RESEARCH ARTICLE

Genetic Epidemiology

OFFICIAL JOURNAL
INTERNATIONAL GENETIC EPIDEMIOLOGY SOCIETY
www.geneticepi.org

Adjusted Sequence Kernel Association Test for Rare Variants Controlling for Cryptic and Family Relatedness

Karim Oualkacha,[1,2,3] Zari Dastani,[1,2] Rui Li,[1] Pablo E. Cingolani,[4,5] Timothy D. Spector,[6] Christopher J. Hammond,[6] J. Brent Richards,[1,2,7,8] Antonio Ciampi,[2] and Celia M. T. Greenwood[1,2,8,9]*

[1]*Lady Davis Institute for Medical Research, Jewish General Hospital, Montreal, QC, Canada;* [2]*Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montreal, QC, Canada;* [3]*Département de Mathématiques, Université du Québec à Montréal, QC, Canada;* [4]*Department of Computer Science, McGill University, Montreal, QC, Canada;* [5]*McGill University and Genome Quebec Innovation Centre, Montreal, QC, Canada;* [6]*Department of Twin Research and Genetic Epidemiology, King's College London, London, United Kingdom;* [7]*Department of Medicine, Jewish General Hospital, McGill University, Montreal, QC, Canada;* [8]*Department of Human Genetics, McGill University, Montreal, QC, Canada;* [9]*Department of Oncology, McGill University, Montreal, QC, Canada*

Received 18 April 2012; Revised 20 February 2013; accepted revised manuscript 25 February 2013.
Published online 25 March 2013 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/gepi.21725

ABSTRACT: **Recent progress in sequencing technologies makes it possible to identify rare and unique variants that may be associated with complex traits. However, the results of such efforts depend crucially on the use of efficient statistical methods and study designs. Although family-based designs might enrich a data set for familial rare disease variants, most existing rare variant association approaches assume independence of all individuals. We introduce here a framework for association testing of rare variants in family-based designs. This framework is an adaptation of the sequence kernel association test (SKAT) which allows us to control for family structure. Our adjusted SKAT (ASKAT) combines the SKAT approach and the factored spectrally transformed linear mixed models (FaST-LMMs) algorithm to capture family effects based on a LMM incorporating the realized proportion of the genome that is identical by descent between pairs of individuals, and using restricted maximum likelihood methods for estimation. In simulation studies, we evaluated type I error and power of this proposed method and we showed that regardless of the level of the trait heritability, our approach has good control of type I error and good power. Since our approach uses FaST-LMM to calculate variance components for the proposed mixed model, ASKAT is reasonably fast and can analyze hundreds of thousands of markers. Data from the UK twins consortium are presented to illustrate the ASKAT methodology.**

Genet Epidemiol 37:366–376, 2013.

KEY WORDS: **gene-based test; rare variant methods; resequencing; mixed models; family-based design; score statistics; linear kernel function**

Introduction

Genome-wide association studies (GWAS) have identified hundreds of common variants that appear associated with complex diseases or related traits, but these common variants explain only a small proportion of heritability for most diseases [Dickson et al., 2010; Eichler et al., 2010]. Re-sequencing is now being used to characterize genetic diversity between individuals, and is well known to identify substantial amounts of rare genetic variation. For example, sequencing of the melatonin receptor MTNR1B in almost 8,000 individuals identified 40 rare variants, and furthermore type II diabetes risk was shown to be associated with the rare genetic variation in this gene [Bonnefond et al., 2012].

As a result of the low frequency of many of the genetic variants identified by re-sequencing, statistical methods that test single genetic variants one at a time have very low power for detecting associations. Consequently, many approaches have been developed recently that consider simultaneously a group of rare variants in a chosen region of interest. In this context, the primary hypothesis is that multiple genetic variants in the chosen region could have influence on the disease or trait of interest.

Many of the proposed methods for testing region-based genetic associations are based on the idea of constructing a linear combination of a sequence of r rare variants that summarizes the information across the variants. Then association is tested between the linear combination of genetic variants and the target phenotype, often linking the two by (generalized) linear models [Li and Leal, 2008; Lin and Tang, 2011; Madsen and Browning, 2009]. Due to the rarity of the variants, usually the weights of the linear combination are chosen in advance, and some methods allow the incorporation of quality scores for each variant [Daye et al., 2012; Lin and Tang, 2011; Wu et al., 2011; Yi and Zhi, 2011]. Wu et al. [2011] proposed a score test (called SKAT [sequence

†Contract grant sponsor: CIHR; Contract grant numbers: MOP-115110; MOP-119462; Contract grant sponsors: MDEIE, CFI, and CQDM.

*Correspondence to: Celia M. T. Greenwood, 3755 Côte Ste Catherine, H414, Montreal, Quebec, Canada, H3T 1E2. E-mail: celia.greenwood@mcgill.ca

kernel association test]) looking for the effects of any of the (rare) genetic variants in the chosen region. They also showed that the significance levels could be evaluated asymptotically, without need for permutations. Their method is essentially a mixed model where the random component is the genetic variation at the region of interest and the fixed component includes any other covariate effects.

In GWAS of single nucleotide polymorphisms (SNPs), analysis of families or related individuals is often performed through the use of linear mixed models (LMMs) that account for the relationships; these LMMs can also be used to control for population structure and cryptic relatedness [Hayes et al., 2009; Kang et al., 2008; Yu et al., 2006]. Relationships between individuals can be either known, or estimated from the marker data. Estimated kinship matrices have been used in several LMMs [Aulchenko et al., 2007a; Hayes et al., 2009; Kang et al., 2010, 2008; Yu et al., 2006]. Yu et al. [2006] generalized the concept by using a fixed effect covariate to correct for population structure, based on a matrix which measures the population admixture proportions, together with a random polygenic effect for marker-based family relatedness.

When the kinship matrix is estimated from marker data, even individuals not known to be related can have nonzero kinship, and so the mixed model depends on a matrix of size $N \times N$, where N is the number of individuals. Hence, implementation with thousands of individuals and millions of markers leads to heavy computations for the estimation of the model parameters (i.e., the variance components). To remedy this problem for family-based designs, Aulchenko et al. [2007a] proposed to first estimate the random component of the mixed model to correct for relatedness, and to construct approximately uncorrelated residuals that can then be used to test for association at a large number of SNP markers using a simple trend test. Kang et al. [2008] proposed an efficient mixed-model association (EMMA) for population-based designs including a random effect to allow for nonspecified genetic effects and fixed effects for an SNP of interest. Compared to Yu et al. [2006], the EMMA method increased computational speed and improved the estimation of the variance components by achieving near-global likelihood optimization. Kang et al. [2010] considered an EMMA expedited approach (EMMAX) that also reduced the computational time for large GWASs. Similar to Aulchenko et al. [2007a], Kang et al. [2010] estimated the variance components only once for a given data set, and then used these components to calculate a generalized least squares estimate of each marker effect. More recently, Lippert et al. [2011] obtained increased computational speed by spectral factoring of the large matrices in a mixed model (FaST-LMM where FaST is factored spectrally transformed). However, all the above methods focused on common variants and single-SNP analyses.

