corresponding 95% upper confidence bounds. The results for the top five SNPs, sorted in descending order according to the upper confidence bound, are given in table 3. We can see that the 95% upper confidence bounds for those five SNPs are all at or above 3.5%, with the estimated contributions to the genetic variance of about 2%. If resources were available and the investigator was willing to investigate SNPs that contributed less than 2% to

the genetic variance, then additional SNPs could be further included in a follow-up study as the ranked list for all SNPs is available from running CSI.

From table 3 we can see that the SNP at the top of the list was rs9989419, located 55,542,640 base pairs from the beginning of chromosome 16 in a recombination hot spot between LD blocks near the *CETP* gene, which is known to regulate HDL levels [31]. This finding is consistent with results from other studies that have identified this particular SNP as a regulating factor for the phenotype of interest [31–34]. The second-ranked SNP was rs11586238 on chromosome 1, which, to the best of our knowledge, has not been directly linked to HDL levels but has been identified as a risk-contributing locus to rheumatoid arthritis [35]. As recent studies have shown a relationship between HDL levels and rheumatoid arthritis [36, 37], this result from our analysis suggests that this SNP might play a regulatory role in both traits. Nevertheless, to the best of our knowledge, this SNP has not been characterized as an expression quantitative trait locus (eQTL). The last three SNPs in table 3 were all located in chromosome 6 inside the *TULP4* gene located at 158 cM from the beginning of the chromosome. This general region has been implicated in the regulation of HDL in previous studies [38, 39], suggesting that these three SNPs would merit further investigation. Note that two of these SNPs (rs1754415 and rs9355648) were in high LD with each other and had almost identical MAFs, whereas the last SNP (rs9356560) was also in high LD with the other two SNPs but had a significantly different MAF. It is possible that at least one of these SNPs is functional, and therefore a functional analysis experiment to distinguish their roles would be of great interest.

Discussion

The increased use of GWAS based on extremely dense genetic maps consisting of hundreds of thousands of SNPs has prompted the need for methods that can be used in preliminary scans to limit the number of SNPs to be investigated in follow-up studies. We presented a novel CSI method that can be used as a first step in association studies to derive a set of SNPs that contribute at least a specific percentage h_0 to the additive variance of a quantitative trait. The CSI method allows the researchers to control the confidence level with which they can identify genes that have a specific effect on the genetic variance of the quantitative trait of interest. It is exactly this property of our method that makes it particularly relevant in preliminary whole-genome scans as it possesses the ability to reduce the number of SNPs to be followed up and, at the same time, ensures with high confidence that a sufficient number of trait loci with specific contribution to the phenotype are carried over to the subsequent analysis. In addition, our method can easily incorporate information on relevant fixed covariates, and it can handle data from extended families of arbitrary size and structure. Furthermore, as it is a family-based association test, it is expected to be robust with respect to population stratification [40].

The current formulation of the CSI method assumes additivity across the different loci which regulate the trait and within the QTL of interest. These assumption appear to be reasonable as additive diallelic models have been shown to be good approximations to more complex underlying genetic mechanism, even in the presence of epistatic effects, especially when the sample size is fairly large [27, 41–43]. Through a simulation study, we extensively investigated the effect of model misspecification on the true coverage of the derived confidence sets under different sample sizes. The results of the analyses demonstrate that, irrespectively of the sample size, our method provides confidence sets of QTLs with true coverage matching the nominal one, when the level h_0 used correctly reflects the actual contribution of the locus to the additive variance of the phenotype. In addition, it seems that mild violation of the additivity assumption within the locus of interest has no significant effect on the coverage of the resulting confidence sets, thereby indicating the robustness of the method with respect to model misspecification.

For the construction of the confidence sets of QTLs, the CSI method requires the value of the total additive variance of the quantitative phenotype and the MAFs of the SNPs. Our results demonstrated that estimation of

this information from the data themselves seems to bear a negligible effect on the coverage of the resulting confidence sets, even when the sample size is as low as 500 nuclear families, i.e., 2,000 individuals. Finally, the CSI method also needs estimates of the MAFs of the SNPs on the genetic map. Such information is usually readily obtained from the data themselves. However, SNPs with very rare variants may not yield accurate estimates, thereby affecting the results of the analyses. Nevertheless, as GWAS usually rely on common SNPs ($MAF > 0.05$) [2], the effect of ill-distributed markers on the confidence sets obtained through the CSI method should be moderated.

