Published in final edited form as:

Hum Hered. 2018 ; 83(4): 175–195. doi:10.1159/000493543.

A Family-Based Rare Haplotype Association Method for Quantitative Traits

Ananda S. Datta¹, Shili Lin², and Swati Biswas¹

¹Department of Mathematical Sciences, University of Texas at Dallas, Richardson, TX 75080.

²Department of Statistics, The Ohio State University, Columbus, OH 43210.

Abstract

Background: The variants identified in genome wide association studies account for only a small fraction of disease heritability. A key to this “missing heritability” is believed to be rare variants. Specifically, we focus on rare haplotype variant (rHTV). The existing methods for detecting rHTV are mostly population-based, and as such, are susceptible to population stratification and admixture, leading to inflated false positive rate. Family-based methods are more robust in this respect.

Methods: We propose a method for detecting rHTVs associated with quantitative traits called family-based Quantitative Bayesian LASSO (famQBL). FamQBL can analyze any type of pedigree and is based on a mixed model framework. We regularize the haplotype effects using Bayesian LASSO and estimate the posterior distributions using Markov chain Monte Carlo methods.

Results: We conduct simulation studies, including analyses of Genetic Analysis Workshop 18 simulated data, to study the properties of famQBL and compare with a standard family-based haplotype association test implemented in the FBAT software. We find famQBL to be more powerful than FBAT with well-controlled false positive rates. We also apply famQBL to the Framingham Heart Study data and detect an rHTV associated with diastolic blood pressure.

Conclusion: FamQBL can help uncover rHTVs associated with quantitative traits.

Keywords

Bayesian LASSO; Framingham Heart Study; Markov chain Monte Carlo; FBAT; quantitative traits; blood pressure; Genetic Analysis Workshop

---

* Address for correspondence: Swati Biswas, PhD, Department of Mathematical Sciences, University of Texas at Dallas, 800 W Campbell Rd., FO 35, Richardson, TX 75080-3021, Tel: (972) 883-6686, swati.biswas@utdallas.edu.

Software

An R code implementing famQBL is available at <http://www.utdallas.edu/~swati.biswas>. It is currently being implemented into an R package, which will be also made available here.

Conflict of Interest

The authors declare that they have no conflict of interest.

Introduction

Rare variants have garnered a great deal of scientific interest in the last decade in the pursuit of explaining missing heritability through the Common Disease Rare Variant (CDRV) hypothesis [1, 2]. Although numerous statistical methods have been proposed for finding rare single nucleotide variants (rSNVs), the most popular one being the so-called “collapsing” methods [3, 4], they rely on next generation sequencing (NGS) data. However, NGS data typically have smaller sample sizes compared to those in genome-wide association studies (GWAS) due to higher costs involved. Moreover, the accuracy of genotype calling in NGS is also an issue [5]. Also, the collapsing methods do not provide any information about which specific variants are important once a significance is declared, which is needed for follow-up validation studies [5]. These limitations pose challenges to fully explore the CDRV hypothesis via NGS data; thus methods that do not rely solely on NGS data and collapsing rSNVs are very much in need.

In this regard, haplotype-based methods are a promising alternative. A haplotype analysis is typically conducted in an association study as a follow-up once approaches based on testing individual single nucleotide polymorphisms (SNPs) indicate that a specific genomic region may harbor associated variants. As haplotypes have biological significance in terms of functionality of a genomic region, detecting associated haplotypes is, in general, of scientific importance [6, 7, 8]. Adding to this, in the past several years, haplotype analysis has also been put forth as a tool for detecting rare variants using GWAS data (in addition to NGS data). As common SNPs from GWAS data combine to form rare haplotype variants (rHTV), we can test the CDRV hypothesis using haplotypes as the basic unit for analysis. Thus, as opposed to a popular view that existing GWAS data are useless for investigating the CDRV hypothesis, they, in fact, provide a tremendous opportunity for exploring rare variants through rHTV analysis by-passing the additional costs involved in collecting NGS data. Although clearly advantageous in a variety of settings, methods for detecting rHTV association has thus far received much lesser attention in this ongoing race to unlock the missing heritability. This is most likely due to the fact that haplotypes are usually not observable (unlike SNVs) and thus the problem is much harder and computationally much more intensive from an analytical point of view. The latter should not be overlooked as more and more data are accumulating at an incredible rate and statistical analysis should not become a bottleneck.

Several haplotype-based methods have been proposed recently for uncovering rare variants [11, 9, 10, 12, 13, 14]. It has been shown that rHTVs are likely to tag rSNVs, and moreover, rHTV methods can detect rSNVs with higher power than their collapsing counterparts in some situations [9, 10, 15, 16, 17]. Broadly speaking, haplotype methods typically test for one of the two types of association: (1) Global: testing overall association of a haplotype block with a disease without testing association with specific (individual) haplotypes in that block (2) Individual haplotype-specific: testing association of specific haplotypes in a block with a disease. As the latter can pin-point specific haplotypes to be associated, it can be more informative for future validation studies. Recently, several haplotype methods of both types have been compared and Logistic Bayesian LASSO (LBL) was shown to be the most powerful among individual haplotype-specific methods [17].

Most of the haplotype association methods proposed to date are predominantly for analyzing population-based data and there is limited work on family-based methods, especially for detecting rHTV. Nonetheless, family-based designs have some unique advantages over their population-based counterparts. Family-based designs are more robust against population admixture and stratification [18, 19]. For finding association with rare variants, a family-based design can have increased power compared to a population-based design as rare variants tend to be enriched in families that have at least one member carrying such a variant [20]. Furthermore, haplotype inference using family data generally provides more reliable results as compared to haplotype inference from unrelated individuals in population data [21]. These features of family-based approaches make them highly attractive for haplotype association studies. Moreover, it has been argued recently that linkage peaks detected by family-based methods may be attributed to rare variants with large effects, and as family data can avoid heterogeneity, they hold great potential for uncovering rare causal variant [22, 23].

A widely used haplotype association method for analyzing family data is implemented in Family-Based Association Test (FBAT) [19]. It can be applied to binary or quantitative phenotype data as well as general pedigree structures; to handle the latter, any given pedigree is broken up into nuclear sub-pedigrees for the analysis. FBAT has also been extended to rare variant analysis; however, this extension is not applicable to testing for association with haplotypes [24]. In recent years, several rHTV methods have been proposed for specific types of family data. For example, Zhu et al. [20] propose a method for analyzing affected sib-pair data by forming a risk haplotype set and using Fisher's exact test. Another recent approach is famLBL (family-based extension of LBL), which uses case-parent triads [25]. However, these two rHTV approaches cannot be used for quantitative traits and general pedigrees. Guo and Shugart [26] have proposed haplotype-based tests for detecting rare variants with general pedigrees using the framework of the Pedigree Disequilibrium Test [27]. In these tests, all haplotypes in a gene (including rare and common) except the most frequent one are collapsed (using weighted combination) and tested as a whole. Thus, these are global haplotype tests and do not provide individual haplotype-specific tests (as provided in FBAT and famLBL).