Methods for the analysis of family-data for region-based rare variant tests have been lacking to date. Lin and Tang [2011] stated without detail that it would be possible to extend their approach to family-based data using generalized LMMs (GLMMs). However, they did not show any results of such an extension. Here, we extend the SKAT approach of Wu et al. [2011] to test for region-based rare variant associations while allowing for correlation between individuals due to relatedness, and combine this with the spectral factoring of Lippert et al. [2011] in FaST-LMM for computational speed. Our approach, which we term ASKAT (adjusted sequence kernel association test), captures sample structure based on a LMM that incorporates the proportion of the genome that is identical by descent (IBD) between pairs of individuals. The IBD proportions can either be calculated using the expectations based on familial relationships, or estimated from the covariances of genome-wide allele counts between pairs of individuals. This latter is termed the realized relationship matrix (RRM) [Amin et al., 2007; Astle and Balding, 2009].

We test the association between a continuous phenotype and a group of variants within a given region using a variance-component score test, and adjust for relatedness through an additional random polygenic effect. We calculate the score statistic using restricted maximum likelihood (REML), and thus our score statistic does not depend on fixed effects. Furthermore, the test depends only on the estimates of the variance components under the null model. Since the IBD proportions are either based on family structures or estimated from whole genome marker data, we assume that the variance of this additional random effect is the same across the genome. Therefore, we need to estimate the variance components only once during the mapping, under the null model.

Performance of this approach is demonstrated in simulated data of families of varying size. In addition to the simulation results, the performance of ASKAT is demonstrated through single-SNP analyses of one chromosome in a cohort of twins, where we show that the type I error is appropriately controlled.

Methods

Notation and Model Setup

Let Y_i, $i = 1, \ldots N$ represent a continuous phenotype measured on N individuals. The individuals may belong to families, but for simplicity we do not introduce a second subscript. For subject i, let $X_i = (X_{i1}, \ldots, X_{im})^T$ be a vector of m covariates. In addition, assume there are V loci that influence the phenotype Y, and that at each locus, there are r_v genetic variants that have been genotyped. Let $G_i^{(v)} = (G_{i1}^{(v)}, \ldots, G_{ir_v}^{(v)})^T$ represent the minor allele counts at r_v variant sites in the vth genomic region of interest. We are interested in constructing a test that considers jointly the r_v variant sites at one locus.

Although testing for genetic associations is usually performed one locus at a time, the "true" model for phenotype Y may depend on many genes [Kang et al., 2010]. Suppose that the true model explaining the genetic contributions to Y can be written as

$$Y_i = \mu + X_i^T \eta + \sum_{v=1}^{V} \left\{ G_i^{(v)} \right\}^T \beta^{(v)} + E_i, \qquad (1)$$

where μ is the general mean, $\eta = (\eta_1, \ldots \eta_m)^T$ is a vector of fixed effects for the covariates, $\beta^{(v)} = (\beta_1^{(v)}, \ldots \beta_{r_v}^{(v)})^T$ is a vector of regression coefficients for the genetic variants, and E_i is a random error which is assumed to be independent and identically distributed across individuals, with mean zero and variance σ_e^2. In many tests of association for either common or rare SNP effects, testing is performed at one locus v_0 and the others are subsumed into the error term, so that $\tilde{E}_i = \sum_{v \neq v_0}^{V} \{G_i^{(v)}\}^T \beta^{(v)} + E_i$. In fact, this implies that the errors of the assumed model are no longer independent and identically distributed, and hence this model is inherently misspecified. We can therefore argue that the misspecified model ignores the polygenic background of the phenotype, and thus the phenotype distribution will depend on the sample structure. It follows that even in the absence of known family relationships, that it may be important to account for polygenic genetic background.

Adjusted Sequence Kernel Association Test (ASKAT)

We consider here a genetic LMM which captures the polygenic inheritance by using kinship coefficients. These can be either the expected kinships based on family data, or the estimated kinship coefficients from genomic data [Kang et al., 2010; Lippert et al., 2011; Yu et al., 2006]. Since we focus on a specific region, we are assuming that the effects of the other regions are absorbed into the polygenic variance. Simplifying the notation a little, the model (1) can be rewritten as

$$Y_i = \mu + X_i^T \eta + G_i^T \beta + P_i + E_i, \tag{2}$$

where $\beta = (\beta_1, \ldots \beta_r)^T$, and P_i denotes the polygenic random effect which accounts for the other loci affecting the trait; P_i is assumed to have a normal distribution with mean zero and variance σ_g^2.

Ignoring dominance effects, the contribution of the genome sharing between individuals i and j in the covariance can be written as

$$\mathrm{Cov}(Y_i, Y_j) = \mathrm{Cov}(P_i, P_j) = \Phi_{ij} \sigma_g^2,$$

where Φ is an $N \times N$ matrix with entries reflecting the proportion of the genome that is IBD between pairs of individuals. If the individuals belong to a set of known families, this could be a block diagonal matrix containing expected IBD proportions. For instance, between full sibs, the predicted proportion of the genome that is IBD is 1/2, 1/4 for grandparent-grand-child pairs, and 1/8 for cousins. However, the matrix can also be estimated from the realized proportion of the genome that is IBD between pairs of individuals (RRM) based on a large number of genotyped markers spanning the genome [Amin et al., 2007; Astle and Balding, 2009]. Provided that a sufficient number of markers is used, the matrix Φ can be estimated with a high degree of precision [Astle and Balding, 2009; Hayes et al., 2009]. In fact, since there is known to be variation in the amount of the genome shared IBD between relatives, the RRM may be a more accurate estimate of the IBD proportion than the prior expectation based on known relationships.

A similar model, without the polygenic random component, forms the basis of the tests proposed by several other groups ([Daye et al., 2012; Lin and Tang, 2011; Wu et al., 2011; Yi and Zhi, 2011] and others). In these methods, the variants in a chosen region are linked to the phenotype under study by

$$Y_i = \mu + X_i^T \eta + G_i^T \beta + E_i. \tag{3}$$

Due to the assumed rarity of the causal variants, it may be impossible to estimate all the coefficients β_r. Therefore, in SKAT [Wu et al., 2011], β is treated as a random vector following an arbitrary distribution with mean zero and variance

$$\mathrm{Var}(\beta) = \tau W, \quad W = \mathrm{diag}(w_1, \ldots, w_r), \tag{4}$$

where τ is a variance component and W is an $r \times r$ diagonal matrix of the weights to be used for the r variants. In this context, testing for $\beta = 0$ is equivalent to testing $\tau = 0$, and Wu et al. [2011] proposed to do this via a variance-component score test. A key advantage of the score test is that it fits only the null model, i.e., a standard regression model in this case. After correcting an error in Equation (3) of Wu et al. [2011], the score test statistic is

$$Q = \frac{(Y - \hat{\mu})^T K (Y - \hat{\mu})}{2\hat{\sigma}_e^2},$$

where $K = G^T W G$, $G = (G_1, \ldots, G_N)$, an $r \times N$ matrix, $\hat{\mu}$ and $\hat{\sigma}_e^2$ are the predicted mean and the mean squared error of Y under H_0, respectively. Wu et al. [2011] noted that the matrix K measures the genetic similarity between individuals in the region by using the r variants.