Using the GAW16 simulated data and focusing on the HDL simulated phenotype, we compared the performance of the CSI method to the commonly used pedigree association method introduced by Chen and Abecasis [22] as implemented in Merlin [23] in terms of their FDRs and TDRs. As it was expected, due to the sample size being relatively small, roughly 2,000 individuals, and the simulated QTLs having very miniscule contribution to the phenotype, both methods had very low TDRs. Nevertheless, the results of the analyses gave encouraging results for the CSI method as our method displayed uniformly higher TDRs than the method of Chen and Abecasis [22] when both methods had identical FDRs, suggesting that CSI can potentially outperform other similar family-based association methods. However, we need to note that the particular simulated data sets had several limitations, such as the use of the same genotypes for all 200 available replicate and the absence of epistasis and gene-gene and gene-environment interactions in the determination of the phenotype. As such, generalization of the results to more general settings needs to be made with caution. Whether the behavior of these methods under other settings would remain the same is unclear, and thus a separate, more extensive study is needed to answer this and other related questions.

Analysis of the FHS data in search of loci contributing to the HDL levels pointed to a known genomic area on chromosome 16 that has already been implicated in the regulation of the phenotype, roughly 55.5 cM from the beginning of the chromosome [31–33]. Specifically, we uncovered strong evidence that SNP rs998919 acts as a QTL with an estimated contribution of 3.9% of the additive genetic variance or 1.7% of the total variance of the quantitative phenotype in the Framingham population. We also identified four additional SNPs, each of which can potentially contribute anywhere between 1.8 and 2.4% of the additive genetic variance (or between 0.8 and 0.9% of the total variance) to the HDL levels. Of these four SNPs, SNP rs11586238 resides on chromosome 1 and

could potentially be a co-regulating factor as it has already been implicated in the development of rheumatoid arthritis [35]. The other three SNPs (rs1754415, rs9355648, and rs9356560) are located on chromosome 6 near a region that has been identified as an interesting region in previous studies of HDL [38, 39].

The construction of the confidence sets by the CSI method requires specification of the threshold h_0 for the contribution of the inferred trait loci to the overall additive genetic variance. Selection of the value of h_0 should be done in advance and should take into consideration the amount of the data available. Higher values of h_0 will be needed for small sample sizes so that the FDR can be controlled at acceptable levels. Perhaps a more appealing strategy for practical purposes is to avoid specifying the value of h_0 . The researcher can use the CSI approach to obtain 95% upper confidence bounds for the contribution to the total additive variance of the trait for each of the SNPs on the map. The SNPs can then be sorted in decreasing order according to these bounds, and the top loci can be selected for further investigation. The availability of resources will usually determine the exact number of such SNPs to be followed up.

Software

The software package CSI-QTL implementing the described method, is freely available at <http://code.google.com/p/papachristou-free-genetics-software/downloads/list>.

Acknowledgments

The Framingham Heart Study project is conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University (N01 HC25195). The Genetic Analysis Workshop is supported by NIH grant R01 GM31575. The simulated data was supported by the Washington University Institute of Clinical and Translational Sciences, NIH grant 1U54RR023496. The authors acknowledge the investigators that contributed the phenotype, genotype, and simulated data for this study. This paper was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI.

