BMC Genetics

SOFTWARE Open Access

A variance component based multi-marker association test using family and unrelated data

Xuefeng Wang[1], Nathan J Morris[2], Xiaofeng Zhu[2*] and Robert C Elston[2]

Abstract

Background: Incorporating family data in genetic association studies has become increasingly appreciated, especially for its potential value in testing rare variants. We introduce here a variance-component based association test that can test multiple common or rare variants jointly using both family and unrelated samples.

Results: The proposed approach implemented in our R package aggregates or collapses the information across a region based on genetic similarity instead of genotype scores, which avoids the power loss when the effects are in different directions or have different association strengths. The method is also able to effectively leverage the LD information in a region and it can produce a test statistic with an adaptively estimated number of degrees of freedom. Our method can readily allow for the adjustment of non-genetic contributions to the familial similarity, as well as multiple covariates.

Conclusions: We demonstrate through simulations that the proposed method achieves good performance in terms of Type I error control and statistical power. The method is implemented in the R package “fassoc”, which provides a useful tool for data analysis and exploration.

Keywords: Association studies, Family data, Score test, Multi-marker test

Background

With the availability of cost-effective next generation sequencing platforms, one hot topic in the field is the analysis of low frequency and rare variants, which are believed to play an important role in the etiology of common complex diseases and may explain a portion of the missing heritability [1,2]. However, the sample sizes investigated in most studies are not large enough to ensure sufficient power for detecting rare variants with small or moderate effect sizes using single-marker tests [3]. Combining both family and unrelated data can improve statistical power over separate analysis of family data and unrelated data [4]. Current methods for testing rare variants are mainly based on aggregation or group tests that first pool together all variants with low minor allele frequencies in a region of interest and then test the association between phenotypes and the combined super-locus. Two of the earliest collapsing methods proposed are the combined multivariate and collapsing (CMC) test [5] and the weighted-sum method [6]. A number of variations of these methods have also been quickly developed [3,7-9]. Despite these developments, challenges remain to identify rare risk variants under different scenarios and assumptions. Because the aggregation test needs to assume homogeneity in the magnitude and direction of the individual effect sizes, it may experience massive loss in power when both protective and risk variants are present in the tested region, or when inappropriate weights/priors (or threshold of allele frequency) are used on rare variants. It is therefore timely to seek more powerful and reliable methods and designs. Motivated by recent works of Feng et al. [10] and Zhu et al. [11] who found that using sib pair data can increase power over using only unrelated samples, here we further explore the performance of methods with family information in searching for rare variants underlying complex traits.

In this work, we present an R package that implements a variance-component (VC) based association test that can test multiple common or rare variants jointly using both family and unrelated samples. The VC or linear mixed

* Correspondence: xzhu1@darwin.epbi.cwru.edu
[2]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA
Full list of author information is available at the end of the article

model (LMM) based approach aggregates or collapses the information across a region based on genetic similarity instead of genotype effects, which avoids the power loss when the effects are in different directions. A comparison study of binary traits [12] has also shown that the similarity-based test can be more powerful than the collapsing test when the rare variants have different association strengths. We propose to include an additional random effect in the mixed model in order to model polygenic effects and familial correlations.

Our method can readily allow adjusting for a non-genetic contribution to the familial similarity (shared environmental effects), as well as multiple covariates such as principal components of population structure. We compare the performance of the proposed method across a range of simulation scenarios with a fixed-effect or sum test based on Feasible Generalized Least Squares (FGLS). We also investigate the factors that influence power for testing rare variants. In this paper, we also show the connection between our method and kernel machine based methods [13,14], which may provide more flexibility in extending the proposed model. Although the simulations in this paper focus on rare variants analysis, our package can be readily applied to common variant association tests without any change.

Implementation

Assume there are q subjects in the sample studied, including both family and unrelated individuals; and suppose for all individuals there is one gene or genetic region genotyped or sequenced that contains n variant sites or SNPs. For the ith individual, y_i denotes the observed quantitative trait value; $X_i = (x_{i,1}, x_{i,2} \dots x_{i,m})'$ denotes an $m \times 1$ vector of covariates (which might include sex, age, environmental factors, and principal components to allow for population stratification); $S_i = (s_{i,1}, s_{i,2}, \dots s_{i,n})'$ denotes an $n \times 1$ genotype score vector for the n SNPs or variants in the region, coded 0, 1, or 2 (i.e., additive coding), reflecting the number of copies of the minor allele; and $Z_i = (z_{i,1}, z_{i,2} \dots z_{i,n})'$ denotes a standardized genotype vector with the ij-th element $z_{i,j} = \left(s_{ij} - 2f_j\right) / \sqrt{2f_i\left(1 - f_j\right)}$, where f_j is the minor allele frequency of the jth SNP or variant site.