The idea of our proposed method is to assume that β satisfies (4), to add the residual polygenic random effect to the model and to test for $H_0 : \tau = 0$ using a variance-component score test in the extended LMM (2). The variance-covariance matrix of Y, the vector of all N individuals, is given as

$$\mathrm{Var}(Y) = \tau K + \sigma_g^2 \Phi + \sigma_e^2 I_N,$$

where I_N denotes the identity matrix. Since we are implementing a variance-component score test, we calculate the score test statistic based on the REML function, which finds a translation invariant score statistic that does not involve fixed effects. Standard elementary algebraic calculations [Searle, 1979; Rao and Kleffe, 1988; Quaas, 1992] show that the REML score statistic can be given under the null model as

$$Q = \frac{1}{2} Y^T P_0 K P_0 Y, \tag{5}$$

where

$$P_0 = \Omega_0^{-1} - \Omega_0^{-1} X_1 \left(X_1^T \Omega_0^{-1} X_1 \right)^{-1} X_1^T \Omega_0^{-1},$$

and

$$\Omega_0 = \sigma_g^2 U S U^T + \sigma_e^2 I_N \quad X_1 = (\mathbb{1}_N, X),$$

where $X = (X_1, \ldots, X_N)^T$ and $\mathbb{1}_N$ refers to the vector of 1's of dimension N. The matrices U and S satisfy $\Phi = U S U^T$, the spectral decomposition of Φ [Lippert et al., 2011].

The computational challenge is the estimation of the variance components σ_g^2 and σ_e^2, since the computing time for solving (2) increases with the cube of N. The advantages of our approach are: first, since it is a score test, only the null model ($\beta = 0_r$) needs to be fit. The statistic Q depends only on the estimates of the variance components σ_g^2 and σ_e^2 under the null model. Second, spectral decomposition of Φ facilitates greatly the calculation of the inverse $\Omega_0^{-1} = U(\sigma_g^2 S + \sigma_e^2 I_N)^{-1} U^T$, since $(\sigma_g^2 S + \sigma_e^2 I_N)$ is a diagonal matrix. Moreover, to estimate the variance components under the null, we suggest the FaST-LMM program which is computationally optimized. Finally, since the estimated kinship matrix is based on genotype data from the whole genome, and since we assume that each locus explains only a very small amounts of the trait heritability, we assume that the polygenic variance component is the same across genomic regions. Therefore, the variance components only need to be estimated once.

The key insight behind the spectral decomposition is that it projects individuals to uncorrelated directions under the null ($\tau = 0$). Indeed, one can easily verify that the variance-covariance matrix of $U^T Y$ can be written as

$$\mathrm{Var}(U^T Y) = \tau U^T K U + \sigma_e^2 D, \qquad (6)$$

where D is a diagonal matrix given by $D = (\sigma_g^2/\sigma_e^2)S + I_N$. Thus, the matrix $\tilde{K} = U^T G^T W G U$ can be viewed as a known weighted linear kernel function which measures similarity between individuals in the whole genome. Hence, $\tilde{K}$ controls for family relationships and population structure simultaneously.

The weight matrix W plays the same role here as in the SKAT model. As discussed by Wu et al. [2011], different formulations for $\tilde{K}$, obtained by changing the weight matrix W, will alter power, and will be optimal for different genetic models. Type I error, however, is protected for any choice of the weights. Wu et al. [2011] discussed in detail several choices for the elements of the matrix, w_r. In the presence of sequencing errors, a quality-weighted multivariate score association test (qMSAT) was proposed by Daye et al. [2012]. This approach incorporated subject-variant sequencing quality scores (e.g., the probability of a given rare variant genotype being correctly called for a given subject) into the unweighted version of the SKAT model, and showed a substantial gain in power over current methods, including the unweighted version of SKAT.

The score statistic Q given by (5) is a nonnegative quadratic form of Y. If we assume that σ_g^2 is known, then under the null hypothesis, Q's distribution is a mixture of chi-square distributions. Under this assumption, both empirical and exact methods can be used to calculate the P-values [Davies, 1980; Duchesne and Lafaye De Micheaux, 2010; Liu et al., 2009]. Although this is not exact since σ_g^2 is unknown, in our simulations, we did not see evidence for departures from this expected distribution.

Lin and Tang [2011] proposed a general framework for detecting disease associations with rare variants. They assumed that the β given in (3) is a fixed effect which satisfies $\beta = \tau w^{1/2}$, $w^{1/2} = (\sqrt{w_1}, \ldots, \sqrt{w_r})^T$. Thus, $R_i = G_i^T w^{1/2}$ is the weighted linear combination of G_{ij}, $j = 1, \ldots, r$ with G_{ij} receiving weight $\sqrt{w_j}$. These authors also proposed testing $H_0 : \tau = 0$ using a fixed effect score test. Their score statistic is given by

$$Q_L = \sum_{j=1}^{r} \sqrt{w_j}(Y - \hat{\mu})^T G_{.j}, \quad G_{.j} = (G_{1j}, \ldots, G_{nj})^T,$$

where $\hat{\mu}$ is the predicted mean under the null given in (5). Note that the SKAT statistic given by (5) can be written in a similar way as

$$Q_{SKAT} = \sum_{j=1}^{r} w_j \left\{ \frac{1}{\sqrt{2}\hat{\sigma}_e}(Y - \hat{\mu})^T G_{.j} \right\}^2,$$

where $\hat{\sigma}_e^2$ is the mean sum of square error of Y under H_0 of the SKAT model (3). Rather than prespecifying the weights of variants as the SKAT approach, Lin and Tang [2011] estimated the weights from the data. If the choice of the weight vector $w^{1/2}$ is not proportional to the true β or is estimated using data, then Q_L is no longer the score statistic. However, Lin and Tang showed that regardless of how $w^{1/2}$ is estimated, the statistic Q_L is asymptotically normal under H_0, as long as $\hat{w}^{1/2}$ converges to a constant vector as the sample size N increases. Although they mentioned that their approach could be extended to family data, no details are given.

Aulchenko et al. [2007a] developed an approach for family-based quantitative trait loci (QTL) association analysis called "genome-wide rapid association using mixed model and regression" (GRAMMAR). This method first fits the background model (2) without any genetic SNP data, then obtains residuals once the polygenic variance component σ_g^2 has been estimated. The association between these residuals and the genetic variants is then estimated using standard least-squares methods (simple regression models). This two-step method leads to conservative results, and this tendency is more pronounced for traits with higher heritability [Aulchenko et al., 2007a]. Although GRAMMAR was developed for single-SNP analyses, it is worth noting that it is possible to use GRAMMAR's residuals in SKAT or in another existing region-based test after adjusting for family structure. However, as for single-SNP tests, the region-based tests are conservative.

Results

Simulation Studies

Here, we show results of several simulation studies to illustrate the methodology presented in the previous sections and demonstrate the performance of the proposed method. Performance here is measured by both type I error and power.

Family-Based QTL Simulation

In all simulations, we assumed the genetic model (2) containing L disease susceptibility loci (SNPs or sequence

variants), where the lth SNP has two alleles denoted by d_l and D_l with frequencies f_l and $(1-f_l)$, respectively, (d_l is the susceptibility allele):

$$Y_i = \mu + \sum_{l=1}^{L} G_{il}\beta_l + P_i + E_i,$$

where μ is the general mean and P_i and E_i are the residual polygenic and the environmental components which are normal with mean zeros and variances σ_g^2 and σ_e^2, respectively. G_{il} is the number of the risk alleles d_l, carried by the ith subject, at the lth locus. The scalar $\beta_l \in \Re$ represents the effect of the susceptibility alleles; thus, the effect of allele d_l on the trait is assumed to be additive (i.e., carrying one copy of d_l adds β_l to the mean of the traits). The effect β_l is assumed to be the same for each family and its magnitude is explained in term of the lth locus QTL-heritability, $h^2_{QTL(l)}$, according to the following relationships:

$$h^2_{QTL} = \sum_{l=1}^{L} h^2_{QTL(l)},$$

$$h^2_{QTL(l)} = \frac{\sigma^2_{QTL(l)}}{\sum_{l=1}^{L} \sigma^2_{QTL(l)} + \sigma_g^2 + \sigma_e^2},$$

$$\sigma^2_{QTL(l)} = 2f_l(1-f_l)\beta_l^2, \quad l = 1, \ldots, L, \tag{7}$$

$$h^2 = \frac{\sum_{l=1}^{L} \sigma^2_{QTL(l)} + \sigma_g^2}{\sum_{l=1}^{L} \sigma^2_{QTL(l)} + \sigma_g^2 + \sigma_e^2},$$

where $\sigma^2_{QTL(l)}$ is the lth locus additive genetic variance, h^2_{QTL} is the proportion of the trait variation that is explained by the region under study, and h^2 is the total trait heritability including both the region and the polygenic effects.