Our goal here is to develop an individual haplotype-specific test using general pedigree data for quantitative traits. As LBL is one of the best rHTV association method for case-control data [17], it may be fruitful to adapt the LBL methodology for developing such a family-based approach. Indeed, as mentioned above, this strategy was successful in developing famLBL for analyzing triad data and binary trait. Thus, here we propose an approach named family-based Quantitative Bayesian LASSO (famQBL) for detecting rHTV association with quantitative traits. Specifically, we consider a mixed effects model for quantitative trait with random polygenic effects to capture the relationships between family members [28]. We carry out several simulation studies, including the analysis of the Genetic Analysis Workshop (GAW) 18 simulated data [29], to study the properties of famQBL and to compare with the haplotype-based test of FBAT [19]. Finally, we analyze the Framingham Heart Study (FHS) data from GAW 16 [30] and find an rHTV association with diastolic blood pressure. This study was approved by the Institutional Review Board of the University of Texas at Dallas.

Methods

To describe the famQBL approach, we first introduce the relevant notations to represent the family data. Then we describe the model and prior distributions. Next we lay out the Markov chain Monte Carlo (MCMC) steps for estimating the posterior distributions. Finally we describe the inference procedure for association. Much of these models and methodologies have been adapted from [28] and [12].

Data and Notations

Suppose we have m pedigrees in a sample with the j th pedigree consisting of n_j related individuals, $j = 1, \dots, m$. Let $N = \sum_{j=1}^m n_j$ be the total number of individuals in the m pedigrees. Suppose there are L SNPs, which form a haplotype block with p possible haplotypes $\{H_1, \dots, H_p\}$ in the population under study.

Throughout this article, we represent the i th member in the j th pedigree by $(ij)^{th}$ individual, $i = 1, \dots, n_j$, $j = 1, \dots, m$. Let G_{ij} and z_{ij} denote the observed genotypic data and missing (phased) haplotype pair of the $(ij)^{th}$ individual, $\mathbf{G}_j = (G_{1j}, \dots, G_{n_jj})$, $\mathbf{Z}_j = (Z_{1j}, Z_{2j}, \dots, Z_{n_jj})$, and $\mathbf{Z} = (\mathbf{Z}_1, \mathbf{Z}_2, \dots, \mathbf{Z}_m)$. Note that z_{ij} 's are typically unobservable because the phase information is generally not deductible unambiguously from the genotype data of an individual. However, for family data, haplotype pairs of some members may be inferred unambiguously using the genotypes of their relatives.

For the $(ij)^{th}$ individual, denote y_{ij} as the quantitative trait value, u_{ij} as the unobserved polygenic effect, and ε_{ij} as the residual (the random environmental effect). Let

$$\mathbf{Y}_j = (y_{1j}, \dots, y_{n_jj})', \mathbf{U}_j = (u_{1j}, \dots, u_{n_jj})', \boldsymbol{\varepsilon}_j = (\varepsilon_{1j}, \dots, \varepsilon_{n_jj})', \mathbf{Y} = (\mathbf{Y}_1, \dots, \mathbf{Y}_m)', \mathbf{U} = (\mathbf{U}_1, \dots, \mathbf{U}_m)', \text{ and } \boldsymbol{\varepsilon} = (\boldsymbol{\varepsilon}_1, \dots, \boldsymbol{\varepsilon}_m)'.$$

Model, Assumptions, and Likelihood

We consider a mixed effects model framework and assume that the quantitative trait value is determined by three components, namely, the fixed effects of the haplotypes, an additive random polygenic component, and a component due to random environmental effect. Let $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_{p-1})'$ denote the intercept and the effects of haplotypes in a block with p haplotypes. Without loss of generality, we assume the p^{th} haplotype H_p to be the baseline. Thus, the trait value of the $(ij)^{th}$ individual can be written as

$$y_{ij} = \mu_{z_{ij}} + u_{ij} + \varepsilon_{ij}, \quad i = 1, \dots, n_j, \quad j = 1, \dots, m, \quad (1)$$

where $\mu_{z_{ij}} = x_{z_{ij}} \boldsymbol{\beta}$, where $x_{z_{ij}} = \{1, h_{(ij)1}, h_{(ij)2}, \dots, h_{(ij)p-1}\}$ is the $(ij)^{th}$ row of the design matrix, $h_{(ij)k}$ is the number of copies of the haplotype H_k in haplotype pair z_{ij} of the $(ij)^{th}$

individual, and $k = 1, \dots, p - 1$. With this specification, we are assuming that the haplotype effects are additive.

We assume $\mathbf{U}_j, j = 1, \dots, m$, are independent, and each follows an independent multivariate normal distribution $N(0, \sigma_a^2 \Sigma_j)$, where Σ_j is the relationship matrix of dimension n_j and σ_a^2 is the additive variance. The random environmental effect vector ϵ is assumed to follow a normal distribution with mean $\mathbf{0}$ and covariance $\sigma_e^2 \mathbf{I}$, where $\mathbf{I}$ is an identity matrix of dimension N and σ_e^2 is the residual or environmental variance. We assume $\mathbf{U}$ and ϵ are independent and $\mathbf{Z}$ and ϵ are also independent. Moreover, $\mathbf{Z}$ and $\mathbf{U}$ are also assumed to be independent given the model parameters. We also assume no recombination between SNPs making up the haplotype block.

Let $\Phi = (\beta, v, \sigma_a^2, \sigma_e^2)$ be the parameter vector of the model, where v consists of the parameters associated with haplotype frequencies, which will be described more explicitly later in our formulation. The complete data likelihood can be written as

$$L_c(\Phi) = f(\mathbf{Y}, \mathbf{Z}, \mathbf{U} | \Phi) = f(\mathbf{Y} | \mathbf{Z}, \mathbf{U}, \Phi) f(\mathbf{Z} | \Phi) f(\mathbf{U} | \Phi).$$

Note that the information from the observed genotype data, $\mathbf{G}_j, j = 1, \dots, m$, is contained completely within $\mathbf{Z}$. It is also noted that with the above assumptions, the quantitative trait values are conditionally independent given the genotypes and polygenic values, and follow a normal distribution. Thus,

$$f(\mathbf{Y} | \mathbf{Z}, \mathbf{U}, \Phi) = \prod_{j=1}^m \prod_{i \in |O_j|} \frac{1}{\sqrt{2\pi}\sigma_e} \exp \left[-\frac{(y_{ij} - \mu_{z_{ij}} - u_{ij})^2}{2\sigma_e^2} \right],$$

where $|O_j|$ is the set of individuals in the j^{th} pedigree whose trait values are observed. The likelihood contribution of the $(ij)^{\text{th}}$ individual is 1 when y_{ij} is missing. That is, an individual with missing trait value does not contribute to the likelihood, and hence has no effect on the results.