Linear mixed model and score test

The setup of our model is similar to the linear mixed model recently proposed to estimate the genetic variance explained by genome-wide SNPs [15,16], in which using all common SNPs was claimed to be able to uncover a large portion of missing heritability. That model required all subjects to be unrelated and assumed the similarity among individuals' phenotype values is completely due to the similarity of their genetic components. The mixed model is written in matrix form as $\mathbf{y} = \mathbf{X\beta} + \mathbf{W\mu} + \boldsymbol{\varepsilon}$ with $\mathrm{var}(y) = WW'\sigma_u^2 + \mathbf{I}\sigma_e^2$, where y is a phenotype vector (assumed to be centered), X is a covariate matrix whose ith row is X_i'; β is a vector of coefficients (fixed effects) for covariates X, μ is a vector of causal variant effects with $\boldsymbol{\mu} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_\mu^2\right)$, W is a standardized genotype matrix, I is an identity matrix and ε is a random error vector with $\boldsymbol{\varepsilon} \sim N\left(\mathbf{0}, \mathbf{I}\sigma_\varepsilon^2\right)$. In the real case when the position and number of causal variants are unknown, a working model can be represented as $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$, where δ is a vector representing random effects of all SNPs, with $\boldsymbol{\delta} \sim N\left(\mathbf{0}, \mathbf{A}\sigma_\delta^2\right)$ and thus $\mathrm{var}(\mathbf{y}) = \mathbf{A}\sigma_\delta^2 + \mathbf{I}\sigma_\varepsilon^2$. A can be interpreted as the genetic relationship matrix (GRM) among individuals and its jk-th element is $\mathbf{A}_{jk} = \sum_{i=1}^{N} \frac{\left(s_{ij} - 2f_i\right)\left(s_{ik} - 2f_i\right)}{2f_i(1 - f_i)} / N$, where N is the total number of genome-wide SNPs. The variance components can be estimated via the restricted maximum likelihood (REML) method [16].

To estimate and test the variance expressed by a gene or a genomic region using both family and unrelated data, intuitively one can extend the above model by

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z\gamma} + \boldsymbol{\delta} + \varepsilon \ = \mathbf{X\beta} + \mathbf{g} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}, \tag{1}$$

where γ is a vector of the random effect of SNPs in the studied region distrtibuted $\sim N\left(\mathbf{0}, \mathbf{I}\sigma_\gamma^2\right)$, Z is the standardized genotype matrix, and δ is a vector of random effect representing the polygenic genic effects over the genome. Under this model, the marginal phenotypic variance V_y can be partitioned into components attributable to the SNPs in the studied region, polygenic and residual variances:

$$\begin{aligned} \mathbf{V_y} &= \mathbf{V}_g + \mathbf{V}_\delta + \mathbf{V}_\varepsilon = \mathbf{ZZ}'\sigma_\gamma^2 + \mathbf{A}\sigma_\delta^2 + \mathbf{I}\,\sigma_\varepsilon^2 \\ &= \mathbf{S}\sigma_g^2 + \mathbf{A}\sigma_\delta^2 + \mathbf{I}\,\sigma_\varepsilon^2, \end{aligned} \tag{2}$$

where $\mathbf{S} = \mathbf{ZZ}'/n$, and σ_g^2 represents the variance explained by the SNPs in the region, i.e., $\sigma_g^2 = n\sigma_\gamma^2$. A and S can thus be interpreted as two genetic similarity matrices. In this model, if two individuals are from different families (unrelated), their corresponding entry in A is calculated genomewide in the same way as above, but excluding the SNPs in the region we are testing. For individuals in the same family, or when the genome-wide SNPs are not available, the corresponding entries in A can be approximately computed by twice their kinship coefficients - which depends only on the relatedness between individuals - in which case $\mathbf{V_y} = \mathbf{S}\sigma_g^2 + 2\mathbf{\Phi}\sigma_\delta^2 + \mathbf{I}\sigma_\varepsilon^2$, where Φ denotes the q × q kinship matrix. To account for the common environmental factors shared by family members, we can include a common environmental factor in the model. Our mixed linear model now becomes $\mathbf{y} = \mathbf{X\beta} + \mathbf{g} + \boldsymbol{\delta} + \boldsymbol{\alpha} + \boldsymbol{\varepsilon}$ with $\mathbf{V_y} = \mathbf{S}\sigma_g^2 + \mathbf{A}\sigma_\delta^2 + \mathbf{C}\sigma_\alpha^2 + \mathbf{I}\,\sigma_\varepsilon^2$, where a is the effect due to the shared common environment factors with $\boldsymbol{\alpha}\tilde{}N\left(\mathbf{0}, \mathbf{C}\sigma_\alpha^2\right)$ and C is a matrix with the the *jk*-th element being 1 if the *j*-th and *k*-th individuals belong to the same family and 0 otherwise. Note that, by adding a variance component common to siblings, it is also easy to allow for the fact that siblings resemble each other more than do parents and their offspring, whether due to dominant effects or common environment.

