KernelPvals022916

Multi-locus Association Analysis in Subjects with Arbitrary Genealogical Relationships

Abstract.

Introduction.

Massively parallel genome sequencing promises new insights into the complex genetic basis of common chronic disease. But realizing the potential of this technology requires powerful study designs and sophisticated statistical tests. The enrichment of rare causal variants in affected relatives gives family-based designs a major advantage. Yet transmission disequilibrium tests (TDTs), while robust to population structure, (Spielman et al 1993) are less powerful than case-control designs using unrelated subjects. (Risch and Teng 1998) TDTs lose power because they rely on a few loci to simultaneously provide evidence of association and guard against population stratification. But since genotype data are now routinely available on a genomic scale, methods that use multiple loci to control for population structure promise to provide more powerful and cost-effective tests for association.

While case-control analyses of unrelated subjects provide more powerful association tests than TDT's, simulations have shown that they are less powerful than analyses of *related* cases and unrelated controls, particularly for rare genetic variants. (Teng and Risch 1999) This power gain is due to the enrichment of affected relatives for rare disease-causing variants. In fact, the power gain increases as the causal variant frequency decreases, which is important for the rare variants (frequencies < 0.01) likely to be identified in massively parallel sequencing studies. Increased power is critical because most genetic variants occur at very low frequencies in the recently exploding human population, and huge sample sizes and external biological information will be needed to detect associations with disease. (Tennessen et al 2012, Nelson et al 2012) Biologically-based contrasts between the multi-locus genotypes of cases and controls are likely to be complex, and we need simple, flexible and powerful test statistics whose null distributions can be evaluated in the presence of relatedness and population stratification among subjects.

FIX If a test statistic can be expressed as a linear combination of subjects' genotypes at the loci of interest, large-sample Gaussian approximations to its null distribution can be extended to include pairwise correlations of the subjects' genotypes. (Thornton and McPeek 2010, Zhu and Xiong 2012) The test statistics we consider involve specifying the correlation matrix of subjects' genotypes, either as determined from pedigree structures or as estimated from subjects' genotypes elsewhere in the genome.

We describe multi-locus statistics that test for association between a binary trait and a specified set of M markers in a gene, a pathway or other chromosomal region. INSERT on burden & kernel statistics, Wu, Schaid, burden reference. INSERT on how relative performance of burden and SKAT statistics varies with specific characteristics of target marker set, such as INSERT. None of these test statistics is uniformly most powerful for all possible marker set configurations; what's needed is flexibility and robustness to the many variations in the true state of nature (Lee ref).

Here we describe a family of test statistics with accurate test size when applied to subjects with cryptically or explicitly correlated genotypes. The family includes the (squared) burden statistic and the linear kernel (SKAT) statistic. It also includes a new case-based kernel statistic with increased power for detecting the effects of some types of marker sets. Because the relative power of these three statistics varies with the specific characteristics of the target set of markers, many of which are unknown to the investigator, we need robust test statistics that perform well regardless of the specific characteristics of the target set of interest. Therefore we describe an optimal ensemble test statistic, formed as a linear combination of the burden, SKAT and case-based statistics. We show with simulations that this optimal test performs robustly well regardless of the specific characteristics of the marker set of interest.

Methods.
We assume that the data to be analyzed consist of the genotypes of N subjects at a large set of diallelic markers located across the autosomal genome. We also assume that the trait of interest is dichotomous with N_1 affected individuals (hereafter called *cases*) and $N_0 = N - N_1$ unaffected individuals (*controls*). Interest focuses on the relation between trait risk and a set of M dichotomous markers of interest.

We wish to test the null hypothesis of independence between a vector $Y = (y_1,...,y_N)^T$, whose elements are the N subjects' binary trait indicators, and an NxM matrix $G = (g_{nm})$, whose elements are the subjects' genotypes at the M markers of interest. To do so, we need test statistics and estimates of their null distributions in the presence of correlation among subjects' genotypes. Analogous to the two types of conditioning for prospective and retrospective analyses of binary data, we can do this either by conditioning on the observed trait vector $Y = y$ and regarding the matrix G as random (hereafter called *G|Y conditioning*) (Thornton and McPeek (REF), Zhu and Xiong (2012), Schaid et al (REF)), or by conditioning on the observed genotype matrix G and treating the vector Y as random (*Y|G conditioning*) REFS. The distributions described here are based on *G|Y* conditioning . As noted by Schaid et al (2013), *G|Y* conditioning avoids difficulties in specifying and modeling the phenotype-based ascertainment process for families with multiple affected subjects.