For modeling $f(\mathbf{Z} | \Phi)$, let $\mathbf{f} = (f_1, \dots, f_p)$ denote the haplotype frequencies of the p haplotypes. Then the haplotype configuration vector $\mathbf{Z}$ can be modeled as

$$f(\mathbf{Z} | \Phi) = \prod_{(ij) \in F} P(z_{ij} | v) \prod_{(ij) \in \bar{F}} P(z_{ij} | z_{F_{ij}}, z_{M_{ij}}),$$

where F is the set of founders and $\bar{F}$ is the set of non-founders, and $z_{F_{ij}}$ and $z_{M_{ij}}$ are the phased haplotype pairs of the father and mother of the $(ij)^{\text{th}}$ individual. Thus, the first product is over the probabilities of the haplotype pairs of founders F and the second product

is over the transmission probabilities of haplotype pairs of non-founders $\bar{F}$. The probability of the haplotype pair of a founder is given by

$$P(z_{ij} = z_k/z_{k'}|v) = a_{z_{ij}}(v) = \delta_{kk'}d f_k + (2 - \delta_{kk'})(1 - d)f_k f_{k'}, \quad (2)$$

where z_k and $z_{k'}$ are the haplotypes making up the haplotype pair z_{ij} , with f_k and $f_{k'}$ being their respective frequencies; $\delta_{kk'} = 1(0)$ if $k = k'(k \neq k')$; and $v = \{f, d\}$. Here $d \in (-1, 1)$ is the within-population inbreeding coefficient and is used for capturing excess/reduction of homozygosity [31]. That is, d measures deviation from the Hardy-Weinberg Equilibrium (HWE). Note that when $d = 0$, $a_{z_{ij}}(v)$ specified in (2) corresponds to HWE; otherwise, it is one way of modeling Hardy-Weinberg disequilibrium. The transmission probability $P(z_{ij}|z_{F_{ij}}, z_{M_{ij}})$ can be calculated using Mendelian rules of inheritance. The distribution of the vector of polygenic values, $\mathbf{U}$, is

$$f(\mathbf{U}|\Phi) = \prod_{j=1}^m \left(\frac{2\pi\sigma_a^2}{a} \right)^{-n_j/2} |\Sigma_j|^{-1/2} \exp \left\{ -\frac{1}{2\sigma_a^2} \mathbf{U}_j' \Sigma_j^{-1} \mathbf{U}_j \right\}.$$

Prior Specification, Posterior Distribution, and MCMC Sampling

In order to weed out the unassociated haplotypes and allow the associated ones, especially the rare ones, to stand out, the regression coefficients, β s, need to be regularized. Following Biswas and Lin [12], we use Bayesian LASSO for this purpose, that is, we assign to each β a prior distribution of double exponential with mean 0 and variance $2/\lambda^2$, where the parameter λ controls the degree of penalty. We let λ follow a Gamma(a, b) distribution with $a = b = 20$, which leads to the unconditional standard deviation (SD) of β to be 1.53 [12].

For the additive variance σ_a^2 and the residual variance σ_e^2 , following [28], we assign conjugate Inverse Gamma (IG) priors. In particular,

$$\pi(\sigma_a^2) \propto (\sigma_a^2)^{-a_0/2} \exp \left\{ \frac{-s_0}{2\sigma_a^2} \right\},$$

where a_0 and s_0 are hyperparameters. Note that this parametrization of IG is slightly different from the standard one. Comparing with the standard IG prior with parameters α and β (which is proportional to $x^{-\alpha-1} \exp\{-\beta/x\}$), we have $a_0/2 = \alpha + 1$ and $\beta = s_0/2$. One suggested value of α and β for IG priors is $\alpha = \beta = 1$ [32], which gives non-informative prior. We use these values by setting $a_0 = 4$ and $s_0 = 2$ in our analyses. The prior for σ_e^2 is also taken to be IG with parameters $e_0 = 4$ and $t_0 = 2$.

For the priors of $\mathbf{f}$ and d , note that $a_{z_{ij}}^{(v)}$, defined in equation (2), must be non-negative. This imposes the constraint that $d > \{-f_k/(1-f_k)\}$ for all k ; with the further constraint that $d < 1$, we have $\max_k \{-f_k/(1-f_k)\} < d < 1$. The prior for $\mathbf{f}$ is set to be Dirichlet(1, 1, ..., 1) consisting of a total of p 1's for the p haplotypes. For d given $\mathbf{f}$, we use Uniform $(\max_k \{-f_k/(1-f_k)\}, 1)$ as the prior distribution. Finally, with all the priors fully specified, the joint posterior distribution of all parameters is given by:

$$\pi(\mathbf{Z}, \mathbf{U}, \Phi | \mathbf{Y}, \mathbf{G}) \propto f(\mathbf{Y} | \mathbf{Z}, \mathbf{U}, \Phi) f(\mathbf{Z} | \Phi) f(\mathbf{U} | \Phi) \pi(\beta | \lambda) \pi(\lambda) \pi(\mathbf{f}) \pi(d | \mathbf{f}) \pi(\sigma_a^2) \pi(\sigma_e^2).$$

We estimate all parameters from the posterior distribution using MCMC sampling methods. The updating steps for all parameters except $\mathbf{Z}_j$ are standard and can be found in the Supplementary File. In the following, we describe how $\mathbf{Z}_j$ is updated.

Updating of $\mathbf{Z}_j$: To ensure that the updated haplotype pair of a family member is compatible with the rest of the family, we update haplotype pairs of all members in the j^{th} family together, that is, we block update $\mathbf{Z}_j$. The conditional posterior distribution for $\mathbf{Z}_j$ is given by

$$\pi(\mathbf{Z}_j | \mathbf{Y}, \mathbf{U}, v, \beta) \propto \prod_{i=1}^{|O_j|} \exp \left[-\frac{(y_{ij} - \mu_{z_{ij}} - u_{ij})^2}{2\sigma_e^2} \right] \left(a_{z_{ij}}^{(v)} \right)^{I_{ij} \in F} \left(P(z_{ij} | z_{F_{ij}}, z_{M_{ij}}) \right)^{I_{ij} \in \bar{F}}.$$

Let L_j be the total number of compatible haplotype configurations for the j^{th} family, where the set of compatible haplotype configurations are obtained using HAPLORE [33]. We find the above conditional probability for each of the haplotype configuration in L_j and use the Gibbs sampler to obtain the updated $\mathbf{Z}_j$. If $L_j = 1$, that is, there is only one compatible haplotype configuration, then this update step is not needed. Note that the $\mathbf{Z}_j$'s are independent across the families; therefore, haplotype configurations are updated for each family separately.

