Briefings in Bioinformatics, 18(6), 2017, 954–961

doi: 10.1093/bib/bbw083
Advance Access Publication Date: 26 September 2016
Paper

Rare variant association test in family-based sequencing studies

Xuefeng Wang, Zhenyu Zhang, Nathan Morris, Tianxi Cai, Seunggeun Lee, Chaolong Wang, Timothy W. Yu, Christopher A. Walsh and Xihong Lin

Corresponding authors: Xuefeng Wang, Department of Biostatistics and Bioinformatics, H. Lee Moffitt Cancer Center & Research Institute, Tampa, FL 33612, USA. Tel.: +1-617-432-2914; Fax: +1-617-432-5619; E-mail: xuefeng.wang@moffitt.org; Xihong Lin, Department of Biostatistics, Harvard School of Public Health, Boston, MA 02511, USA; E-mail: xlin@hsph.harvard.edu

Abstract

The objective of this article is to introduce valid and robust methods for the analysis of rare variants for family-based exome chips, whole-exome sequencing or whole-genome sequencing data. Family-based designs provide unique opportunities to detect genetic variants that complement studies of unrelated individuals. Currently, limited methods and software tools have been developed to assist family-based association studies with rare variants, especially for analyzing binary traits. In this article, we address this gap by extending existing burden and kernel-based gene set association tests for population data to related samples, with a particular emphasis on binary phenotypes. The proposed approach blends the strengths of kernel machine methods and generalized estimating equations. Importantly, the efficient generalized kernel score test can be applied as a mega-analysis framework to combine studies with different designs. We illustrate the application of the proposed method using data from an exome sequencing study of autism. Methods discussed in this article are implemented in an R package 'gskat', which is available on CRAN and GitHub.

Key words: association test; family; GEE; Kernel machine; mega analysis; score test; perturbation; rare variants; sequencing

Introduction

The high-throughput next-generation sequencing (NGS) technology has revolutionized genetic research, not only by dramatically decreasing sequencing costs but also by increasing the scale of genomic sequencing. The NGS advances open up the entire spectrum of genomic variation for the genetic analysis of complex diseases and traits. In genome-wide association studies (GWAS) era, we typically focus on common causal variants with moderate effects, or more precisely, the single-nucleotide polymorphisms (SNPs) in linkage disequilibrium with the causal ones. Given the whole-genome sequencing, the interest has shifted toward identifying rare variants that are associated with diseases. Rare variants are potential contributors of the 'missing heritability' that was widely debated after the 1st wave of GWAS. However, the costs of NGS remain high for large-scale whole-genome sequencing projects and thus hinder a widespread application. As a result, it is more practical to consider cost-efficient designs by targeting specific genomic regions (e.g.

Xuefeng Wang is currently an Assistant Member at Moffitt Cancer Center, Adjunct Assistant Professor at Stony Brook University and a visiting faculty member at Yale School of Medicine. His research focuses on statistical genomics and bioinformatics.
Zhenyu Zhang is a PhD student in the Department of Applied Math & Statistics at Stony Brook University.
Nathan Morris is an assistant professor in the Department of Epidemiology and Biostatistics at Case Western Reserve University.
Tianxi Cai is a professor in Biostatistics at Harvard T.H. Chan School of Public Health.
Seunggeun Lee is an assistant professor in Biostatistics at University of Michigan, Ann Arbor.
Chaolong Wang is a Principal Investigator at the Genome Institute of Singapore. He is also an Adjunct Assistant Professor at the Centre for Computational Biology, Duke-NUS Medical School.
Timothy W. Yu is an assistant professor in Pediatrics at Children's Hospital Boston.
Christopher A. Walsh is a professor in the Division of Genetics at Children's Hospital Boston.
Xihong Lin is Henry Pickering Walcott professor of Biostatistics and Chair of Department of Biostatistics at Harvard T.H. Chan School of Public Health. She also serves as the coordinating director of the Program of Quantitative Genomics (PQG) of Harvard School of Public Health.
Submitted: 6 May 2016; **Received (in revised form):** 23 August 2016

exome sequencing) and specific samples. Current strategies include extreme-phenotype sampling, family-based sampling as well as the more involved two-stage designs. In this study, we restrict our attention to family-based sequencing studies.