This model can be readily applied to haplotype-based analysis with the design matrix for genotype scores Z replaced by a haplotype matrix H, where a vector H_i records the *i*-th individual's haplotype pair via a given scoring rule [17]. Hence, Model (1) becomes $\mathbf{y} = \mathbf{X\beta} + \mathbf{H}\boldsymbol{\gamma}_h + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$ with $\mathbf{V_y} = \mathbf{S}_h\sigma_h^2 + \mathbf{A}\sigma_\delta^2 + \mathbf{I}\,\sigma_\varepsilon^2$, where γ_h represents the random effect of haplotypes; S_h is a matrix of pair-wise similarity scores between the haplotype pairs of two individuals, with the *ij*-th element equal to $\sum_{h,k} H_{i,h}H_{j,k} \times s(h,k)$ [18], where $s(h,\ k)$ is a similarity matrix measuring the similarity between haplotypes *h* and *k*. If we set s(*h*, *k*) as the proportion of matching alleles between two haplotypes, the *ij*-th element of S_h will be equivalent to the average allelic sharing across multiple markers between two individuals and thus phase information is not required.

Our primary interest lies in detecting whether there is an effect of a genomic region on the phenotype, which is assessed by testing the null hypothesis $H_0 : \sigma_g^2 = 0$. In the following, we construct a fast score test based on the MLE and REML framework as an extension of that proposed by Tzeng and Zhang [17]. For the sake of demonstration, we first assume there is no shared environmental effect within families. Assuming a normally distributed quantitative trait, the log-likelihood function and its REML version for the variance component model are written as

$$\ell\left(\sigma_g^2, \sigma_\delta^2, \sigma_e^2; \mathbf{y}\right) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log\left|\mathbf{X^TV^{-1}X}\right| \\ -\frac{1}{2}(\mathbf{y} - \mathbf{X\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X\beta})$$

$$\ell_{\text{REML}}\left(\sigma_g^2, \sigma_\delta^2, \sigma_e^2; \mathbf{y}\right) = -\frac{1}{2}\log|\mathbf{V}| - \frac{1}{2}\log\left|\mathbf{X^TV^{-1}X}\right| \\ -\frac{1}{2}\mathbf{y}^T\mathbf{P}^{-1}\mathbf{y},$$

where $\mathbf{V} = \mathbf{S}\sigma_g^2 + \mathbf{A}\sigma_\delta^2 + \mathbf{I}\,\sigma_\varepsilon^2$, and $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$ is the projection matrix under the linear mixed model (1).

It will be convenient to denote the parameter of interest σ_g^2 by τ, and the nuisance parameters $\left(\boldsymbol{\beta}, \sigma_\delta^2, \sigma_e^2\right)$ by η. Under the null hypothesis, the score statistic with respect to τ is given by

$$U_\tau(\hat{\eta}) = \frac{\partial\ell(\tau, \eta; y)}{\partial\tau}\bigg|_{\tau=0,\eta=\hat{\eta}} \\ = -\frac{1}{2}\mathrm{tr}\left(\mathbf{V}_0^{-1}\mathbf{S}\right) + \frac{1}{2}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\mathbf{V}_0^{-1}\mathbf{S}\mathbf{V}_0^{-1}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)$$

where $\mathbf{V}_0 = \mathbf{A}\hat{\sigma}_\delta^{\ 2} + \mathbf{I}\,\hat{\sigma}_\varepsilon^{\ 2}$, and $\hat{\eta} = \left(\hat{\boldsymbol{\beta}}, \hat{\sigma}_p^{\ 2}, \hat{\sigma}_{\mathrm{e}}^{\ 2}\right)$ is the maximum likelihood estimate of η under the null linear mixed model $\mathbf{y} = \mathbf{X\beta} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$. These estimates can be obtained using the regular statistical software that implement mixed-model functionality, or even more easily in some genetic analysis packages that can directly read in a kinship matrix, such as EMMA [19] (http://mouse.cs.ucla.edu/emma/) and GenABEL (http://www.genabel.org/).

However, the asymptotic distribution of the above score statistic is not a typical standard normal distribution (neither does the corresponding LRT statistic converge to a mixture of χ_0^2 and χ_1^2). This is because, in contrast to IBD, the genotype-based similarity matrix $\mathbf{S} = \mathbf{ZZ}'/n$ does not have a block diagonal structure, and thus the statistic cannot be written in a form of a sum of independent variables that meets the asymptotical conditions indicated in Lin [20]. Instead, we can construct the test on the basis of the second term of $U_\tau(\hat{\eta})$, following the approach proposed by Zhang and Lin [21]. Letting $\mathbf{M} = \frac{1}{2}\mathbf{V}_0^{-1}\mathbf{S}\mathbf{V}_0^{-1}$ and $\tilde{}\mathbf{y}\tilde{} = \mathbf{V}_0^{-1/2}\left(\mathbf{y}–\mathbf{X}\hat{\boldsymbol{\beta}}\right)$, the new statistic becomes

$$T_\tau = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)^T\mathbf{M}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) = \widetilde{\mathbf{y}}^T\mathbf{V}_0{}^{1/2}\mathbf{M}\mathbf{V}_0{}^{1/2}\widetilde{\mathbf{y}},$$

Because asymptotically $\mathbf{y}\tilde{}\sim N(\mathbf{0}, \mathbf{I})$, T_τ follows a weighted sum of chi-square variables: $T_\tau \sim \sum_{i=1}^m \lambda_i\chi_{1i}^2$, where χ_{1i}^2 are independent chi-square random variables with one degree of freedom, and the weights λ_i are the *i*-th ordered nonzero eigenvalues of $\mathbf{V}_0^{1/2}\mathbf{M}\mathbf{V}_0^{1/2}$. A good approximation may be obtained using only *r* (<<*q*) dominant eigenvalues as λ usually decays rapidly toward zero.