Simulation Under a Single Locus Model

For our first simulations, we set $L = 1$, and we simulated a single quantitative trait locus influencing the continuous trait Y. Note that since standard family-based association tests do not handle simultaneously a group of SNPs, this single locus setting is used to compare our approach with two existing family-based association tests, FaST-LMM [Lippert et al., 2011] and GRAMMAR [Aulchenko et al., 2007a], which scale efficiently to GWAS but which analyze only a single SNP at a time.

Specifically, for this simulation, we assumed

$$L = 1, \quad f_1 = 0.01, \quad \sigma_e^2 = 1, \quad \mu = 3.$$

To calculate the type I error and the power, we assumed the QTL-heritability $h^2_{QTL(1)} = 0$ under H_0, and $h^2_{QTL(1)} = 0.01$ under H_1. We conducted simulation studies with a sample size of $N = 1{,}160$ and we varied the total trait heritability: $h^2 \in \{5\%, 35\%, 60\%\}$. Using these parameters, we generated 160 families; 40 families of two parents and one child, 40 families of two parents and two children, 40 families of two generations (eight subjects per family), and 40 families of three generations (14 subjects per family). We performed 1,000 simulations for power and 5,000 simulations for type I error. To simulate genotypes at the SNP locus, we used the SIMULATE program [Terwilliger et al., 1993], and we used the theoretical kinship matrices to simulate the normal polygenic effect P.

For each method, we calculated the proportion of simulations rejecting the null hypothesis at the given significance threshold α (the empirical P-value or type I error), and also the power at the given α, where the P-values are calculated using the Davies approximation.

From Table 1, it can be seen that our method and FaST-LMM control the type I error well. However, as expected, SKAT rejects the null hypothesis more often than expected, and the inflation of type 1 error worsens as h^2 increases. In contrast, the GRAMMAR approach is conservative and fewer simulations reject the null hypothesis than would be expected. Table 2 demonstrates the power to detect single-SNP associations, assuming that the asymptotic thresholds for rejection are accurate. ASKAT and FaST-LMM have similar power, both much better than GRAMMAR, especially as h^2 increases. Of course, part of the improvement is due to the conservativeness of the GRAMMAR approach. However, both ASKAT and FaST-LMM can be seen to be comparable

Table 1. Type I error rates for single-SNP association tests in simulated family data. The table shows the proportion of 5,000 simulations where the test statistic exceeded the expected quantile at level α. Our method (ASKAT) is compared to SKAT, FaST-LMM (F-L), and GRAMMAR (GR)

	$h^2 = 5\%$				$h^2 = 35\%$				$h^2 = 60\%$			
α Level	ASKAT	SKAT	F-L	GR	ASKAT	SKAT	F-L	GR	ASKAT	SKAT	F-L	GR
0.1	0.101	0.109	0.102	0.091	0.096	0.149	0.097	0.053	0.112	0.195	0.112	0.018
0.05	0.048	0.056	0.048	0.045	0.052	0.089	0.052	0.019	0.052	0.127	0.053	0.005
0.025	0.023	0.028	0.024	0.027	0.025	0.056	0.025	0.008	0.025	0.08	0.025	0.001
0.01	0.01	0.012	0.011	0.008	0.01	0.029	0.01	0.003	0.008	0.04	0.009	0.0002
0.005	0.005	0.006	0.005	0.002	0.005	0.015	0.005	0.0006	0.004	0.026	0.004	0.0000
0.0025	0.0028	0.0036	0.0028	0.0016	0.0026	0.0098	0.0026	0.0000	0.0028	0.016	0.0028	0.0000
0.001	0.0012	0.0018	0.0012	0.0003	0.0016	0.0052	0.0016	0.0000	0.001	0.009	0.001	0.0000
0.0005	0.0002	0.0006	0.0002	0.0000	0.0008	0.0032	0.0008	0.0000	0.0002	0.006	0.0002	0.0000

Table 2. Estimated power for single-SNP association tests in simulated family data. The table shows the proportion of 1,000 simulations where the test statistic exceeded the theoretical quantile at level α. ASKAT is compared to SKAT, FaST-LMM (F-L), and GRAMMAR (GR)

	$h^2_{QTL} = 0.01$											
	$h^2 = 5\%$				$h^2 = 35\%$				$h^2 = 60\%$			
α Level	ASKAT	SKAT	F-L	GR	ASKAT	SKAT	F-L	GR	ASKAT	SKAT	F-L	GR
0.1	0.939	0.927	0.929	0.899	0.897	0.904	0.897	0.597	0.897	0.893	0.897	0.296
0.05	0.881	0.885	0.882	0.850	0.833	0.842	0.833	0.455	0.831	0.835	0.833	0.173
0.025	0.820	0.827	0.821	0.783	0.769	0.797	0.769	0.349	0.750	0.779	0.750	0.097
0.01	0.743	0.764	0.746	0.670	0.682	0.740	0.683	0.226	0.653	0.710	0.654	0.044
0.005	0.658	0.689	0.659	0.594	0.604	0.672	0.605	0.149	0.584	0.644	0.586	0.021

or slightly more powerful than SKAT, and this despite the fact that SKAT's tests are slightly liberal in these simulations.

Simulation Using Multiple SNPs

The previous set of simulations compared the efficiency of our proposed ASKAT with respect to SKAT, FaST-LMM, and GRAMMAR at a single SNP. Since rare variants are typically studied in groups, we conducted several simulations with several causal variants, i.e., with $L > 1$. We simulated 160 families using the same family sizes as before. The trait heritability ranged $h^2 \in \{5\%, 35\%, 60\%\}$, and we assumed $\sigma^2_e = 1$ and $\mu = 3$. For each subject, $L = 5$ rare variants were simulated using the SIMULATE program, with allele frequencies: $(f_1, f_2, f_3, f_4, f_5) = (0.01, 0.01, 0.003, 0.01, 0.001)$.

To calculate the type I error under H_0, we assumed the total QTL-heritability $h^2_{QTL} = 0$. Under the alternative hypothesis, we generated three different parameter settings for groups of causal SNPs. Setting I included a group of five SNPs with only one rare causal variant. Setting II included a group of five SNPs with two rare causal variants, where the two causal variants have opposite effects on the trait Y. Finally, Setting III included a group of five rare variants where all were causal. In each of the three settings, two values were used for the total QTL-heritability: $h^2_{QTL} \in \{0.01, 0.02\}$. In addition, to simulate variants in close proximity, we allowed for linkage disequilibrium (LD) among the variants. The SIMULATE program parameterizes the LD by the square of the correlation, r^2, between two adjacent SNPs. In our simulations, we varied the strength of LD between SNPs using: $r^2 \in \{0, 0.4, 0.8\}$.