The family-based design is increasingly popular in sequencing studies because of its significant advantages in detecting rare variant associations. Family-based designs also provide unique opportunities to identify genetic variants that complement studies of unrelated individuals. Mendelian inheritance information in family data can also be used in genotyping error checks and can improve imputation of variants in unsequenced family samples. For a general review of family-based designs for sequencing studies, see [1]. On the other hand, applying family-based sequencing analysis may be the only option available for some cases. It is not unusual that family samples are easier to collect compared with unrelated samples for some projects. Also, many current genome sequencing studies tend to focus on individuals ascertained from families based on previous linkage analyses. There are, however, some limitations and challenges, which must be considered in the use of family-based studies. With common variants, an association analysis using a family-based design is less powerful than an unrelated case-control design with the same sample size, i.e. the statistical efficiency per individual is typically lower unless it includes the phenotypes of the ungenotyped relatives [2, 3]. It is also known that, in sequencing studies, family-based designs can be less efficient for discovering novel variants than population-based designs using the same sample size.

There are two main approaches to handling the correlated or clustered structure in the family data. The first is to model the dependence structure explicitly by specifying a random effect in the framework of the linear mixed model (LMM) and generalized LMM for binary traits (GLMM) [4–8]. The second is to fit a marginal model with generalized estimating equations (GEE) [9]. The basic idea of GEE is to replace the covariance matrix in the GLMM with a 'working' covariance matrix that reflects the cluster dependencies. The resulted estimator and testing are more robust to model misspecification than GLMM. The proposed approach in this article is based on our previous work describing a common variant association test [10], which blends the strengths of kernel machine (KM) [4, 11, 12] methods and GEE. The KM-GEE score test retains the quadratic statistics form as in the typical KM score test but has a more complicated null distribution. Our work was mainly motivated by the challenges in the current mental disease projects where only dichotomous traits are available for analysis. Examples include the autism data used in this article, as well as alcohol, cocaine and other dependencies. Although KM methods such as sequence kernel association test (SKAT) have been extended to incorporate family structure [5, 13], none of them is readily applicable to binary traits. It is thus the aim of the current article to present a valid analysis workflow to address this gap, where both dichotomous and continuous traits will be allowed in a KM testing framework. The GEE techniques offer several advantages over the GLMM-based methods for handling family data, including computational rapidity and numerical reliability. A significant advantage of GEE framework is that it does not require a fully parameterized dependence structure and does not require assumptions regarding the joint family distribution. Therefore, GEE framework is more robust and flexible for large-scale studies with a complex data structure.

Importantly, this article presents a framework, which can be readily applied to mega-analysis, in which raw (genotyping) data from multiple sources and mixed designs are pooled and processed in a uniform processing and analysis pipeline. The common practice in the current large-scale multiple-center GWAS is to perform analyses separately within each data set or different population groups. The results are then combined using meta-analysis techniques. For single-site association tests, the meta-analysis can achieve the same statistical efficiency and should yield equivalent results as compared with the mega-analysis [14] that aggregates individuals from all groups. However, mega-analysis provides significant advantages in practice—which allows for more consistent data processing and quality control, as well as a more sensible choice of control subjects. More importantly, it was recently demonstrated that, by theory and simulations, mega-analysis could provide better power than meta-analysis in gene-based association tests [15]. Mega-analysis can assess the pattern of signals at the variant level across sites and then combine them at the gene level across studies, whereas meta-analysis can only pool information at the gene level. This feature makes mega-analysis appealing for sequencing data, where rare variants often need to be collapsed on a gene or region level and tested based on a burden or SKAT method. The mega-analysis strategy will provide not only a more powerful tool but also a valuable complement to the results from the initial meta-analyses, and is likely to shed new insight into the genetic mechanisms of complex diseases. The remainder of this report is organized as follows. We first describe the proposed model and the KM test in the GEE framework for sequencing studies. We then present simulation settings and results to evaluate the finite-sample performance of the proposed method and compare the proposed approach with the single-SNP-based minimum P-value analysis. Finally, we apply the proposed method to a whole-exome sequencing study to identify inherited causes of autism.

GEE kernel association test for family-based sequencing data

We assume there are n families, and family i has m_i members. Suppose a single-nucleotide variant (SNV) set contains p variants. Let $\mathbf{y}_{ij}$ denote the continuous or discrete phenotype for the j th individual in the i th family; $\mathbf{X}_{ij}$ denote a vector of covariates such as sex, age, environmental factors and the intercept; $\mathbf{Z}_{ij}$ denote a $p_{\times 1}$ genotype vector for the SNVs in the set, coded 0, 1, 2, reflecting the number of copies of the minor allele. We model the phenotypic value using a marginal generalized linear model:

$$g(\mathrm{E}(\mathbf{y}_{ij}|\mathbf{X}_{ij},\mathbf{Z}_{ij})) = \mathbf{X}_{ij}^{\mathrm{T}}\boldsymbol{\alpha} + \mathbf{Z}_{ij}^{\mathrm{T}}\boldsymbol{\beta} \quad (1)$$