Significance of a test can be evaluated empirically through simulating a large set of sums of chi-squared random variables, where the *p*-value is obtained by calculating the proportion of the generated random variables that are greater than the observed statistic. However, this is considerably slower than computing the eigenvalues of $\mathbf{V}_0^{1/2}\mathbf{M}\mathbf{V}_0^{1/2}$ when the sample size is large. Furthermore, to ensure reliable results for a large effect size or small α level, one needs to run a huge number of simulations. For instance, when α is set at 1×10^{-5}, at least 10^7 simulations are needed for each test. This becomes computationally infeasible for a genome-wide scan. Here we consider Satterthwaite's procedure to approximate the null distribution of T_τ by a scaled chi-square distribution $k\chi^2_\upsilon$ or a gamma distribution Gamma (a, b). The two parameters in the approximate distribution are calculated by matching the first and second moments (mean and variance) with those of the score statistic. Taking a Gamma distribution as an example, we attempt to obtain $ab = \mu_T$ and $ab^2 = \nu_T \Leftrightarrow a = \mu^2_T/\nu_T$ and $b = \nu_T/\mu_T$. Due to its quadratic form, it is easy to obtain the mean and variance of $T\tau$:

$$\begin{aligned}\mu_T &= E\left(\tilde{\mathbf{y}}^T \mathbf{V}_0{}^{1/2}\mathbf{M}\mathbf{V}_0{}^{1/2}\tilde{\mathbf{y}}\right)\\ &= \mathrm{tr}\left(\mathbf{V}_0{}^{1/2}\mathbf{M}\mathbf{V}_0{}^{1/2}\right) = \frac{1}{2}\mathrm{tr}\left(\mathbf{V}_0{}^{-1}\mathbf{S}\right)\\ \nu_T &= \mathrm{var}\left(\tilde{\mathbf{y}}^T \mathbf{V}_0{}^{1/2}\mathbf{M}\mathbf{V}_0{}^{1/2}\tilde{\mathbf{y}}\right)\\ &= 2\mathrm{tr}\left[\left(\mathbf{V}_0{}^{1/2}\mathbf{M}\mathbf{V}_0{}^{1/2}\right)^2\right] = \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{V}_0{}^{-1}\mathbf{S}\right)^2\right]\end{aligned}$$

To account for the fact that the nuisance parameters η are estimated and replaced by their MLEs $\hat{\eta}$, ν_T may be replaced by the partial information $I_\tau = I_{\tau\tau} - I_{\tau\eta}I^{-1}_{\eta\eta}I_{\eta\tau}$ (to subtract the loss of information in the data due to η being unknown), where $I_{\tau\tau} = \frac{1}{2}\mathrm{tr}\left[\left(\mathbf{V}_0{}^{-1}\mathbf{S}\right)^2\right]$, $I_{\tau\eta} = \frac{1}{2}\mathrm{tr}\left[\mathbf{V}_0{}^{-1}\mathbf{S}\mathbf{V}_0{}^{-1}\frac{\partial \mathbf{V}}{\partial \eta}\right]$, $I_{\eta\eta} = \frac{1}{2}\mathrm{tr}\left[\mathbf{V}_0{}^{-1}\frac{\partial \mathbf{V}}{\partial \eta}\mathbf{V}_0{}^{-1}\frac{\partial \mathbf{V}}{\partial \eta}\right]$ and $I_{\eta\tau} = I^T_{\tau\eta}$. When the estimation and score test is based on the REML, the above formulas remain the same but with $\mathbf{V}_0^{-1}$ replaced by the projection matrix $\mathbf{P}_0 = \mathbf{V}_0^{-1} - \mathbf{V}_0^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}_0^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}_0^{-1}$.

Satterthwaite's procedure is fairly fast but may not have desirable performance in the extreme tails of the distribution. An alternative procedure would be to fit a distribution for which the first three moments are estimated, rather than only the first two. Possibilities would be to assume the distribution is a multiple of a non-central chi-square distribution, estimating the multiple and the two parameters of the non-central chi-square distribution from the empirical first three moments; alternatively, one could fit a distribution that is a multiple of a power of a chi-square distribution, estimating the multiple, the power and the *d.f.* from the first three moments. A similar strategy of utilizing higher moments/cumulants has been proposed by Liu et al. [22], in which the parameters of the approximate distribution are determined in such a way that skewness is matched while the difference in kurtosis is minimized. Other possible methods include the Davies method [23] (based on numerical inversion of the characteristic function), Farebrother's [24] and Imhof's methods [25,26]. These methods are available in an R package called "CompQuadForm".