For this *G|Y* conditioning, the null mean of *G* is the *NXM* matrix $E_0[G] = 2\pi^T \otimes 1_N$, where $\pi^T = (\pi_1,\ldots,\pi_M)$ is the *MX1* vector of minor allele frequencies (MAFs) at the *M* SNPs, 1_N is the NX1 vector all of whose elements are one, and $\otimes$ denotes the Kronecker matrix product. The null covariance of G is the MNxMN matrix $Cov_0[G] = \Gamma \otimes \Psi$, where $\Gamma = (\gamma_{mm'})$ is the MxM null covariance matrix of one subject's genotypes at the M SNPs, and $\Psi = (\psi_{nn'})$ is the NxN correlation matrix for subjects' genotypes at one SNP (Thornton and McPeek, Schaid et al, Zhu and Xiong (AJHG 90: 1028-1045, 2012)). The entries $\gamma_{mm'}$ of Γ are $\gamma_{mm'} = \sigma_m \sigma_{m'} r_{mm'}$ where $\sigma_m = \sqrt{2\pi_m(1-\pi_m)}$ and $r_{mm'}$ is the correlation coefficient between SNPs *m* and *m', m,m' = 1,...,M*. We assume that external data can be used to specify the MAFs $\pi_1,\ldots,\pi_M$, the pairwise marker correlation coefficients $r_{mm'}$, and the interpersonal genotype correlation coefficients $\psi_{nn'}$. The latter can estimated from known family pedigree structures and/or from the subjects' genotypes at markers independent of those in the target set. Finally, we introduce user-specified trait probabilities p_n satisfying

$$\sum_{n=1}^{N} p_n = \sum_{n=1}^{N} y_n = N_1. \quad (1)$$

For example, p_n might be the predicted value of y_n obtained by fitting a logistic regression model of trait phenotypes against nongenetic covariates or principle components of ancestry.

Test statistics. We consider the family of test statistics *Q* having the quadratic form

$$Q = z^T A z, \quad (2)$$

where for G|Y conditioning, z is an Lx1 vector whose asymptotic null distribution is multivariate Gaussian with mean μ and positive definite covariance matrix V, and A is an LXL symmetric matrix of constants. We focus on statistics (2) with $L = 2M$ and

$$z^T = (y^T GW,\ p^T GW), \quad (3)$$

where $y^T = (y_1,...,y_N)$, $p^T = (p_1,...,p_N)$ and $W = \Delta(w_1,...,w_M)$ denotes an *MxM* diagonal matrix whose diagonal entries are user-specified positive marker weights $w_1,...,w_M$. Note that the *2Mx1* random vector *z* of (3), whose elements are weighted sums of the N subjects' genotypes, has a multivariate Gaussian asymptotic null distribution as $N \to \infty$. The user-specified null mean and covariance matrix of this distribution are

$$(E_0[z])^T = 2N_1(\pi^T W, \pi^T W) \equiv \mu^T, \quad (4)$$

and

$$\mathrm{Cov}_0(z) = \left[F^T \Psi F\right] \otimes \left(W \Gamma W\right) \equiv V, \tag{5}$$

where *F = (y,p)* has dimension NX2.

We consider the statistics determined by three choices for the symmetric matrix A:

$$A_B = \begin{pmatrix} J_M & -J_M \\ -J_M & J_M \end{pmatrix}, \; A_S = \begin{pmatrix} I_M & -I_M \\ -I_M & I_M \end{pmatrix} \text{ and } A_C = \begin{pmatrix} I_M & 0 \\ 0 & -I_M \end{pmatrix}, \tag{6}$$

where J_M is the *MxM* matrix whose elements are all one, and I_M is the *MxM* identity matrix. The matrix A_B corresponds to the (squared) burden statistic Q_B, while A_S gives the SKAT statistic Q_S, and A_C gives a new case-based kernel statistic, denoted Q_C. We shall show with simulations that Q_C can be more powerful than Q_B and Q_S when applied to some marker sets. For example, suppose we have equal numbers $N_1 = N/2$ of cases and controls, with genotypes coded as indicators of variant carriage, and with all trait probabilities $p_n = 0.5$, $n = 1,\ldots,N$. Then the trait residuals $y_n - p_n$ of cases and controls are 0.5 and -0.5, respectively. In this case it can be shown that Q_S is proportional to the number of variants shared by phenotypically concordant (case-case and control-control) pairs of subjects, minus twice the number shared by phenotypically discordant (case-control) pairs. Lack of allele-sharing by pairs of control subjects may attenuate the magnitude of Q_S. In contrast, the test statistic Q_C avoids this problem by contrasting observed genotype similarities among case pairs with their null expectations based on all pairs of subjects.