Table 3 shows the type 1 error for the region-based tests using ASKAT (using 10,000 simulated data sets) and SKAT (using 1,000 data sets). As was seen in Table 1 for single SNPs, the ASKAT approach adjusting for polygenic variance leads to good control of type I error even when heritability is very high and the α levels are small. This accuracy is reassuring, given that σ^2_g is estimated and hence the distributional convergence is not exact.

To examine power for a multivariant region in families, we have compared the power of the multimarker-based ASKAT test to the minimum P-value from FaST-LMM across the five single-SNP tests. An empirical estimate of the significance level for the FaST-LMM minimum P-value was obtained via permutation. To preserve family relatedness, each permutation randomly assigned one of the two parental haplotypes at each meiosis, consecutively down the pedigree. We assumed no recombination within the five-SNP region. Founders' haplotypes were kept the same as in the original simulated genotype data.

Table 3. Type I error rates for region-based tests using five SNPs in simulated family data. The table shows the proportion of simulations where the test statistic exceeded the expected quantile at level α, using 10,000 simulations for ASKAT and 1,000 simulations for SKAT

	$h^2 = 5\%$		$h^2 = 35\%$		$h^2 = 60\%$	
α Level	ASKAT	SKAT	ASKAT	SKAT	ASKAT	SKAT
0.1	0.112	0.116	0.093	0.252	0.117	0.348
0.05	0.045	0.046	0.053	0.163	0.056	0.248
0.025	0.032	0.034	0.026	0.114	0.031	0.182
0.01	0.016	0.017	0.011	0.060	0.014	0.113
0.005	0.009	0.009	0.007	0.042	0.006	0.078
0.001	0.001	–	0.001	–	0.001	–
0.0005	0.0006	–	0.0004	–	0.0008	–
0.0001	0.0001	–	0.0001	–	0.0000	–

Table 4 shows power in 1,000 simulated data sets, for both multiple-variant ASKAT and for the FaST-LMM minimum P-value approach. In all the three settings, the gains in power for the multiple-variant ASKAT vs. the adjusted FaST-LMM minimum P-value approach can be quite substantial, especially when the locus-specific heritability is higher and in the presence of LD.

Note that the simulation studies conducted here were based on the theoretical kinship matrix calculated directly from the pedigree structure. Even when pedigrees are known, the estimated kinship matrix, based on marker data, may better reflect the true unobserved genomic sharing than the expectation computed from pedigree structure, because of the variability in actual genetic sharing in families [Myles et al., 2009; Zhu et al., 2009]. Thus, we can expect that the model (2) based on the estimated kinship matrix, Φ, may be more powerful than the model using theoretical kinship values.

Analysis of a Large Data Set Containing Related Individuals: The UK Twins Consortium

The Twins UK cohort was collected to study the genetic and environmental etiology of complex traits and diseases

Table 4. Power comparisons for tests of association using five rare variants in simulated family data. The multiple-variant ASKAT test (ASK) is compared to the empirical power of the minimum *P*-value across five single-variant FaST-LMM tests (F-L). The table shows the proportion of 1,000 simulations where the test statistic exceeded the expected quantile at level α. The LD among SNPs is based on the square of the correlation coefficient between two SNPs, r^2. The total trait heritability and the region QTL-heritability are, respectively, h^2 and h^2_{QTL}

		$\alpha = 0.05$						$\alpha = 0.005$					
		$r^2 = 0$		$r^2 = 0.4$		$r^2 = 0.8$		$r^2 = 0$		$r^2 = 0.4$		$r^2 = 0.8$	
h^2_{QTL}	h^2	ASK	F-L	ASK	F-L	ASK	F-L	ASK	F-L	ASK	F-L	ASK	F-L
Setting I													
	0.05	0.43	0.42	0.64	0.38	0.82	0.62	0.21	0.16	0.44	0.28	0.65	0.27
0.01	0.35	0.37	0.41	0.64	0.39	0.82	0.60	0.20	0.16	0.40	0.29	0.61	0.30
	0.60	0.38	0.44	0.62	0.41	0.81	0.64	0.18	0.17	0.39	0.32	0.61	0.34
	0.05	0.60	0.56	0.79	0.63	0.92	0.77	0.42	0.28	0.67	0.34	0.84	0.50
0.02	0.35	0.57	0.57	0.78	0.62	0.92	0.77	0.38	0.29	0.65	0.35	0.83	0.51
	0.60	0.56	0.59	0.79	0.66	0.92	0.80	0.37	0.32	0.62	0.37	0.83	0.55
Setting II													
	0.05	0.37	0.32	0.50	0.46	0.60	0.47	0.16	0.10	0.22	0.12	0.28	0.17
0.01	0.35	0.34	0.30	0.45	0.40	0.54	0.47	0.12	0.09	0.17	0.12	0.23	0.19
	0.60	0.33	0.36	0.45	0.34	0.53	0.49	0.14	0.12	0.19	0.14	0.24	0.18
	0.05	0.58	0.47	0.70	0.52	0.81	0.64	0.35	0.19	0.47	0.23	0.58	0.32
0.02	0.35	0.54	0.47	0.67	0.53	0.77	0.62	0.31	0.19	0.41	0.24	0.52	0.32
	0.60	0.54	0.50	0.65	0.54	0.79	0.65	0.31	0.23	0.42	0.26	0.52	0.34
Setting III													
	0.05	0.34	0.23	0.43	0.27	0.47	0.36	0.18	0.05	0.18	0.07	0.23	0.11
0.01	0.35	0.27	0.22	0.37	0.31	0.44	0.35	0.08	0.05	0.16	0.08	0.19	0.11
	0.60	0.19	0.26	0.41	0.30	0.42	0.38	0.04	0.06	0.16	0.09	0.20	0.12
	0.05	0.92	0.39	0.91	0.47	0.92	0.57	0.70	0.11	0.72	0.18	0.77	0.24
0.02	0.35	0.86	0.40	0.89	0.50	0.90	0.55	0.64	0.11	0.67	0.20	0.75	0.26
	0.60	0.87	0.48	0.86	0.52	0.89	0.60	0.63	0.15	0.67	0.21	0.73	0.29

[Andrew et al., 2001; Richards et al., 2008]. Using 2,197 individuals from this cohort for whom genome-wide genotype data and adiponectin measurements were available, we analyzed the associations between this trait and SNPs on chromosome 3, organized by gene. The adiponectin protein is coded by the ADIPOQ gene on chromosome 3. The complete Twins UK cohort contains both monozygotic and dizygotic twin pairs, but here we analyzed 928 dizygotic twin pairs and 341 individuals with no known family relationships to other cohort members. Associations between SNPs in the ADIPOQ gene and adiponection levels have been demonstrated by several GWAS [Heid et al., 2010; Hivert et al., 2008; Ling et al., 2009; Menzaghi et al., 2007; Richards et al., 2009], and these associations were confirmed recently by a very large meta-analysis for adiponectin, undertaken by the ADIPOGEN consortium [Dastani et al., 2012]. The data analyzed here were part of this meta-analysis, and hence we expect to see association at this locus.