Inference on Association

Once we have estimated the posterior distributions of the model parameters using the MCMC sampling scheme, we draw inference about association by testing for significance of each β coefficient. For this, we can calculate the Bayes Factor (BF), which is defined as the ratio of posterior odds to prior odds of H_a to H_0 . If $\text{BF} > 2$, it can be considered as providing significant evidence for association [12]. Alternatively, a 95% credible set (CS) estimated using MCMC output may be used [14]. If 95% CS for a β excludes 0, then association with that specific haplotype may be inferred. In our simulations, we found $\text{BF} > 2$ to be conservative so we use 95% CS for inference. However, in real data analysis we report both BF and CS. Note that this BF cutoff and CS level are not adjusted for multiple testing. For genome-wide testing, a 99% CS may be more suitable [34]. An estimate of the effect of a haplotype, $\hat{\beta}$, is given by the mean of the posterior distribution of β .

Simulation Study 1

Data generation procedure

To evaluate the performance of famQBL, we conduct a wide variety of simulations. We consider three settings with different haplotype diversities as listed in Table 1 with each haplotype being made up of five SNPs coded as a string of 0s and 1s. Under each setting, we simulate two types of association, each consisting of two causal haplotypes (having $\beta > 0$) with (1) both being rare (RR) and (2) one rare and one common (RC). A haplotype is considered rare if its frequency is less than 5%. Figure 1 (adapted from [35]) shows three types of pedigree structures used in our simulations: (a) nuclear family; (b) three-generation cousin family; and (c) extended pedigree with four generations. A sample consists of one hundred pedigrees of each type. Thus, there are a total of three hundred pedigrees with a total of 2500 individuals in each sample. For each simulation, 500 replicates are generated.

We *jointly* generate the haplotype pairs for all members in each pedigree under HWE using the software SimPed [36]. For this, we provide SimPed with the pedigree structure, number of replicates, number of marker loci, recombination fraction values of all zeros, and haplotype information (number of haplotypes and their frequencies and names). Next we generate the polygenic and environmental effect values randomly from the distributions shown in Table 2. In particular, we set the additive variance σ_a^2 and residual variance σ_e^2 to be 5 and 20, respectively, corresponding to 20% heritability, given by $\sigma_a^2 / (\sigma_a^2 + \sigma_e^2)$. Then we generate the trait values using the mixed model with the β values as given in Table 1, β_0 (intercept) set equal to 50, and generated polygenic and environmental effect values. The trait values are finally standardized by subtracting the sample mean and dividing by sample SD of the trait values. This standardization makes it reasonable to assign to β_0 (baseline trait value) the prior mean of 0 in analyses, and helps in convergence of β_0 in MCMC sampling.

As described in Table 2, we simulate genotypes and phenotypes under three simulation scenarios. Scenario 2 is devised to investigate the robustness of the method to model misspecification by using t distributions rather than normal distributions for data generation. Scenario 3 allows for missing genotypes and phenotypes. In particular, for each pedigree of type (b), either zero or one individual (member 1) is set to be missing randomly with equal probability. For each pedigree of type (c), either zero, one, or two individuals (members 4 and 8) is/are set to be missing randomly with equal probability. For pedigrees of type (a), there is no missing information.

Results

For each sample generated using SimPed, after the phase information is deleted, we run it using HAPLORE [33] to obtain all possible compatible haplotype configurations for each family. Each (multivariate) configuration consists of haplotype pairs of *all* members that are compatible with the pedigree structure and the observed genotypes of all family members. HAPLORE also provides maximum likelihood estimates of haplotype frequencies, which are used as starting values in the MCMC procedures. The most frequent haplotype is set as reference (baseline). With these information from HAPLORE, we analyze the sample using

famQBL. For scenarios 1 and 2, we run a total of 70,000 MCMC iterations including 20,000 burn-in. On the other hand, scenario 3 has to be run for a larger number of iterations (150,000) with same burn-in in order to achieve satisfactory convergence as determined by trace plots of parameter estimates and $\hat{R}$ statistics [37]. Briefly, for making these plots and computing $\hat{R}$ statistics, we run multiple MCMC chains from different starting points. The $\hat{R}$ statistic compares the estimated between-chains and within-chain variances for each parameter, and large discrepancies between these two variances indicate non-convergence. $\hat{R}$ values less than 1.1 or 1.2 typically suffices for convergence [37] especially when the same is supported by the trace plots as well. Supplementary Figure 2 shows the trace plots of the estimators at each iteration obtained from three chains with three different starting points and the corresponding $\hat{R}$ statistics for one particular sample. In addition to applying famQBL, we also analyze each sample with FBAT using the command “hbat -p” for testing haplotype association.

To summarize the simulation results for famQBL and FBAT, we calculate the percentage of times a regression coefficient, and hence the corresponding haplotype, is found to be significant. Thus this percentage is essentially a power or a type I error rate for that haplotype depending on whether the haplotype is truly associated or non-associated. For FBAT, if the p -value of an individual regression coefficient is less than 5%, it is considered to be significant.

Figure 2 shows the results for the three scenarios and three haplotype settings for RR association. The exact numerical values of the type I error rates can be found in Supplementary Table 1. In general, famQBL has high power and well-controlled type I error rates. FBAT has slightly higher type I error rates compared to famQBL (as seen more explicitly in Supplementary Table 1). Further, as seen in Figure 2, famQBL is much more powerful than FBAT in detecting the two rare associated haplotypes. We can see that scenario 3 has reduced power compared to scenario 1, as expected, due to missing data. Interestingly and unexpectedly, scenario 2 has higher powers than scenario 1 for both famQBL and FBAT. Figure 3 shows the results for RC association and Supplementary Table 2 shows the exact numerical values of the type I error rates. The results are similar as in RR association except that the inflated type I error rates for FBAT are now even more pronounced especially in scenario 2 (see Supplementary Table 2). However, as seen in Supplementary Table 2, famQBL appears to be reasonably robust to model mis-specification with false positive rates generally still under control.

Simulation Study 2 — Analysis of GAW18 Simulated Data

In the above simulations, the disease model was haplotype-based. We now consider a SNP-based disease model wherein the traits are governed by SNPs rather than haplotypes. For this, we use GAW18 simulated data in which the causal variants are individual SNPs rather than haplotypes. These data were simulated to mimic real data on hypertension related traits [29]. The underlying trait generating model is extremely complicated, consisting of 1458 functional variants interacting with each other and other factors in a complex manner. Thus, it portrays reality closely and serves as a useful test dataset. We consider two quantitative traits — systolic and diastolic blood pressure (BP) levels, which were recorded at four

different time points. There are 200 replicates, each consisting of the same genotypes but with different (simulated) phenotype values.