The VC score approach described above has a special advantage of being easily extended to, and compatible with, the kernel machine regression that allows for more flexible modeling of genetic effects. Methods like least-square kernel machines (LSKM) and their variants have been successfully applied in multi-marker association tests with both common and rare variants [13,14,27]. Under the framework of LSKM, the outcome of an individual can be described by the following semiparametric regression model:

$$y_i = \mathbf{x}_i^T\boldsymbol{\beta} + h(\mathbf{s}_i) + \delta_i + \varepsilon_i\,,$$

where $h(.)$ is a nonparametric smoothing function that allows a flexible modeling of the influence of the genotype information s_i on the trait value. The function space that $h(.)$ lies in is fully determined by a positive semidefinite kernel function $K(.,.)$. A kernel function can implicitly map input data to a higher-dimension inner product space, and thus defines the complexity level of the relationship between the genotypes and the trait. Intuitively, a kernel function $K(\mathrm{s}_i,\mathrm{s}_j)$ can also be thought as a similarity measure between the genotypes of individuals i and j (in the genomic region tested). Three types of kernel used most often are the linear, quadratic and Gaussian kernels. Note that the linear kernel $K\left(\mathbf{s}_i, \mathbf{s}_j\right) = \mathbf{s}_i^T\mathbf{s}_j$ will be analogous to a covariance when s is centered. One can choose an appropriate non-linear kernel to accommodate interaction and nonlinear genetic effects.

Given the close relationship between the LSKM and GLMM framework, Liu et al. [28] found that it is much more convenient to test the null hypothesis $H_0 : h(\mathbf{z}) = 0$ based on the related linear mixed model. The corresponding model in our method is

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\delta} + \varepsilon\,,$$

where h is regarded as a random effect with mean zero and variance $\tau\mathbf{K}$, where $\mathbf{K}$ is an $n \times n$ matrix with the ij-th element equal to $K(\mathrm{s_i},\mathrm{s}_j)$. It can be shown that the best-linear unbiased estimators (BLUP) of h and β have the same form as those derived via LSKM estimation [27]. The equivalence implies that we can directly use the above likelihood functions and the test procedures that are constructed on Model (1), but with g replaced by h, and the similarity matrix $\mathbf{S}$ replaced by $\mathbf{K}$.

To accommodate rare variant SNPs, a weighted kernel function might be used so that similarity in rare variants will be emphasized. Assuming additive genotypic coding, a weighted IBS kernel can be written as $K\left(\mathbf{s}_i, \mathbf{s}_j\right) = \sum_{l=1}^p w_l\left(2 - \left|s_{i,l} - s_{j,l}\right|\right)$. One such weight is $w_l = 1/\sqrt{p_l}$, where p_l is the minor allele frequency of the l^{th} SNP or variant. A more flexible way is based on the density function of a beta distribution: $w_l = Beta(p_l; a, b)$ [14]. Note that, when $a = b = 0.5$, w_l will be equal to $1/p_l(1 -$

p_l), in which case the weighted IBS kernel with the original genotype scores will be analogous (but not exactly identical) to using standardized genotypes in model (1). Under this formulation, the VC score test can be viewed as a special case of the LSKM approach.

Simulations

We performed simulation studies to examine the type I error and power of the proposed score approach for detecting genetic variants under a range of scenarios, especially when rare variants are the cause of the phenotypic variation. We began by simulating 10,000 haplotypes of a 500 kb genomic region under the coalescent model using the software "cosi" (http://www.broadinstitute.org/~sfs/cosi/), with an effective population size of 10^4, mutation rate set at 1.5×10^{-8} per bp per generation, and the recombination rate varying across the region with a local window size of 100 kb. A total of 2883 variant locations were generated using this setting, of which 73 % had minor allele frequencies < 0.05. We randomly picked a region of 500 variants as our test region. In determining causal variants and risk haplotypes, we used a procedure similar to that described in Feng et al. [10]. Specifically, we assumed that only the variants with MAF < 2% can be causal variants, and considered a collapsing risk model in which the risk of one haplotype is determined by the presence of a minor allele at any risk location within the region. We then randomly drew causal variants from the pool of locations with MAF <2% until the accumulated frequency of risk haplotypes reached 10%. In each simulation, this procedure led to around 5%-8% of the variants being risk variants. In other words, we considered as risk haplotypes those that include at least one causal variant and assumed that their contributions to the phenotype are identical, i.e., the phenotype of an individual depends upon a genomic region only through the number of risk haplotypes she/he carries. The genotypes of unrelated individuals or those of founders in family data were simulated by randomly sampling with replacement two haplotypes from the 10,000 haplotypes. The haplotype data within one individual were then combined and converted into unphased genotype data. For illustration purposes, we only considered nuclear families for family data in our simulations, in which the number of children in each family was a random number drawn from a Poisson distribution with mean λ = 2. To simulate the genotypes of the second generation, we randomly drew one of the two haplotypes from each parent and then transmitted them to his/her offspring.