References

- Yang HH, Hu N, Taylor PR, Lee MP: Whole genome-wide association study using affymetrix SNP chip: a two-stage sequential selection method to identify genes that increase the risk of developing complex diseases. *Methods Mol Med* 2008;141:23–35.
- Manolio TA: Genomewide association studies and assessment of the risk of disease. *N Engl J Med* 2010;363:166–176.
- Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB: Rare variants create synthetic genome-wide associations. *PLoS Biol* 2010;8:e1000294.
- Yoo YJ, Bull SB, Paterson AD, Waggott D, Sun L: Were genome-wide linkage studies a waste of time? Exploiting candidate regions within genome-wide association studies. *Genet Epidemiol* 2010;34:107–118.
- Stewart WC, Peljto AL, Greenberg DA: Multiple subsampling of dense SNP data localizes disease genes with increased precision. *Hum Hered* 2010;69:152–159.
- Bull SB, John S, Briollais L: Fine mapping by linkage and association in nuclear family and case-control designs. *Genet Epidemiol* 2005;29(suppl 1):48–58.
- Papachristou C, Lin S: A comparison of methods for intermediate fine mapping. *Genet Epidemiol* 2006;30:677–689.
- Rohlf RV, Taylor C, Mirea L, Bull SB, Corey M, Anderson AD: One-stage design is empirically more powerful than two-stage design for family-based genome-wide association studies. *BMC Proc* 2007;1(suppl 1):S137.
- Sinha R, Igo RP, Saini SK, Elston RC, Luo Y: Bayesian intervals for linkage locations. *Genet Epidemiol* 2009;33:604–616.
- Liang KY, Chiu YF, Beaty TH: A robust identity-by-descent procedure using affected sib pairs: multipoint mapping for complex diseases. *Hum Hered* 2001;51:64–78.
- Hosjer O: Assessing accuracy in linkage analysis by means of confidence regions. *Genet Epidemiol* 2003;25:59–72.
- Schaid DJ, Sinnwell JP, Thibodeau SN: Robust multipoint identical-by-descent mapping for affected relative pairs. *Am J Hum Genet* 2005;76:128–138.
- Biernacka JM, Sun L, Bull SB: Tests for the presence of two linked disease susceptibility genes. *Genet Epidemiol* 2005;29:389–401.
- Ott J: *Analysis of Human Genetic Linkage*, ed 3. Baltimore, Johns Hopkins University Press, 1999.
- Lander ES, Botstein D: Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics* 1989;121:185–199.
- Walling GA, Visscher PM, Haley CS: A comparison of bootstrap methods to construct confidence intervals in QTL mapping. *Genet Res* 1998;71:171–180.
- Ooijen JW: Accuracy of mapping quantitative trait loci in autogamous species. *Theor Appl Genet* 1992;84:803–811.
- Mangin B, Goffinet B, Rebai A: Constructing confidence intervals for QTL location. *Genetics* 1994;138:1301–1308.
- Bennewitz J, Reinsch N, Kalm E: Improved confidence intervals in quantitative trait loci mapping by permutation bootstrapping. *Genetics* 2002;160:1673–1686.
- Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: Complement factor H polymorphism in age-related macular degeneration. *Science* 2005;308:385–389.
- De Gobbi M, Viprakasit V, Hughes JR, Fisher C, Buckle VJ, Ayyub H, Gibbons RJ, Vernimmen D, Yoshinaga Y, de Jong P, Cheng JF, Rubin EM, Wood WG, Bowden D, Higgs DR: A regulatory SNP causes a human genetic disease by creating a new transcriptional promoter. *Science* 2006;312:1215–1217.