We analyzed 61,227 common SNPs grouped into 1,135 genes on chromosome 3. Although this method has been developed for analysis of rare genetic variation such as would be found from sequencing analysis, it can also analyze SNPs with larger minor allele frequencies. To calculate the average of IBD sharing estimates between all pairs of individuals, we used GenABEL [Aulchenko et al., 2007b] to obtain the estimated kinship matrices for each chromosome, and then we averaged these matrices across the 22 autosomes. Following Dastani et al. [2012], the adiponectin phenotype was first adjusted for BMI, sex, and age, and the residuals were used in the analysis. We then performed gene-based association analysis using our approach as well as SKAT for the 2,197 individuals. Moreover, we created a smaller data set including only unrelated individuals, by removing one subject from each of the 928 dizygotic twin pairs, and we used SKAT to analyze this new data set (unrSKAT). Figure 1 shows the QQ plots for these three methods across the genes on chromosome 3. As expected, inflation of significance can be seen for SKAT, but for ASKAT, the distribution follows the expected null very well. The smaller data set containing only unrelated individuals (unrSKAT) displays control of type I error, however, substantial loss in power can be noticed due to the reduced sample size. The signal at the ADIPOQ gene, which contains 19 SNPs in this study, is clearly visible (ASKAT *P*-value = 1.69×10^{-7}).

In addition, we took the five genes with the strongest signals on chromosome 3, and directly compared the ASKAT multisite test to the empirical significance for the minimum *P*-value across all variants in the gene. We used FaST-LMM to perform the single-SNP analyses, thereby ensuring that we adjusted for the familial relationships, and using permutation to obtain the empirical significance level for the minimum *P*-value. Table 5 demonstrates a small power increase for each of these genes by using the multivariant test rather than a series of univariate tests, and the benefit of ASKAT can be expected to improve if there are more associated variants with small minor allele frequencies where univariate power is negligible.

The data used for analysis included mainly common SNPs. However, for the ADIPOQ gene, there were two markers with minor allele frequency less than 5% among the 19 markers, and so we also used ASKAT to analyze only these two rare

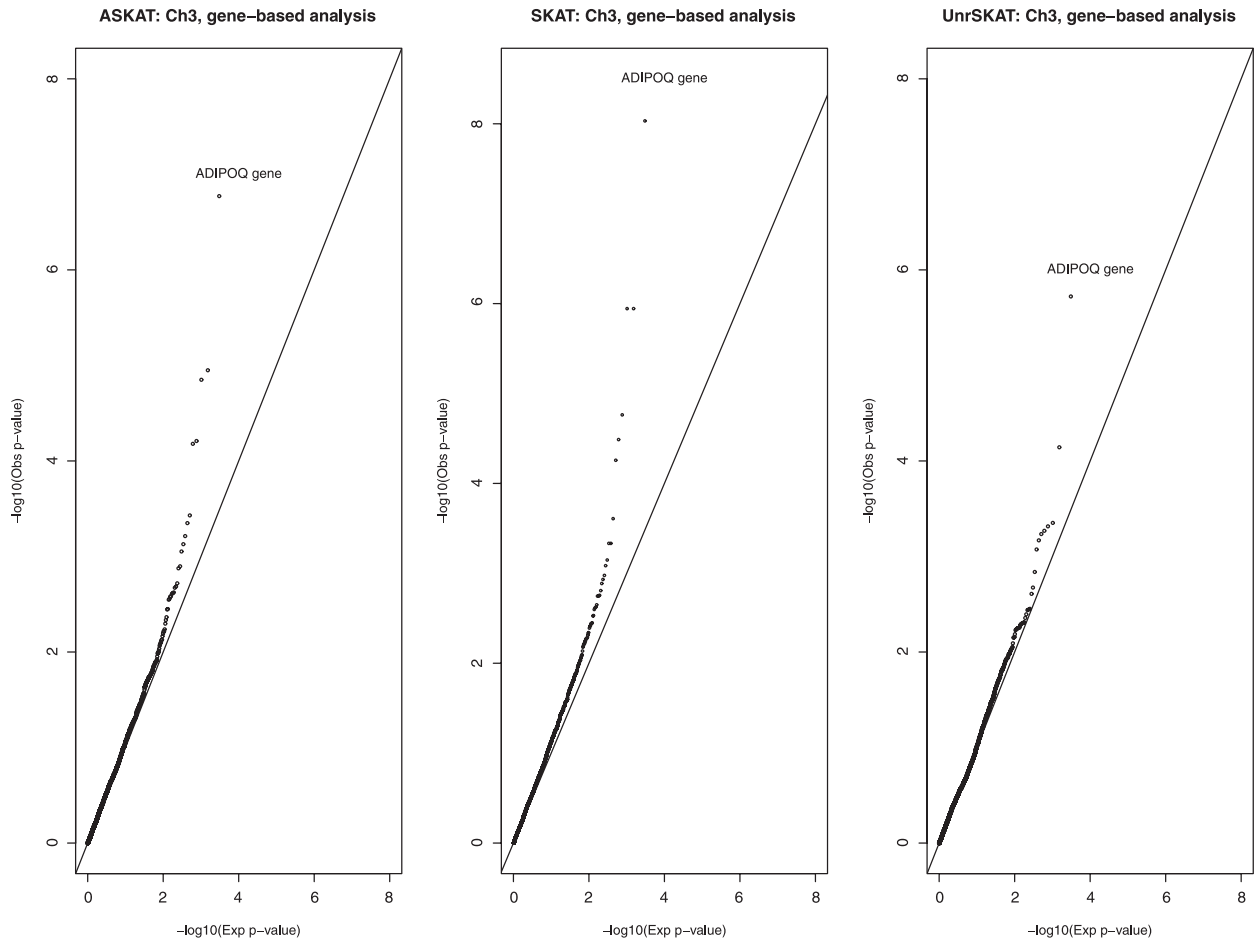

Figure 1. QQ plots comparing *P*-value distributions, obtained from gene-based association tests for adiponectin levels using ASKAT, SKAT, and SKAT in unrelated individuals (unrSKAT), across chromosome 3 data from the Twins UK study. $-\log_{10}(P)$-values are plotted against their expectations.

Table 5. Comparison of the five most significant gene-based ASKAT *P*-values to the corresponding *P*-values obtained by the minimum *P*-value approach. The minimum *P*-value is calculated across the single-variant ASKAT tests of all SNPs within a given gene, and empirical significance for the minimum is obtained by permutation analysis

Gene	No. of SNPs within gene	ASKAT[a]	Minimum *P*-value[a]
ST6GAL1	25	4.18	3.78
PVRL3	31	4.21	3.31
KNG1	22	4.85	4.48
EIF4A2	10	4.95	4.79
ADIPOQ	19	6.77	6.12

[a] Results are shown as $-\log_{10}(P\text{-value})$.

variants at this gene. Our resulting Q statistic was 89,063 (asymptotic *P*-value of 0), whereas the same two SNPs analyzed by SKAT gave a Q statistic of 2,086 (*P*-value 1.1×10^{-11}), indicating very strong evidence for association between these two variants and adiponectin.

Discussion

We have developed and implemented a new method for rare variant association tests that adjusts for related individuals. This method is based on the variance components idea of Wu et al. [2011] and also incorporates calculation efficiencies used in Fast-LMM Lippert et al. [2011]. The method can use either a theoretical or an estimated kinship matrix to adjust for the relatedness between individuals in the sample, and we have used an efficient one-step LMM approach for estimation. As illustrated in our simulation study, ASKAT's type I error is well controlled under a range of different heritabilities, and we have further demonstrated that our approach has good power.

We have shown that ASKAT can be more powerful than a permutation-corrected minimum *P*-value approach, especially if there are multiple signals in the same region of analysis. Since our proposal is an extension of the SKAT test for rare variants to family-based designs or sets of related individuals, all features of SKAT remain available for ASKAT. For example, the ASKAT model can include covariates. Moreover,

ASKAT could be combined with collapsing approaches to improve power (i.e., collapsing some of the rare variants based on some prior knowledge or expectation, and then applying ASKAT to the collapsed variants). Different weights can be incorporated to give more emphasis to particular genotypic patterns. Permutation may not be needed for evaluation of significance levels for ASKAT, since the Davies approximation seems to give acceptable approximations to the P-values in our simulations; however, it should be noted that the Davies approximation assumes that the locus-specific variance is known. In contrast, the minimum P-value approach always requires the use of permutations within families.