We determined the quantitative trait values based on a normal distribution. Specifically, we first calculated the causal genetic score (g) of an individual by g = zu, where u is the effect size and z_i is coded as 0, 1, or 2 indicating the number of risk haplotypes. Next we generated the overall residual variance by var(g)($1/h^2 - 1$), in which h^2 is the proportion of phenotypic variance explained by a genomic region, and var(g) is the theoretical variance of the genetic score calculated as var(g) = var(z)$u^2 - 2r(1 - r)u^2$, where r is the proportion of risk haplotypes (in 10,000 samples). The variances of the polygenic effect (p) and random error effect (e) were split from the overall residual variance to meet two conditions: var(g) + (p) = 0.4 and var(e) = 1-var (g)-var(p). For founders and unrelated individuals, we generated values of g, p, and e from normal distributions with means zero and variances var(g), var(p) and var(e), respectively. For children, p was generated by $p_c = \frac{1}{2}\left(p_m + p_f\right) + \frac{1}{\sqrt{2}}p_i$, where p_m and p_f are the values of the parents and $p_i \sim N(0, \text{var}\ (p))$. The phenotypic value of each individual was then calculated as $y = g + p + e$, and all y were centered before any analysis. For simplicity, here we did not simulate covariates or shared environment effects.

We designed various simulation scenarios by changing parameters such as h^2, sample sizes, and the proportion of risk haplotypes. Each set-up consisted of 200 independent replications (by updating each time not just phenotypes, but also genotypes). To compare with fixed-effect sum tests, each data set was also analyzed by the feasible generalized least squares regression model (FGLS). FGLS is very similar to generalized least squares except that it uses an estimated variance-covariance matrix (which can be obtained under the null model) [29]. We used the 'FGLS' function in the R package "MixABEL" (http://www.genabel.org/packages/MixABEL) to implement this analysis.

We have evaluated the type I error for the proposed method by generating 400,000 replicates under the H_0.

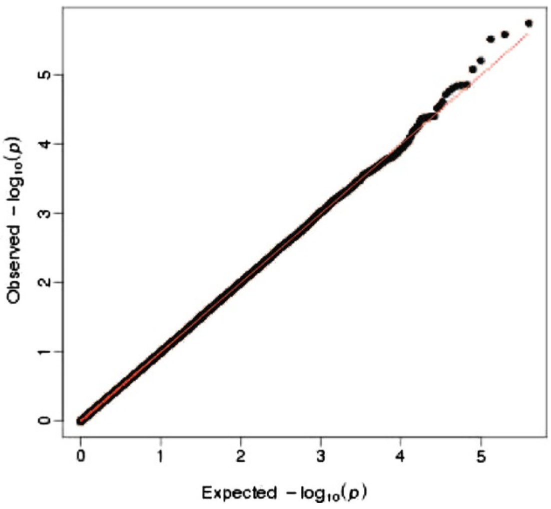

Figure 1 Quantile-quantile plot comparing empirical p-values (based on 400,000 simulations under the no assocation model) against those expected under the null.

Figure 1 shows a quantile-quantile plot of observed p values against those expected under the null.

Results and discussion

In our primary set of simulation for power comparison, 500 nuclear families and 2,000 unrelated individuals were generated, based on the simulation procedures described above, where the proportion of phenotypic variance explained by a region was set at (0, 0.01,. . ., 0.05). Each data set was analyzed by four different strategies: (1) the proposed VC-score test with all 500 variants; (2) the FGLS test using the genotype sum of 500 variants; (3) the VC-score test with only rare variants (with minor allele frequency (MAF) < 0.02 in the sample) included; (4) the FGLS test using the genotype sum of rare variants. Because we used standardized genotypic scores and true MAF thresholds for rare variants, results from method (4) should represent the best results that a weighted-sum aggregation test could possibly reach. The power was assessed at the 0.05 and 1×10^{-6} significance levels using 200 replications. When α was set at 0.05 and h^2 (heritability) set at 0, all analysis strategies maintain type I error rates around the nominal level. The power of the VC-score test is close to or higher than the FGLS method under all scenarios. The VC-score method also demonstrated great robustness to the number of noise markers. Results indicate that excluding common variants (all non-causal) results in noticeable power increase when using the FLGS method, but has nearly no effect on the VC method. We also tried the VC method using the genotype sum of rare variants only. Results are not presented here because they are exactly the same as those from method (4) in view of the equivalence of the two statistics when the genotype sum is used.