Previous simulations have shown that the relative power of the burden and SKAT statistics Q_B and Q_S varies with the characteristics of the marker set. (REFS) To address this variation, Lee et al (2012) proposed a test statistic Q_{ρ_*} obtained by optimizing a set of linear combinations $Q_\rho = \rho Q_S + (1-\rho) Q_B$ of the SKAT and burden statistics, where the parameter ρ ranges over a grid on the unit interval. Assuming Y|G conditioning, the authors chose the optimal statistic Q_{ρ_*} as that value ρ_* with minimum P-value among those on the grid. Simulations described in the next section indicate that the relative power of the statistics Q_B, Q_S, Q_C also varies with characteristics of the marker set of interest. To accommodate this variation, we propose adapting and extending the approach of Lee et al (2012) by optimizing (under G|Y conditioning) the set of all linear combinations

$$Q_\alpha = \alpha_B Q_B + \alpha_S Q_S + \alpha_C Q_C, \tag{8}$$

whose coefficients $\alpha = (\alpha_B, \alpha_S, \alpha_C)$ are nonnegative and sum to one. Statistics of the form (8) have the quadratic form (2) with *z* given by (3) and

$$A = A_\alpha = \begin{pmatrix} \alpha_B J_M + (\alpha_S + \alpha_C) I_M & -(\alpha_B J_M + \alpha_S I_M) \\ -(\alpha_B J_M + \alpha_S I_M) & \alpha_B J_M + (\alpha_S - \alpha_C) I_M \end{pmatrix}. \tag{9}$$

Figure 1 shows the closed triangle ζ containing points $\alpha = (\alpha_B, \alpha_S, \alpha_C)$ whose values determine the statistics Q_α of (8). The three corners of this triangle correspond to the three test statistics Q_B, Q_S, Q_C. The three edges correspond to linear combinations of the three pairs of statistics: (*B,S*) (Lee et al 2012), (*B, C*) and (*S, C*), and the interior points of the triangle correspond to weighted sums of all three statistics.

Null distributions of test statistics.

To estimate the asymptotic null distribution of the test statistics Q_{α_*} under *G|Y* conditioning, we use the following three steps, whose implementation is described in the Appendix:

Step 1. Estimate the asymptotic null distribution of any test statistic satisfying (2,3), with fixed matrix A_α. **INSERT** on how the estimated tail probabilities of this distribution differ from those described by Schaid et al (REF).
Step 2. Choose a grid of points α in the triangle ζ and calculate the corresponding test statistics Q_α of (8). For each point α in the grid, use the estimated tail probabilities in Step 1 to calculate a significance level $P(\alpha)$ for the statistic Q_α. Define Q_{α_*} as the statistic corresponding to the value α_* that minimizes the values $P(\alpha)$ over the grid.
Step 3. Use a genome-wide resampling strategy to estimate the null distribution of the optimized statistic Q_{α_*}.

Simulations.
We use simulations to evaluate the size and power of the seven test statistics shown in Table 1. Because the relative power of the three test statistics Q_B, Q_S, Q_C varies with the many characteristics of the targeted marker set, and because the some of these characteristics are unknown to the investigator, we want robust test statistics that perform well regardless of the specific characteristics of the target set of interest.

Data generation.
Generating genotypes for a target set of markers. We generated sequence data of European ancestry from 5,000 chromosomes covering a 50 kb region using the calibrated coalescent model of Schaffner et al (2005). We then selected the set of all markers with MAFS between 0.005 and 0.03 and such that all pairwise correlation coefficients were less than 0.95; from these, we selected two subsets of markers, based on their overall pairwise correlations: pairs of markers in subset 1 (N = X) were highly correlated, while those in subset 2 (N = X) were only weakly correlated. This sampling yields a source matrix S of dimension 5000x102, with each row corresponding to a haplotype whose column entries are indictors for the minor allele at the 102 loci.

Application to prostate cancer data.

REFERENCES

Schaffner S.F., Foo, C., Gabriel, S., Reich, D., Daly, M.J., and Altshuler, D. (2005). Calibrating a coalescent simulation of human genome sequence variation. Genome Res. 15, 1576–1583.