- 22 Chen WM, Abecasis GR: Family-based association tests for genomewide association scans. *Am J Hum Genet* 2007;81:913–926.
- 23 Abecasis GR, Cherny SS, Cookson WO, Cardon LR: Merlin-rapid analysis of dense genetic maps using sparse gene flow trees. *Nat Genet* 2002;30:97–101.
- 24 Kraja AT, Culverhouse R, Daw EW, Wu J, Van Brunt A, Province MA, Borecki IB: The Genetic Analysis Workshop 16 Problem 3: simulation of heritable longitudinal cardiovascular phenotypes based on actual genome-wide single-nucleotide polymorphisms in the Framingham Heart Study. *BMC Proc* 2009;3(suppl 7):S4.
- 25 Jaquard A: The genetic structure of populations. New York: Springer-Verlag, 1974.
- 26 Abney M, McPeck MS, Ober C: Estimation of variance components of quantitative traits in inbred populations. *Am J Hum Genet* 2000;66:629–650.
- 27 Haseman JK, Elston RC: The investigation of linkage between a quantitative trait and a marker locus. *Behav Genet* 1972;2:3–19.
- 28 Kathiresan S, Manning AK, Demissie S, D'Agostino RB, Surti A, Guiducci C, Gianiny L, Burt NP, Melander O, Orho-Melander M, Arnett DK, Peloso GM, Ordovas JM, Cupples LA: A genome-wide association study for blood lipid phenotypes in the Framingham Heart Study. *BMC Med Genet* 2007;8(suppl 1):S17.
- 29 Papachristou C: Confidence set of putative quantitative trait loci in whole genome scans with application to the Genetic Analysis Workshop 17 simulated data. *BMC Proc* 2011;5(suppl 9):S58.
- 30 Leal SM, Yan K, Muller-Myhsok B: SimPed: a simulation program to generate haplotype and genotype data for pedigree structures. *Hum Hered* 2005;60:119–122.
- 31 Heid IM, Boes E, Muller M, Kollerits B, Lamina C, Coassin S, Gieger C, Doring A, Klopp N, Frikke-Schmidt R, Tybjaerg-Hansen A, Brandstatter A, Luchner A, Meitinger T, Wichmann HE, Kronenberg F: Genome-wide association analysis of high-density lipoprotein cholesterol in the population-based KORA study sheds new light on intergenic regions. *Circ Cardiovasc Genet* 2008;1:10–20.
- 32 Ma L, Yang J, Runesha HB, Tanaka T, Ferrucci L, Bandinelli S, Da Y: Genome-wide association analysis of total cholesterol and high-density lipoprotein cholesterol levels using the Framingham heart study data. *BMC Med Genet* 2010;11:55.
- 33 Piccolo SR, Abo RP, Allen-Brady K, Camp NJ, Knight S, Anderson JL, Horne BD: Evaluation of genetic risk scores for lipid levels using genome-wide markers in the Framingham Heart Study. *BMC Proc* 2009;3(suppl 7):S46.
- 34 Waterworth DM, Ricketts SL, Song K, et al: Genetic variants influencing circulating lipid levels and risk of coronary artery disease. *Arterioscler Thromb Vasc Biol* 2010;30:2264–2276.
- 35 Raychaudhuri S, Thomson BP, Remmers EF, et al: Genetic variants at CD28, PRDM1 and CD2/CD58 are associated with rheumatoid arthritis risk. *Nat Genet* 2009;41:1313–1318.
- 36 Toms TE, Panoulas VF, Smith JP, Douglas KM, Metsios GS, Stavropoulos-Kalinoglou A, Kitas GD: Rheumatoid arthritis susceptibility genes associate with lipid levels in patients with rheumatoid arthritis. *Ann Rheum Dis* 2011;70:1025–1032.
- 37 van Halm VP, Nielen MM, Nurmohamed MT, van Schaardenburg D, Reesink HW, Voskuyl AE, Twisk JW, van de Stadt RJ, de Koning MH, Habibuw MR, van der Horst-Bruinsma IE, Dijkmans BA: Lipids and inflammation: serial measurements of the lipid profile of blood donors who later developed rheumatoid arthritis. *Ann Rheum Dis* 2007;66:184–188.
- 38 Arya R, Lehman D, Hunt KJ, Schneider J, Almasy L, Blangero J, Stern MP, Duggirala R: Evidence for bivariate linkage of obesity and HDL-C levels in the Framingham Heart Study. *BMC Genet* 2003;4(suppl 1):S52.
- 39 Charlesworth JC, Peralta JM, Drigalenko E, Goring HH, Almasy L, Dyer TD, Blangero J: Toward the identification of causal genes in complex diseases: a gene-centric joint test of significance combining genomic and transcriptomic data. *BMC Proc* 2009;3(suppl 7):S92.
- 40 Borecki IB, Province MA: Genetic and genomic discovery using family studies. *Circulation* 2008;118:1057–1063.
- 41 Rosenthal EA, Wijsman EM: Joint linkage and segregation analysis under multiallelic trait inheritance: simplifying interpretations for complex traits. *Genet Epidemiol* 2010;34:344–353.
- 42 Vieland VJ, Hodge SE, Greenberg DA: Adequacy of single-locus approximations for linkage analyses of oligogenic traits. *Genet Epidemiol* 1992;9:45–59.
- 43 Vieland VJ, Greenberg DA, Hodge SE: Adequacy of single-locus approximations for linkage analyses of oligogenic traits: extension to multigenerational pedigree structures. *Hum Hered* 1993;43:329–336.