We perform separate analyses for the systolic and diastolic BP (SBP and DBP) values recorded at the 1st (baseline) time point. A related variable, hypertension medication status, is taken into consideration in defining the trait values because a person taking such medication must have had a high BP value before taking the medication. Thus, if an individual was under medication for hypertension but had a BP value less than the commonly used threshold value for hypertension (90 for DBP and 140 for SBP), then that individual was assigned the threshold value of the corresponding BP.

The data are available on 20 extended families with 1389 individuals. The family sizes range from 27 to 107 (mean = 69.5 and SD = 22.5). However, there is a large proportion of members with missing genotypes (430 individuals) making the number of compatible haplotypes configurations (obtained using HAPLORE) for some of the families to be extremely large ($\sim 10^5$ or even $\sim 10^6$). Moreover, 540 individuals have missing phenotypes. To alleviate this problem, we extract smaller sized pedigrees from each family that are manageable yet still multi-generational. We use the software PedStr [38] to perform this extraction. Specifically, in PedStr, we provide maximum bit size in the range of 14 – 30 and obtain 20 smaller pedigrees with a total of 333 individuals for our analyses. Among those 333 individuals, there are 46 individuals with missing genotype information and 47 individuals with missing trait values. Most of the sub-pedigrees include multiple generations, and the number of compatible haplotype configurations returned by HAPLORE ranges from tens to thousands (the largest one is 6601).

Following Biswas and Papachristou [39], we focus on analyzing two genomic regions in the MAP4 gene. Each region is of 8000 base pairs (bp) in length and is centered at the following two “causal” (functional) variants from the simulation model: (1) the most frequent variant at location 47957741 bp with minor allele frequency (MAF) = 0.378 (Region 1) and (2) the variant with the strongest effect size at location 48040283 bp with MAF = 0.032 (Region 2). The MAP4 gene has a total of 15 functional variants with MAF ranging from 0.0016 to 0.378. Out of them, five are in Region 1 and two are in Region 2. As in Biswas and Papachristou [39], we included SNPs with MAF > 0.01 in forming the two haplotype blocks in these regions. Each haplotype block consists of a total of nine SNPs. The nine SNPs in each block form 7 and 12 haplotypes in the two regions, respectively, with several of them being rare.

Due to a large proportion of missing data resulting in a large number of compatible haplotype configurations for many families, a substantially larger number of MCMC iterations are needed for convergence. We run a total of 400,000 iterations with 20,000 burn-in. If a haplotype contains at least one causal variant, we consider detection of that haplotype as a true positive (and hence it counts towards the power calculation) while detection of other haplotypes are considered false positive.

Tables 3 and 4 show the SNPs and haplotypes in the two genomic regions as well as the powers and type I error rates for famQBL (using 95% CS) and FBAT (using 5% cutoff for p-

value). FamQBL has much higher power than FBAT. The type I error rates of famQBL are practically zero while FBAT has inflated type I error rates, most notably for Hap 5 in Region 2. To compare powers at similar type I error rates, in Figure 4, we present ROC curves (average type I error rate vs power). For this, we vary the cutoffs of BF for famQBL and of p-value for FBAT. The type I error rates are obtained by permuting the trait values among family members within each family to create a null scenario of no association. The power shown is for the haplotype with the strongest association in each block (Hap 6 in Region 1 and Hap 11 in Region 2). We see that famQBL has high power in both regions and it outperforms FBAT by a substantial margin, consistent with our earlier simulation results. In Supplementary Figure 1, we also plot the average power (over all haplotypes containing causal variants) versus average type I error rate. We see that famQBL has higher average power than FBAT for average type I error rates below 20%.

Application to the Framingham Heart Study Data

To illustrate the utility of famQBL, we analyze the FHS genotype data that were made available in GAW16 and can be accessed through dbGaP [30]. In particular, we consider haplotype blocks that span the SNP rs2229188 at position 86371063 bp of the CYP51A1 gene on chromosome 7. This region has been shown earlier to be associated with the qualitative trait hypertension [15, 40]. Thus, it is of interest to study if blood pressure as a quantitative trait is associated with any haplotype, especially rHTV, in this region. In particular, we consider the trait DBP (in log scale) measured at baseline visit. For individuals with missing baseline DBP value, we used the maximum of the DBP values measured at subsequent visits, if available.

FHS has three cohorts of participants — the original cohort (enrolled between 1948 and 1953), the offspring cohort (enrolled between 1971 and 1975), and the third generation cohort (grandchildren of the original cohort; enrolled between 2002 and 2005). This longitudinal study consists of information on numerous phenotypes potentially related to cardiovascular risk. However, the DNA samples were not collected until late 1980s and 1990s on surviving FHS participants and dense SNP genotyping was conducted in 2007 [30]. Thus, there is a substantial amount of missing genotype data. Most genetic studies using FHS data either use unrelated case-controls extracted from the family data or specific types of relatives such as parent-offspring triads or sib-pairs [15].

As famQBL can handle, in principle, any family structure, we start with the set of all available families. However, due to extensive amount of missing genotype data, we pre-process the data to select pedigrees/sub-pedigrees that can be run in a reasonable amount of time. Specifically, we discard pedigrees in which 90% or more members have missing genotypes. Then we run PedStr [38] to select sub-pedigrees with a manageable number of members with missing genotypes and with either both parents' genotype information available or both missing as this is a requirement for HAPLORE. Each selected pedigree is run using HAPLORE to obtain a list of compatible haplotype pairs for all members. The pedigrees that give more than 10^6 compatible haplotype configurations (due to presence of many individuals with missing genotypes) are discarded as including them in the analysis will lead to an extreme amount of computational burden. Some pedigrees are eliminated due

to genetic errors such as incompatible haplotype between child and parent or obligatory recombinations between SNPs as indicated in error messages by HAPLORE. After pre-processing, the final dataset consists of 2998 individuals from 652 pedigrees. Of these, 804 individuals have completely missing genotype and 818 individuals have missing DBP values.

Due to the large amount of missing data, as in the GAW18 data, we run the MCMC for a total of 400,000 iterations to achieve satisfactory convergence. We analyze five haplotype blocks, each consisting of 5 SNPs, with the target SNP rs2229188 being placed in the first, second, third, fourth, and fifth position in each block. The convergence diagnostic plots and $\hat{R}$ statistics for one of the blocks can be found in Supplementary Figures 3 and 4. Of the five blocks, famQBL detects association in two wherein the target SNP is positioned at the fourth and fifth positions, as reported in Supplementary Tables 3 and 4. In each of these two blocks, an rHTV is found to be strongly associated with DBP as evidenced by a large BF and 95% CS for β excluding 0. As the results from the two blocks are consistent with each other, we also analyze a 6-SNP block consolidating the two 5-SNP blocks by putting the target SNP at the fifth position. The results are in Table 5 with an rHTV 100111 (combining the two significant rare haplotypes found in the two 5-SNP blocks) having a strong protective effect on DBP with β estimated to be -1.57 . Due to the rarity of the haplotype, the corresponding CS is relatively wide, as expected. On the other hand, FBAT detects two significant common haplotypes with p-values less than 0.05. The effects of these haplotypes are relatively modest as reflected in $\hat{\beta}$ values given by famQBL. Although FBAT also reports a p-value of 0 for an extremely rHTV, this result (as well as those for other extreme rHTVs) is not reliable as the number of families informative for this haplotype is 1 (as reported in FBAT output — not shown here), and FBAT authors recommend using 15 informative families as a threshold for validity of results [19]. No other haplotype is found to be significant with either method in the other three 5-SNP blocks.