The test of association uses a score test statistic, and hence calculation only requires fitting the null model. With the REML variance-component score test, the null model is a one-way ANOVA random model which depends only on variance component parameters that can be estimated efficiently with FaST-LMM. Our algorithm currently takes 10 sec to analyze a single region of 100 rare variants with 1,000 individuals, and 20 min to analyze these 100 variants with 5,000 individuals, on an Intel Xeon 2.4 GHz blade (Supermicro and DELL). Furthermore, we have developed an ASKAT wrapper to enable calculations in parallel. Using this wrapper, we have recently performed gene-based analysis at 7,654 genes on approximately 800 individuals from large families. This analysis, on a multiblade linux system, took 10 h and used 49 h of computation time. However, we acknowledge that the program will run more slowly for larger data sets of related individuals, since the computation time scales with the cube of the kinship matrix dimension. ASKAT software is available at http://www.mcgill.ca/statisticalgenetics/software.

We have illustrated ASKATs performance in gene-based tests of chromosome 3 genotyping data from the Twins UK consortium. We acknowledge that our illustrative example is not ideal, but data sets containing sequencing data from large numbers of related individuals are still uncommon. Although we have been developing the ASKAT method for use in sequencing studies that include individuals from Twins UK, those data are not ready at this time.

Some extensions of this model can be considered. Since ASKAT is an LMM-based approach for family designs, it could be generalized to analyze multivariate phenotypes. Unfortunately, estimating variance components under the null model in the multivariate case is a challenge, since the null model becomes an unbalanced one-way random model. However, explicit, and computationally efficient ANOVA estimates of these variance components in multivariate family-based designs are given by Oualkacha et al. [2012].

LMMs have also been used to control confounding due to population structure and cryptic relatedness [Kang et al., 2010, 2008; Lippert et al., 2011; Yu et al., 2006; Zhang et al., 2010]. In this paper, we have focused on family-based designs, however, it should be possible to extend ASKAT in order to control for population stratification while performing rare variant association tests. Lin and Tang [2011], in their population-based data analysis, controlled for population structure by using the first five principal components, constructed from GWAS SNP data, as covariates in the model (3). Kang et al. [2010] demonstrated that principal component methods can capture some, but not all, of the sample structure, and they further showed that LMMs can provide better control of population structure and hidden relatedness than principal component approaches. Moreover, when dealing with rare variants, the relevance of the principal component analysis approach to population stratification adjustment is not clear [Baye et al., 2011]. Our ASKAT uses LMMs to handle region-based rare variant association for family-based designs and the extension to population-based designs will be a concern of future work. A naive implementation of LMMs in this context may not be adequate, however, since the confounding due to population substructure may be different for rare and common variants [Mathieson and McVean, 2012].

Rare variant analysis may be adversely affected by sequencing errors. For instance, a few minor alleles mistaken as major can involve a loss of power [Daye et al., 2012]. Daye et al. [2012] proposed a powerful qMSAT that incorporates subject-variant sequencing qualities. Extending ASKAT to allow quality weights might lead to additional gains in power.

In this paper, we focused on quantitative traits. For dichotomous traits, the model (2) is a GLMM with two random effects (i.e., two variance components, τ and σ_g^2). The null model is also a generalized linear model with the polygenic random effect P, and the maximum likelihood estimate of the variance component σ_g^2 under H_0 requires the evaluation of difficult integrals [Goldstein, 2003]. For standard GLMMs, global and individual variance component score tests are derived by Lin [1997]. However, taking into account the family structure and using IBD sharing between individuals to control for the family effects in a GLMM requires careful thought and development. This is an ongoing project.

Our approach does not correct for selected ascertainment, i.e., when the studied families are not randomly selected from the population, but instead are sampled based on the phenotypes of some family members. Appropriate ascertainment corrections are challenging [Vieland and Hodge, 1995], and analysis of an ascertained sample without appropriate correction can lead to inflated type I error rates for tests of association. When families are sampled based on disease state, but a secondary (continuous) trait is of interest, recent simulation studies by Schifano et al. [2012] have shown that, when there is only weak association between disease and the secondary trait, there is only a small difference in the null estimates of the variance components σ_g^2 and σ_e^2. Regardless of the sampling mechanism (i.e., random or nonrandom ascertainment), the authors showed that there was little to no inflation in type I error rate. This is in agreement with previous results given by de Andrade and Amos [2000], who showed that ascertainment bias had no effect when assessing major-gene effects in variance-component linkage analysis. When both the genetic markers and secondary trait are associated with disease, appropriate techniques that can account for selected ascertainment are needed, and this is an area for future research.

Since our initial submission of this manuscript, at least two similar methods have been published [Chen et al., 2013; Schifano et al., 2012]. Schifano et al. [2012] proposed the

same LMM equation, although they used the theoretical kinship matrix instead of estimated kinships. Also, their kernel matrix formulation had a general form, whereas we worked with the linear kernel of Wu et al. (2011). Estimation of variance components was performed by maximum likelihood with the R function lmekin(), whereas we used FaST-LMM which implements REML estimation. Despite this, the form of their test statistic is essentially the same as ours. Chen et al. [2013] started from the same LMM and then developed their statistic through maximum likelihood principles rather than REML, hence obtaining a different test statistic. Like us, they used the linear kernel matrix, but they used theoretical kinship estimates. Good control of type 1 error was demonstrated by all three papers. A comparison of speed and performance of these three algorithms would be interesting.

The limitations of all rare variant methods also apply to ASKAT. For example, we do not address the question of how best to choose the region to be analyzed. This decision will include how many variants to analyze together, what threshold for rarity, the potential overlap of regions, and how to use annotation information about the functional consequences of nucleotide changes. These issues apply to all region-based tests of rare genetic variation, and all rare variant methods suffer from a loss of power when the region analyzed contains a mixture of causal and noncausal variants. Nevertheless, our approach provides good power for analysis of rare genetic variation in a sample containing related and unrelated individuals.

Acknowledgments

K.O. was supported by CIHR grant MOP-115110, awarded to C.M.T.G. and A.C. J.B.R. was supported by CIHR, MDEIE, CFI, and CQDM. Z.D. was supported by the CIHR. The Twins UK study acknowledges financial support from the Wellcome Trust, the Department of Health via the National Institute for Health Research (NIHR) comprehensive Biomedical Research Centre award to Guy's & St Thomas' NHS Foundation Trust in partnership with King's College London/Arthritis Research Campaign.

References

Amin N, van Duijn CM, Aulchenko YS. 2007. A genomic background based method for association analysis in related individuals. *PLoS ONE* 2(12):e1274.

Andrew T, Hart DJ, Snieder H, de Lang M, Spector TD, MacGregor AJ. 2001. Are twins and singletons comparable? A study of disease-related and lifestyle characteristics in adult women. *Twin Res* 4(6):464–477.

Astle W, Balding DJ. 2009. Population structure and cryptic relatedness in genetic association studies. *Stat Sci* 24(4):451–471.

Aulchenko YS, de Koning DJ, Haley C. 2007a. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics* 177(1):577–585.