The simulation results indicate that, under the current simulation settings and sample sizes, the proposed method will have adequate power to detect a genomic region with h^2 around 0.01 in a candidate gene analysis, or a region with h^2 around 0.02 in a genome-wide scan. Table 1 summarizes the results from the simulations with increasing sample sizes, in which the power was evaluated at significance levels of .05, 1×10^{-5}, and 1×10^{-6}, respectively. Three different designs were considered. In design I we included an additional 1,000 unrelated individuals, while in design II we added another 250 families (approximately the same genotyping effort as 1000 unrelated individuals). Both designs gave apparent power increase compared to previous simulations (around 15% more when h^2 is below 0.03), but the increase in design I is slightly greater than that in design II. Our preliminary simulations show that the difference can be more significant when using a smaller base sample size. As generally accepted, an association analysis using related individuals is less informative than one using the same number of unrelated individuals, and is thus less powerful. In practice, families are not randomly sampled but often selected through probands or because of existing linkage evidence. We explored this effect in design III. Rather than going through the complex modeling of the ascertainment process, we created an enriched risk haplotype pool by directly removing 2,000 non-risk haplotypes. Therefore, each risk haplotype has a little more than 1/8 chance to be assigned to a family founder instead of about 1/10. As shown in Table 1, design III had much better performance than design I.

Table 1 Power of VC-score tests under different sample sizes

Design	h^2	Significance level (α)		
		0.05	**1×10^{-5}**	**1×10^{-6}**
I. 500/3000				
	0.01	0.840	0.355	0.270
	0.02	1	0.855	0.780
	0.03	1	0.985	0.980
II. 750/2000				
	0.01	0.890	0.345	0.235
	0.02	1	0.825	0.725
	0.03	1	0.975	0.970
III. 750[a]/2000				
	0.01	0.94	0.435	0.335
	0.02	1	0.92	0.880
	0.03	1	1	0.990

Note.—The design column indicates # of families / # of unrelated individuals. Only nuclear families are simulated, with each family having two parents with a mean of two children.
a. 750 families simulated with enriched risk haplotypes.

We also indirectly compared the performance of the VC and FGLS methods by varying parameters that can affect the effect sizes. We calculated the power of the two methods when the proportion of risk haplotypes was set at 5%, i.e., only 500 haplotypes were tagged as risk in the 10,000 haplotype pool. Although each individual has less chance to carry a risk haplotype, there would be fewer causal variants with larger effect size simulated (if the variance explained by a region is fixed). It was found that both methods had substantial power increase compared to the first simulation, but the VC method had greater improvement than the FGLS. In a simulation set-up where causal SNPs (rare variants only) were not assigned independently (but pairs of SNPs close to each other, and thus correlated, were selected), we found the VC method had a slight power improvement while the FGLS had a small loss in power. Detailed results are listed in Additional file 1. In this work, we did not simulate data with a polygenic term but analyzed the data ignoring it because the results from such a comparison are quite predictable. Because the polygenic terms are correlated among individuals within a family, ignoring such correlations in the analysis will cause a deflated type I error rate and thus render any power comparison invalid.

Many extensions are possible for improved implementation of the proposed model and testing procedure. This method can be easily extended to incorporate nonlinear and interaction effects. As discussed previously, our method can be considered as a special case in the framework of the kernel machine method. Interaction and nonlinear effects among markers can be further included in the model through specifying a valid kernel function or similarity metric. Also, more flexible weights may be incorporated into the kernel function according to allele frequencies or other prior information. Although a normally distributed trait was assumed throughout this study, the derived score statistic is also appropriate for non-normal traits [17]. For binary traits, we can construct the score test analogously, based on the logistic version of the mixed variance model (1) with the outcome y replaced by logit[P($y = 1$)], or via extending the logistic kernel machine method [13]. When there are several correlated traits available, the multivariate variance component model will be very useful because it can have more power than univariate analysis.

Conclusions

We propose a multi-marker VC-based association test using both family and unrelated data. A fast score test has been built on the ML and REML framework, in which only the parameters in the null model need to be estimated. Owing to the non-block-diagonal structure of the genotype-based similarity matrix, the score statistic derived has a different form from that based on the typical VC model for linkage analysis. We demonstrate through simulations that the proposed method achieves good performance in terms of Type I error control and statistical power. The method is implemented in the R package “fassoc”. We believe that “fassoc” will be a useful tool to complement existing software for family-based association studies.

Availability and requirements

Project name: fassoc package
Project home page: https://r-forge.r-project.org/R/?group_id=1379
Operating system(s): Linux, Mac OS X, Windows
Programming language: R
Other requirements: R (≥2.15.1)
License: GNU GPL
Any restrictions to use by non-academics: none except those posed by the license

Additional file

Additional file 1: Additional simulation results and software.

Competing interests

The authors declare that they have no competing interests.

Authors’ contributions

XW and NJM participated in the design of the study and implementation of the method. XW drafted the manuscript. XZ and RCE participated in the conception and design of the study and in editing the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The work was supported by the National Research Foundation of Korea, funded by the Korean Government, grant number NRF-2011-220-C00004, and by the National Institutes of Health, grant numbers HL086718 from the National Heart, Lung and Blood Institute, HG003054 and U01HG006382 from the National Human Genome Research Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Human Genome Research Institute or the National Institutes of Health.

Author details

[1]Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115, USA. [2]Department of Epidemiology and Biostatistics, Case Western Reserve University, Cleveland, OH 44106, USA.