where $g(.)$ is the corresponding link function, which can be the identify function for a continuous trait and the logistic function for a dichotomous trait. $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are regression coefficient vectors for the covariates and genotypes, respectively. To test whether there is an overall genetic effect of the SNV set, i.e., $H_0 : \boldsymbol{\beta} = 0$, we assume the individual components of the regression coefficient $\beta_j (j = 1, \ldots, p)$ follow an arbitrary distribution with mean 0 and common variance τ. We can test $H_0 : \tau = 0$ using a GEE kernel association test as proposed in Wang's work [10]:

$$T = \tilde{\mathbf{U}}^{\mathrm{T}} W W \tilde{U}, \quad (2)$$

where $\tilde{\mathbf{U}}$ is the GEE score estimated under the null, i.e. $\tilde{\mathbf{U}} = \sum_{i=1}^{n} \mathbf{Z}_i^T \mathbf{D}_i \mathbf{V}_i^{-1} (\mathbf{y}_i - \tilde{\mu}_i)$ and $\tilde{\mu}_i = g(\mathbf{X}_{ij}^{\mathrm{T}} \boldsymbol{\alpha}_0)$. $\mathbf{W} = diag(\omega_1, \ldots, \omega_p)$ are variant weights that are based on external functional information or the minor allele frequency (MAF) of a variant, i.e.

based on a beta density function Beta(MAF;a1 = 1, a2 = 25) [11]. The GEE-SKAT statistics retains the quadratic form in the original KM test. For continuous traits, it will be equivalent to the KM score test for family data based on the LMM framework [4] when the kinship working correlation is applied. Similarly, it can be shown that the null distribution of the score statistic T can be approximated asymptotically by a weighted sum of χ^2 distributions: $\sum_{k=1}^{p} \lambda_k \chi^2_{k,1}$, where $\chi^2_{k,1}$ are independent χ^2 distributed random variables with one degree of freedom and $(\lambda_1, \ldots \lambda_p)$ are eigenvalues of the matrix defined in the Supplementary Material. It is important to distinguish the null distribution of the quadratic score statistics (i.e. the linear combination of χ^2 random variables) from the mixture of χ^2 distributions with different degrees of freedom, as often used in one-sided variant component tests and constrained hypothesis tests.

Instead of using asymptotic approximations, an empirically adjusted P-value can be calculated approximately by matching the calculated small-sample moments. Instead of matching the first two moments as in Satterthwaite's method or the first three moments as proposed in Liu's method, we match the mean, variance and kurtosis (the fourth moment about the mean) to improve the tail fit. Therefore, the small-sample P-valued is calculated by the following formula, which has also been used in [16]:

$$1 - F((T - \widehat{\mu}_T)\sqrt{2df}/\sqrt{\widehat{\nu}_T} + df|\chi^2_{df}) \quad (3)$$

where $F(.|\chi^2_{df})$ is the cumulative distribution function (CDF) of a χ^2 distribution χ^2_{df}, with the modified degrees of freedom $df = 12/\widehat{\gamma}$. $\widehat{\mu}_T$,$\widehat{\nu}_T$ and $\widehat{\gamma}$ are the estimated sample mean, variance and kurtosis of the statistic under the null, respectively. These moments can be calculated by resampling methods. Alternatively, one can perform a direct comparison of the statistics with the resampled null distribution, in which much more permutations will be needed to approximate the extreme tail probability of the null distribution. As shown in the previous work, the small-sample mean is close to the asymptotic mean of T; however, the small-sample variance and kurtosis usually differ from the ones computed from the asymptotic distribution. When there are no covariates and all families have the same pedigree structure, a simple permutation method can be used as described in the previous work [17]. For more realistic settings where there are covariates and different pedigree structures, we propose a new resampling method. In this new perturbation procedure, we form the blockwise perturbed statistic $\tilde{U}_b$ through assigning a random weight to each family, i.e. $\tilde{U}_b = \sum_{i=1}^{n} \mathbf{Z}_i^T \mathbf{D}_i \mathbf{V}_i^{-1}(\mathbf{y}_i - \tilde{\mu}_i) r_i$, where r_i is a random variable generated from a Gaussian or Rademacher distribution [18]. The main idea of this perturbation procedure is thus similar to the block bootstrap [19], which preserves the correlation structure by keeping all the individuals who belong to the same family together. Suppose a total of B samples of the perturbed score are generated, the sample kurtosis can be calculated as described in Supplementary Material. The adjusted P-value can then be calculated using equation (3). This procedure is computationally much more efficient than the traditional resampling method, as it avoids repeated estimation of randomly perturbed data, and the tail of the distribution can be estimated well based on a sampling of realistic size. We recommend B = 10 000 perturbation samples for a genome-wide scan and B = 100 000 on top hits for finer approximation.