Discussion

In this article, we have proposed famQBL, a family-based association method for detecting rHTVs associated with quantitative traits. The approach is based on a mixed effects model and allows pedigrees of any structure or size. Our simulation studies indicate that famQBL is a powerful method and controls type I error rates satisfactorily. It performs well even in the presence of moderate amount of missing data (scenario 3) and is reasonably robust to deviation from the normality assumption (scenario 2). Further, our results clearly show that famQBL outperforms haplotype association test of FBAT by a substantial margin when the associated variants are rHTVs. We have also run an additional simulation under a scenario that combines scenarios 2 and 3, and obtained similar results as shown in Supplementary Figure 5.

Missing data is an unavoidable issue in family data. An advantage of family-based data is that for a person with missing genotype data, the set of compatible haplotype configurations can greatly reduce the number of possible haplotypes or even be completely deduced based on genotypes of other members. Thus, we can obtain a list of compatible haplotype configurations of all members (including those with missing genotypes). However,

depending on the extent and nature of the missing data and the size of the pedigree, the number of such configurations can be in the order of tens of thousands or more, as we found in the analysis of the GAW18 and the FHS data, which contain some extremely large pedigrees. The computational intensity of famQBL increases tremendously with such large number of configurations and thus we had to extract sub-pedigrees. Although this leads to a loss of information, the use of a software such as PedStr ensures that the extracted sub-pedigrees are optimal and objective according to certain criteria. As indicated from the GAW data analysis, famQBL can handle the number of haplotype configurations in the order of several thousands for a moderate number of families. However, if there are hundreds of families, each with such a large number of compatible configurations, then the computational intensity can become prohibitive.

The analysis of the FHS data using famQBL revealed an association of rHTV with DBP. This is consistent with an earlier report of rare haplotype association with the binary trait hypertension [15]. Although FBAT found association with common haplotypes in this region, those effects are weaker and none of them were found to be significant by famQBL. However, as results from our simulations under the RC association model show, FBAT tends to have larger type I error rates, especially when the model assumptions are violated.

Notwithstanding the advantages of famQBL, it has a number of limitations. First, we need to find compatible haplotype configurations, which may lead to loss of information dictated by the software selected; a case in point is the HAPLORE software which cannot include genotypes available only for one, and not both parents. Second, as famQBL is an MCMC-based procedure and involves updating of haplotype pairs of all members in a family jointly, it is computationally intensive. In particular, for simulation scenarios 1 and 2 with no missing data, it takes roughly 14.5, 15, and 15.25 minutes to run a sample of 300 families for settings 1, 2, and 3, respectively. For scenario 3, which has missing data and substantially longer MCMC run, the corresponding times are 40.5, 41.2, and 42.5 minutes. For the FHS data analysis with 5 SNPs, the computing time is 154.3 minutes due to presence of much larger families and an excessive amount of missing data as described earlier. These computing times are for a 3.5 GHz Xeon processor under Linux operating system with 32.74 GB RAM. Due to the computational intensity, famQBL is not intended for a genome-wide search. Instead, it will be most useful as a follow-up tool to zoom into regions found to be promising in an initial screening of the genome for association or linkage peaks suggested in family studies. Nonetheless, it will certainly be worth exploring ways to reduce the computational burden of famQBL to enhance its practical utility, such as parallel computing, which seems to be naturally suited for updating the haplotypes since each family can be processed individually and separately.

Several avenues for future work appear promising. This includes extension of famQBL to handle case-control or qualitative response, environmental covariates, and their interactions with haplotypes. Further, a unified approach to analyze family and case-control data together will be profitable from a practical perspective as that will allow the uncovering of more modest effects through the use of larger samples, similar to the joint analysis scheme for triad families and case-control data [41]. Finally, a comprehensive comparison of family-

based haplotype association approaches, similar to a comparison of population-based haplotype approaches [17], will provide useful practical guidance for practitioners.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by the grant R03CA171011 from the National Cancer Institute, NIH. The GAW16 FHS data were obtained through dbGaP accession number phs000128.v3.p3 (https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000128.v3.p3). We thank the anonymous reviewers for providing constructive comments, which led to an improved version of the paper.

References

- [1]. Maher B Personal genomes: the case of the missing heritability. *Nature* 2008;456:18–21. [PubMed: 18987709]
- [2]. Manolio TA, Collins FS, Cox NJ et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747–53. [PubMed: 19812666]
- [3]. Dering C, Hemmelmann C, Pugh E et al. Statistical Analysis of Rare Sequence Variants: An Overview of Collapsing Methods. *Genet Epidemiol* 2011;35:(Suppl 1):S12–S17. [PubMed: 22128052]
- [4]. Basu S, Pan W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 2011;35:606–19. [PubMed: 21769936]
- [5]. Goldstein DB, Andrew A, Jonathan K et al. Sequencing studies in human genetics: design and interpretation. *Nat Rev Genet* 2013;14:460–70. [PubMed: 23752795]
- [6]. Schaid DJ, Rowland CM, Tines DE et al. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am J Hum Genet* 2002;70:425–34. [PubMed: 11791212]
- [7]. Schaid DJ. Genetic epidemiology and haplotypes. *Genet Epidemiol* 2004;27:317–20. [PubMed: 15543637]
- [8]. Morris RW, Kaplan NL. On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genet Epidemiol* 2002;23:221–33. [PubMed: 12384975]
- [9]. Li Y, Byrnes AE, Li M. To Identify Associations with Rare Variants, Just WHaIT: Weighted Haplotype and Imputation-Based Tests. *Am J Hum Genet* 2010; 87:728–35. [PubMed: 21055717]
- [10]. Lin WY, Yi N, Zhi D et al. Haplotype-based methods for detecting uncommon causal variants with common SNPs. *Genet Epidemiol* 2012;36:572–582. [PubMed: 22706849]
- [11]. Guo W, Lin S. Generalized Linear Modeling with Regularization for Detecting Common Disease Rare Haplotype Association. *Genet Epidemiol* 2009;33:308–16. [PubMed: 19025789]
- [12]. Biswas S, Lin S. Logistic Bayesian LASSO for identifying association with rare haplotypes and application to age-related macular degeneration. *Biometrics* 2012;68:587–597. [PubMed: 21955118]
- [13]. Biswas S, Xia S, Lin S. Detecting rare haplotype-environment interaction with logistic Bayesian LASSO. *Genet Epidemiol* 2014;38:31–41. [PubMed: 24272913]
- [14]. Zhang Y, Lin S, Biswas S. Detecting rare haplotype-environment interaction under uncertainty of gene-environment independence assumption. *Biometrics* 2017;73:344–355. [PubMed: 27478935]
- [15]. Wang M, Lin S. Detecting associations of rare variants with common diseases: collapsing or haplotyping? *Brief in Bioinform* 2015;16:759–768.
- [16]. Datta AS, Zhang Y, Zhang L, Biswas S. Association of rare haplotypes on ULK4 and MAP4 genes with hypertension. *BMC Proceed* 2016;10(Suppl 7):44.
- [17]. Datta AS, Biswas S. Comparison of haplotype-based statistical tests for disease association with rare and common variants. *Brief Bioinform* 2016;17:657–671. [PubMed: 26338417]