Received: 9 July 2012 Accepted: 11 February 2013
Published: 4 March 2013

References

1. Bodmer W, Bonilla C: **Common and rare variants in multifactorial susceptibility to common diseases.** *Nat Genet* 2008, **40:**695–701.
2. Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747–753.
3. Morris AP, Zeggini E: **An evaluation of statistical approaches to rare variant analysis in genetic association studies.** *Genet Epidemiol* 2010, **34:**188.
4. Zhu X, Li S, Cooper RS, Elston RC: **A unified association analysis approach for family and unrelated samples correcting for stratification.** *Am J Hum Genet* 2008, **82:**352–365.
5. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** *Am J Hum Genet* 2008, **83:**311–321.
6. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** *PLoS Genet* 2009, **5:**e1000384.
7. Han F, Pan W: **A data-adaptive sum test for disease association with multiple common or rare variants.** *Hum Hered* 2010, **70:**42–54.
8. Zawistowski M, Gopalakrishnan S, Ding J, Li Y, Grimm S, Zöllner S: **Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes.** *Am J Hum Genet* 2010, **87:**604–617.
9. Price AL, Kryukov GV, de Bakker PIW, Purcell SM, Staples J, Wei LJ, Sunyaev SR: **Pooled association tests for rare variants in exon-resequencing studies.** *Am J Hum Genet* 2010, **86:**832–838.
10. Feng T, Elston RC, Zhu X: **Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS).** *Genet Epidemiol* 2011, **35:**398–409.
11. Zhu X, Feng T, Li Y, Lu Q, Elston RC: **Detecting rare variants for complex traits using family and unrelated data.** *Genet Epidemiol* 2010, **34:**171–187.
12. Basu S, Pan W: **Comparison of statistical tests for disease association with rare variants.** *Genet Epidemiol* 2011, **35:**606–619.
13. Wu M, Kraft P, Epstein M, Taylor D, Chanock S, Hunter D, Lin X: **Powerful SNP-Set analysis for case–control genome-wide association studies.** *Am J Hum Genet* 2010, **86:**929–942.
14. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: **Rare variant association testing for sequencing data using the sequence kernel association test.** *Am J Hum Genet* 2011, **89:**82–93.

15. Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, Nyholt DR, Madden PA, Heath AC, Martin NG, Montgomery GW, Goddard ME, Visscher PM: **Common SNPs explain a large proportion of the heritability for human height.** *Nat Genet* 2010, **42:**565–569.
16. Yang J, Lee SH, Goddard ME, Visscher PM: **GCTA: a tool for genome-wide complex trait analysis.** *Am J Hum Genet* 2011, **88:**76–82.
17. Tzeng JY, Zhang D: **Haplotype-based association analysis via variance-components score test.** *Am J Hum Genet* 2007, **81:**927–938.
18. Tzeng JY, Zhang D, Chang SM, Thomas DC, Davidian M: **Gene trait similarity regression for multimarker based association analysis.** *Biometrics* 2009, **65:**822–832.
19. Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E: **Efficient control of population structure in model organism association mapping.** *Genetics* 2008, **178:**1709–1723.
20. Lin X: **Variance component testing in generalised linear models with random effects.** *Biometrika* 1997, **84:**309.
21. Zhang D, Lin X: **Hypothesis testing in semiparametric additive mixed models.** *Biostatistics* 2003, **4:**57–74.
22. Liu H, Tang Y, Zhang HH: **A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables.** *Comput Stat Data Anal* 2009, **53:**853–856.
23. Davies RB: **Algorithm AS 155: the distribution of a linear combination of 2 random variables.** *J R Stat Soc Ser C Appl Stat* 1980, **29:**323–333.
24. Farebrother R: **Algorithm AS 204: the distribution of a positive linear combination of 2 random variables.** *J R Stat Soc Ser C Appl Stat* 1984, **33:**332–339.
25. Imhof J: **Computing the distribution of quadratic forms in normal variables.** *Biometrika* 1961, **48:**419–426.
26. Duchesne P, Lafaye De Micheaux P: **Computing the distribution of quadratic forms: further comparisons between the Liu-tang-zhang approximation and exact methods.** *Comput Stat Data Anal* 2010, **54:**858–862.
27. Kwee L, Liu D, Lin X, Ghosh D, Epstein M: **A powerful and flexible multilocus association test for quantitative traits.** *Am J Hum Genet* 2008, **82:**386–397.
28. Liu D, Lin X, Ghosh D: **Semiparametric Regression of Multidimensional Genetic Pathway Data: Least-Squares Kernel Machines and Linear Mixed Models.** *Biometrics* 2007, **63:**1079–1088.
29. Li X, Basu S, Miller MB, Iacono W, McGue M: **A rapid generalized least squares model for a genome-wide quantitative trait association analysis in families.** *Hum Hered* 2011, **71:**67–82.

doi:10.1186/1471-2156-14-17
Cite this article as: Wang *et al.*: A variance component based multi-marker association test using family and unrelated data. *BMC Genetics* 2013 **14**:17.