Aulchenko YS, Ripke S, Isaacs A, van Duijn CM. 2007b. GenABEL: an r library for genome-wide association analysis. *Bioinformatics* 23(10):1294–1296.

Baye T, He H, Ding L, Kurowski B, Zhang X, Martin L. 2011. Population structure analysis using rare and common functional variants. *BMC Proc* 5(Suppl 9):S8.

Bonnefond A, Clément N, Fawcett K, Yengo L, Vaillant E, Guillaume JL, Dechaume A, Payne F, Roussel R, Czernichow S and others. 2012. Rare MTNR1B variants impairing melatonin receptor 1b function contribute to type 2 diabetes. *Nat Genet* 44(3):297–301.

Chen H, Meigs J, Dupuis J. 2013. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37(2):196–204.

Dastani Z, Hivert MF, Timpson N, Perry JRB, Yuan X, Scott RA, Henneman P, Heid IM, Kizer JR, Lyytikäinin LP and others. 2012. Novel loci for adiponectin levels and their influence on type 2 diabetes and metabolic traits: a multi-ethnic meta-analysis of 45,891 individuals. *PLoS Genet* 8(3):e1002607.

Davies RB. 1980. Algorithm as 155: the distribution of a linear combination of ξ^2 random variables. *J R Stat Soc Ser C* 29:323–333.

Daye Z, Li H, Wei Z. 2012. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res* 40(8):e60.

de Andrade M, Amos CI. 2000. Ascertainment issues in variance components models. *Genet Epidemiol* 19:333–344.

Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB. 2010. Rare variants create synthetic genome-wide associations. *PLoS Biol* 8(1):e1000294.

Duchesne P, Lafaye De Micheaux P. 2010. Computing the distribution of quadratic forms: further comparisons between the Liu-TangZhang approximation and exact methods. *Comput Stat Data An* 54(4):858–862.

Eichler EE, Flint J, Gibson G, Kong A, Leal SM, Moore JH, Nadeau JH. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet* 11(6):446–450.

Goldstein H. 2003. *Multilevel Statistical Models* (3rd ed.). London: Arnold.

Hayes BJ, Visscher PM, Goddard ME. 2009. Increased accuracy of artificial selection by using the realized relationship matrix. *Genet Res* 91: 47–60.

Heid IM, Henneman P, Hicks A, Coassin S, Winkler T, Aulchenko YS, Fuchsberger C, Song K, Hivert MF, Waterworth DM and others. 2010. Clear detection of adipoq locus as the major gene for plasma adiponectin: results of genome-wide association analyses including 4659 european individuals. *Atherosclerosis* 208:412–420.

Hivert MF, Manning AK, McAteer JB, Florez JC, Dupuis J, Fox CS, O'Donnell CJ, Cupples LA, Meigs JB. 2008. Common variants in the adiponectin gene (adipoq) associated with plasma adiponectin levels, type 2 diabetes, and diabetes-related quantitative traits: the Framingham offspring study. *Diabetes* 57:3353–3359.

Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S, Freimer NB, Sabatti C, Eskin E. 2010. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42(4):348–354.

Kang HM, Zaitlen NA, Wade CM, Kirby A, Heckerman D, Daly MJ, Eskin E. 2008. Efficient control of population structure in model organism association mapping. *Genetics* 178(3):1709–1723.

Li B, Leal SM. 2008. Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.

Lin DY, Tang ZZ. 2011. A general framework for detecting disease associations with rare variants in sequencing studies. *Am J Hum Genet* 89(3):354–367.

Lin X. 1997. Variance component testing in generalised linear models with random effects. *Biometrika* 84(2):309–326.

Ling H, Waterworth DM, Stirnadel HA, Pollin TI, Barter PJ, Kesäniemi YA, Mahley RW, McPherson R, Waeber G, Bersot TP and others. 2009. Genome-wide linkage and association analyses to identify genes influencing adiponectin levels: the GEMS study. *Obesity (Silver Spring)* 17(4):737–744.

Lippert C, Listgarten J, Liu Y, Kadie CM, Davidson RI, Heckerman D. 2011. Fast linear mixed models for genome-wide association studies. *Nat Methods* 8(10):833–835.

Liu H, Tang Y, Zhang HH. 2009. A new chi-square approximation to the distribution of non-negative definite quadratic forms in non-central normal variables. *Comput Stat Data An* 53(4):853–856.

Madsen BE, Browning SR. 2009. A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5(2):1–11.

Mathieson T, McVean G. 2012. Differential confounding of rare and common variants in spatially structured populations. *Nat Genet* 44(3):243–246.

Menzaghi C, Trischitta V, Doria A. 2007. Genetic influences of adiponectin on insulin resistance, type 2 diabetes, and cardiovascular disease. *Diabetes* 56: 1198–1209.

Myles S., Peiffera J, Browna PJ, Ersoza ES, Zhanga Z, Costicha DE, Bucklera ES. 2009. Association mapping: critical considerations shift from genotyping to experimental design. *Plant Cell* 21:2194–2202.

Oualkacha K, Labbe A, Ciampi A, Roy MA, Maziade M. 2012. Principal components of heritability for high dimension quantitative traits and general pedigrees. *Stat Appl Genet Mol Biol* 11(2): Article 4, doi: 10.2202/1544-6115.1711

Quaas RL. 1992. *REML Notebook. Mimeo.* Ithaca, NY: Cornell University.

Rao CR, Kleffe J. 1988. *Estimation of Variance Components and Applications.* North-Holland series in statistics and probability. Amsterdam: North-Holland.

Richards JB, Rivadeneira F, Inouye M, Pastinen TM, Soranzo N, Wilson SG, Andrew T, Falchi M, Gwilliam R, Ahmadi KR, Valdes AM, Arp P, Whittaker P, Verlaan DJ, Jhamai M and others. 2008. Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet* 371(9623):1505–1512.

Richards JB, Waterworth D, O'Rahilly S, Hivert MF, Loos RJ and others. 2009. A genome-wide association study reveals variants in arl15 that influence adiponectin levels. *PLoS Genet* 5:e1000768.

Schifano ED, Epstein MP, Bielak LF, Jhun MA, Kardia SLR, Peyser PA, Lin X. 2012. SNP set association analysis for familial data. *Genet Epidemiol*, doi: 10.1002/gepi.21676

Searle SR. 1979. Notes on variance component estimation. A detailed account of maximum likelihood and kindred methodology. Technical Report BU-673-M, Biometrics Unit, Cornell University, Ithaca, New York.

Terwilliger JD, Speer M, Ott J. 1993. Chromosome-based method for rapid computer simulation in human genetic linkage analysis. *Genet Epidemiol* 10: 217–224.

Vieland V, Hodge S. 1995. Inherent intractability of the ascertainment problem for pedigree data: a general likelihood framework. *Am J Hum Genet* 56(1):33–43.

Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X. 2011. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89(1):82–93.

Yi N, Zhi D. 2011. Bayesian analysis of rare variants in genetic association studies. *Genet Epidemiol* 35(1):57–69.

Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB and others. 2006. A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nat Genet* 38(2):203–208.

Zhang Z, Ersoz E, Lai CQ, Todhunter RJ, Tiwari HK, Gore MA, Bradbury PJ, Yu J, Arnett DK, Ordovas JM and others. 2010. Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42(4):355–360.

Zhu L, Zhang Z, Friedenberg S, Jung SW, Phavaphutanon J, Vernier-Singer M, Corey E, Mateescu R, Dykes N, Sandler J and others. 2009. The long (and winding) road to gene discovery for canine hip dysplasia. *Vet J* 181(2):97–110.