- [18]. Rabinowitz D, Laird N. A unified approach to adjusting association tests for population with arbitrary pedigree structure and arbitrary missing marker information. *Hum Hered* 2000;504:227–233.
- [19]. Horvath S, Xu X, Lake SL et al. Family-Based Tests for Associating Haplotypes with General Phenotype Data: Application to Asthma Genetics. *Genet Epidemiol* 2004;26:61–69. [PubMed: 14691957]
- [20]. Zhu X, Feng T, Li Y et al. Detecting Rare Variants for Complex Traits Using Family and Unrelated Data. *Genet Epidemiol* 2010;34:171–187. [PubMed: 19847924]
- [21]. Li X, Li J. Comparison of haplotyping methods using families and unrelated individuals on simulated rheumatoid arthritis data. *BMC Proceed* 2007;1(Suppl I):S55.
- [22]. Bowden DW. Will family studies return to prominence in human genetics and genomics? Rare variants and linkage analysis of complex traits. *Gene and Genomics* 2011;33:1–8.
- [23]. Ionita-Laza I, Cho M, Laird N Statistical challenges in sequence-based association studies with population-and family-based designs. *Stat in Biosciences* 2013;5:54–70.
- [24]. De G, Yip W-K, Ionita-Laza I. et al. Rare Variant Analysis for Family-Based Design. *PLoS ONE* 2013; 8:e48495. [PubMed: 23341868]
- [25]. Wang M and Lin S. FamLBL: detecting rare haplotype disease association based on common SNPs using case-parent triads. *Bioinformatics* 2014;30:2611–2618. [PubMed: 24849576]
- [26]. Guo W and Shugart YY. The power comparison of the haplotype-based collapsing tests and the variant-based collapsing tests for detecting rare variants in pedigrees. *BMC Genomics* 2014;15:632. [PubMed: 25070353]
- [27]. Martin ER, Monks SA, Warren LL, and Kaplan NL. A test for linkage and association in general pedigrees: the pedigree disequilibrium test. *Am J Hum Genet* 2000;67:146–154. [PubMed: 10825280]
- [28]. Lin S Monte Carlo Bayesian methods for quantitative traits. *Comput Statist & Data Analysis* 1999;31:89–108.
- [29]. Almasy L, Dyer TD, Peralta JM et al. Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. *BMC Proceed* 2014; 8(Suppl 1):S2.
- [30]. Cupples LA, Heard-Costa N, Lee M et al. Genetics Analysis Workshop 16 Problem 2: the Framingham Heart Study data. *BMC Proceed* 2009;3(Suppl 7):S3.
- [31]. Weir BS. *Genetic Data Analysis II* Sunderland, MA, USA, Sinauer Associates Inc, 1996.
- [32]. Gelman A Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis* 2006; 1(3): 515–533.
- [33]. Zhang K, Sun F, Zhao H. HAPLORE: a program for haplotype reconstruction in general pedigrees without recombination. *Bioinformatics* 2005;21:90–103. [PubMed: 15231536]
- [34]. Fuchsberger C, Flannick J, Teslovich TM et al. The genetic architecture of type 2 diabetes. *Nature* 2016;536:41–47. [PubMed: 27398621]
- [35]. Yang J and Lin S. Likelihood Approach for Detecting Imprinting and In Utero Maternal Effects Using General Pedigrees from Prospective Family-Based Association Studies. *Biometrics* 2012;68:477–485. [PubMed: 22008205]
- [36]. Leal SM, Kai Yan K, Muller-Myhsok B, SimPed: A Simulation Program to Generate Haplotype and Genotype Data for Pedigree Structures. *Hum Hered* 2005;60:119–122. [PubMed: 16224189]
- [37]. Gelman A, Carlin JB, Stern HS, Rubin DB. *Bayesian Data Analysis* Boca Raton, FL, USA, Chapman and Hall/CRC, 2003.
- [38]. Kirichenko AV, Belonogova NM, Aulchenko YS, et al. PedStr software for cutting large pedigrees for haplotyping, IBD computation and multipoint linkage analysis. *Ann Hum Genet* 2009;73(Pt 5):527–531. [PubMed: 19604226]
- [39]. Biswas S, Papachristou C. Evaluation of logistic Bayesian LASSO for identifying association with rare haplotypes. *BMC Proceed* 2014;8:S54.
- [40]. Han M, Hu YQ, and Lin S. Joint detection of association, imprinting and maternal effects using all children and their parents. *Eur J Hum Genet* 2013;21:1449–1456. [PubMed: 23531864]

- [41]. Zhou X, Wang M, Stewart WCL, and Lin S. Logistic Bayesian LASSO for detecting association combining family and case-control data BMC Proceed, 2018; 12 (Suppl 9):54.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Figure 1:

Three pedigree types used in the Simulation Study 1. A sample consists of 100 pedigrees of each type. Figure adapted from [35].

Figure 2:

Type I error rates and powers of famQBL and FBAT under the RR association model and scenarios 1 (top), 2 (middle), and 3 (bottom). In each plot, type I error rates for non-causal haplotypes are plotted on the left and power for causal haplotypes are plotted on the right. The three settings are described in Table 1. The three scenarios are described in Table 2.

Figure 3:

Type I error rates and powers of famQBL and FBAT under the RC association model and scenarios 1 (top), 2 (middle), and 3 (bottom). In each plot, type I error rates for non-causal haplotypes are plotted on the left and power for causal haplotypes are plotted on the right. The three settings are described in Table 1. The three scenarios are described in Table 2.

Figure 4:
Results for GAW18 simulated data

Table 1.

Three haplotype settings and two types of association (RR and RC) for Simulation Study 1.