Lee et al Biostatistics 13: 762, 2012.

APPENDIX: Estimating the null distribution of the optimized test statistic Q_{α_*}

APPENDIX: Estimating the null distribution of the optimized test statistic Q_{α_*}

Step 1. Null distribution of statistics Q_α for fixed α. Under G|Y conditioning and for fixed α, a closed-form representation of the asymptotic null distribution of Q_α of (8) can be obtained from the theory for quadratic forms in Gaussian vectors *z*. Specifically, dropping the subscript α, we rewrite equation (2) as

$$Q = \tilde{z}^T \tilde{A} \tilde{z}, \qquad \text{(A.1)}$$

with $\tilde{z} = V^{-1/2} z$ & $\tilde{A} = V^{1/2} A V^{1/2}$, where *V* is the nonsingular covariance matrix of *z* given by equation (5), and $(V^{1/2})V^{1/2} = V$. We perform a singular value decomposition of $\tilde{A}$ as

$$\tilde{A} = U^T \Lambda U, \qquad \text{(A.2)}$$

where *U* is the LxL orthonormal matrix of eigenvectors of $\tilde{A}$ and Λ is the *LxL* diagonal matrix whose diagonal entries are the eigenvalues λ_ℓ of $\tilde{A}$, $\ell = 1,\ldots,L$. Substituting (A.2) into (A.1) yields

$$Q = x^T \Lambda x = \sum_{\ell=1}^{L} \lambda_\ell x_\ell^2, \qquad \text{(A.3)}$$

where $x = U\tilde{z}$ has covariance equal to the identity matrix of dimension *L*. Thus *Q* is a mixture of the independent noncentral chi-squared variables x_ℓ^2, $\ell = 1,\ldots,L$. We determined significant levels and critical points of its distribution using the Davies exact method (Davies 1987).

Step 2. Determining the optimal statistic Q_{α_*} We proceed as follows:

a. Select a grid of *I* points $\{\alpha_i = (\alpha_{iB}, \alpha_{iS}, \alpha_{iaS})\}$ in the closed triangle ζ shown in Figure 1. Compute $Q_{\alpha_i} = Q_{\alpha_i}(G)$, for $i = 1,\ldots,I$.

b. Use Step 1 to compute significance levels $P(\alpha_i)$ equal to the null probability of observing a value greater than $Q_{\alpha_i}(G)$.

c. Define Q_{α_*} as the statistic corresponding to $\alpha_* = \arg\min_{i=1}^{I} \{P(\alpha_i)\}$, and define $X_{\alpha_*}(G) = -\log_{10} P(\alpha_*)$.

Step 3. Null distribution of optimized statistics Q_{α_*}. To determine the significance level of an optimized statistic Q_{α_*} under *G*|*Y* conditioning, we repeatedly resample subjects' genotypes at markers elsewhere in the autosomal genome that are in linkage equilibrium with all known trait-related loci, and with the target markers. Specifically, we first construct, for each marker in the target set, a sampling set of 500-1000 other markers with minor allele frequencies equal to that of the target marker (+/- 0.0x). Then in resampling replication s, we select one marker at random from each of the M sampling sets and create: i) the NxM pseudo-genotype matrix $\tilde{G}^s$ whose rows contain the subjects' genotypes at the M resampled loci, and ii) the corresponding statistic $\tilde{Q}_\alpha^s$. We then repeat Step 2 to compute $X_{\alpha_*}(\tilde{G}^s)$. We repeat this procedure *S* times to estimate the null distribution of $X_{\alpha_*}(G)$ and to find the significance level of Q_{α_*}. The value of S depends on the nominal test size used. For test size P = 0.01 we used S = 1000 replications.

Table 1. Test Statistics evaluated in Simulations

Test Statistic	Q_B	Q_S	Q_C	Q_{BS}	Q_{BC}	Q_{SC}	Q_{BSC}
Weights $(\alpha_B, \alpha_S, \alpha_{aS})$	(1,0,0)	(0,0,1)	(1,0,0)	$(1-\alpha_S^*, \alpha_S^*, 0)^a$	$(0, \alpha_S^*, 1-\alpha_S^*)$	$(\alpha_B^*, 0, 1-\alpha_B^*)$	$(\alpha_B^*, \alpha_S^*, \alpha_C^*)$

a) asterisks on weights indicate their optimized values