Setting 1				Setting 2				Setting 3			
Hap	Freq	$\beta(RR)$	$\beta(RC)$	Hap	Freq	$\beta(RR)$	$\beta(RC)$	Hap	Freq	$\beta(RR)$	$\beta(RC)$
10011	0.420	0	0	01010	0.060	0	0	00111	0.070	0	0
01100	0.300	0	0	01100	0.250	0	0	01000	0.020	0	0
11100	0.155	0	0	10000	0.080	0	0.8	01011	0.050	0	0
11111	0.110	0	0.8	10100	0.005	4	4	01101	0.060	0	0
11011	0.010	2	0	11011	0.010	2	0	01110	0.140	0	0
10100	0.005	4	4	11100	0.090	0	0	10010	0.080	0	0.8
				11101	0.085	0	0	10100	0.005	4	4
				11111	0.100	0	0	11011	0.010	2	0
				10011	0.320	0	0	11101	0.090	0	0
								11110	0.130	0	0
								11111	0.100	0	0
								10001	0.245	0	0

Table 2.

Three simulation scenarios for Simulation Study 1. A simulation scenario combined with a haplotype setting and an association type (both described in Table 1) gives one simulation set-up. MVN: Multivariate normal distribution with parameters of mean vector and covariance matrix; MVT: Multivariate t distribution with parameters of degrees of freedom and covariance matrix.

	Scenario 1	Scenario 2	Scenario 3
Environmental effect	$N(0, 20)$	$t^*_{2.105}$	$N(0, 20)$
Polygenic effect for j th family	$MVN(\mathbf{0}, 5\Sigma_j)$	$MVT(n_j - 1, 5\Sigma_j)$	$MVN(\mathbf{0}, 5\Sigma_j)$
Missing data	No	No	Yes

* $t_{2.105}$ distribution has variance of 20, where 2.105 is the degrees of freedom.

Table 3.

Results for Region 1 surrounding SNP at 47,956,424 bp. The last four rows, two each for DBP and SBP, show the proportion of replicates with 95% CS excluding 0 for famQBL and with p-value < 0.05 for FBAT. There are 7 possible haplotypes (Hap) with the minor allele of each SNP denoted by 1. The minor alleles in bold typeface are functional variants. The haplotypes containing these variants (in bold) correspond to power while the rest correspond to type I error rate. The most frequent haplotype (Hap1) is the baseline. Freq represents haplotype frequency.

SNP(bp)	MAF	Hap1	Hap2	Hap3	Hap4	Hap5	Hap6	Hap7
47,952,843	0.018	0	0	0	0	0	0	1
47,953,405	0.370	0	1	1	1	1	1	1
47,953,733	0.320	0	0	0	0	1	1	0
47,956,424	0.359	0	0	1	1	1	1	1
47,956,506	0.317	0	0	0	0	1	1	0
47,957,996	0.021	0	0	0	0	0	1	0
47,958,037	0.317	0	0	0	0	1 *	1 *	0
47,959,770	0.367	0	1	1	1	1	1	1
47,959,977	0.011	0	0	0	1	0	0	0
Hap Freq		0.640	0.006	0.011	0.011	0.280	0.034	0.017
famQBL (DBP)			0.0	0.0	0.0	0.0	64.0	0.0
FBAT (DBP)			4.0	8.5	4.0	1.0	35.0	6.0
famQBL (SBP)			0.0	0.0	0.0	0.0	38.5	0.0
FBAT (SBP)			6.0	4.0	8.5	2.0	17.0	6.5

* The % variability explained by this variant is almost negligible (<0.0001) as per the simulation model provided in GAW.

Table 4.

Results for Region 2 surrounding SNP at 48,040,283 bp. The last four rows, two each for DBP and SBP, show the proportion of replicates with 95% CS excluding 0 for famQBL and with p-value < 0.05 for FBAT. There are 12 possible haplotypes (Hap) with minor allele of each SNP denoted by 1. The minor alleles in bold typeface are functional variants. The haplotypes containing these variants (in bold) correspond to power while the rest correspond to type I error rate. The most frequent haplotype (Hap1) is the baseline. Freq represents haplotype frequency.

SNP(bp)	MAF	Hap1	Hap2	Hap3	Hap4	Hap5	Hap6	Hap7	Hap8	Hap9	Hap10	Hap11	Hap12
48,036,889	0.011	0	0	0	0	0	0	0	0	0	0	0	1
48,037,078	0.363	0	0	0	0	0	1	1	1	1	1	1	0
48,038,714	0.373	0	0	0	0	1	1	1	1	1	1	1	1
48,039,908	0.246	0	0	1	1	1	0	0	0	0	0	0	0
48,040,283	0.025	0	0	0	0	0	0	0	0	0	0	1	0
48,040,284	0.021	0	0	0	1	0	0	0	0	0	0	0	0
48,041,471	0.025	0	0	0	0	0	0	0	1	1	1	0	0
48,042,192	0.018	0	0	0	0	0	0	0	0	1	1	0	0
48,043,058	0.331	0	1	0	0	0	0	1	0	0	1	1	0
Hap Freq		0.397	0.006	0.204	0.017	0.006	0.011	0.291	0.011	0.011	0.006	0.034	0.006
famQBL (DBP)		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	60.0	0.0
FBAT (DBP)		3.0	1.5	3.0	3.0	23.0	4.5	1.5	3.5	6.0	NA	36.5	5.0
famQBL (SBP)		0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	32.5	0.0
FBAT (SBP)		13.5	0.0	2.5	2.5	31.0	9.0	1.5	1.5	6.5	NA	16.0	4.5

NA as this particular haplotype was not reported by FBAT.

Table 5.

Results from FHS data analysis. The target SNP rs2229188 is at the fifth position. For famQBL, BF, posterior mean of β ($\hat{\beta}$), and 95% CS for β are reported. For FBAT, p-values are reported in the last column. The numbers in bold represent significant results.

Haplotype	Haplotype Freq	famQBL (BF)	famQBL ($\hat{\beta}$)	famQBL (CS)	FBAT
011011	0.2669	0.003	-0.034	(-0.102, 0.034)	0.217
100001	0.1587	0.022	-0.062	(-0.142, 0.017)	0.556
011001	0.1397	0.008	-0.040	(-0.119, 0.037)	0.026
000011	0.0753	0.034	0.063	(-0.042, 0.170)	0.004
000001	0.0411	0.014	0.022	(-0.097, 0.143)	0.090
111011	0.0037	0.227	-0.155	(-0.564, 0.219)	0.186
111001	0.0034	0.240	0.020	(-0.559, 0.643)	0.603
001011	0.0027	0.230	0.140	(-0.294, 0.607)	0.250
100111	0.0017	17.525	-1.573	(-2.999, -0.301)	0.236
010001	0.0008	0.194	0.034	(-0.432, 0.516)	0.664
010011	0.0006	0.407	-0.218	(-1.059, 0.476)	NA
001001	0.0004	0.548	2.018	(-1.803, 44.311)	0.000*
011101	0.00003	0.548	-0.057	(-2.391, 1.802)	0.498

*The number of informative families for this haplotype in FBAT analysis is 1 and hence the result is unreliable.