Optimal test and perturbation procedures

The proposed test can be easily generalized to the optimal test for maximizing the power over a broader range of scenarios. Previous studies [16, 17] suggest that the proposed kernel association test can be less powerful than burden tests—when the target region has a high proportion of casual variants with the effects in the same direction. The unified test can be based on the following statistic:

$$T_\rho = \tilde{\mathbf{U}}^T \mathbf{W} \mathbf{R}_\rho \mathbf{W} \tilde{U}, \quad (4)$$

where $\mathbf{R}_\rho = (1-\rho)\mathbf{I} + \rho 11'$, $0 \le \rho \le 1$. 1 is the column vector of one here, such that $11'$ is a matrix of ones everywhere. It can be shown that T_ρ is equivalent to a weighted sum of GEE burden and kernel score statistics, i.e. $T_\rho = (1-\rho)T_{\text{GEE-kernel}} + \rho T_{\text{GEE-burden}}$. Therefore, the GEE-based kernel and the burden test are special cases of the optimal test when ρ equals 0 or 1. Under H_0, the parameter ρ disappears and thus is not identifiable. For a fixed ρ, the null distribution of T_ρ can be approximated by the moment-matching procedure described above.

Under the null hypothesis and for a fixed ρ, T_ρ is asymptotically distributed as $\sum_{k=1}^{p} \lambda_k \chi^2_{k,1}$, where $\chi^2_{k,1}$ are independent χ^2_1 random variables and $(\lambda_1, \ldots, \lambda_p)$ are eigenvalues of $\mathbf{B}^{1/2}\mathbf{C}^T\mathbf{W}\mathbf{R}_\rho\mathbf{W}\mathbf{C}\mathbf{B}^{1/2}$ as defined in [10]. The optimal unified test can be constructed based on $T_{\text{optimal}} = \inf_{0 \le \rho \le 1} p_\rho$, where p_ρ is the P-value from T_ρ. $T_{optimal}$ can be obtained by a grid search through a finite number of ρ: $0 = \rho_1 < \ldots. < \rho_l = 1$, and choose the value of ρ that yields the smallest value, i.e. $T_{\text{optimal}} = \min(p_{\rho_1}, \ldots, p_{\rho_l})$. To get its P-value, we propose a perturbation procedure as follows:

1. Set a grid of equally spaced points, we use 11 points here $\rho_1 = 0, \rho_2 = 0.1, \ldots, \rho_{11} = 1$.
2. Compute $T_{\rho_1}, \ldots, T_{\rho_l}$, and calculate P-values based on a modified moment-matching approximation based on the perturbed scores as described in our previous paper [10].
3. Find the minimum P-value: $p_{\min} = \min\{p_{\rho 1}, \ldots, p_{\rho 11}\}$.
4. Calculate the perturbed score $\tilde{U}_b = \sum_{i=1}^{n} \mathbf{Z}_i^T \mathbf{D}_i \mathbf{V}_i^{-1}(\mathbf{y}_i - \tilde{\mu}_i) r_i$, where r_i is a random variable generated from the standard normal distribution. Calculate the perturbed statistic T_b for each grid point of ρ, and compute the corresponding P-value by using the modified moment-matching method (based on the same perturbed scores used in step 2). Set $\widehat{p}^{(b)}_{\min} = \min\{p^{(b)}_{\rho 1}, \ldots, p^{(b)}_{\rho 11}\}$.
5. Repeat step (4) B times to obtain $\widehat{p}^{(1)}_{\min}, \ldots, \widehat{p}^{(B)}_{\min}$.
6. The final P-value for $T_{optimal}$ is calculated by comparing $p_{\min}$ with its empirical null distributions $\widehat{p}^{(b)}_{\min}$, i.e. $p = B^{-1} \sum_{b=1}^{B} I(\widehat{p}^{(b)}_{\min} \le p_{\min})$.

Population structure adjustment

Because the proposed GEE-KM model models test an association unconditional on parental information, it is by nature not robust to potential population stratification. To control for the stratification, the eigenvectors from the principal component analysis for population should be adjusted as covariates in the model. For family-based design, one solution is to use principal component estimates derived from the unrelated samples, e.g. all the parents [20]. Alternatively, we may apply principal component analysis based on outside ancestry informative populations as reference panels such as HapMap and HGDP [21]. A step-by-step protocol is well summarized in [22], which is originally proposed to identify individuals of divergent ancestry.

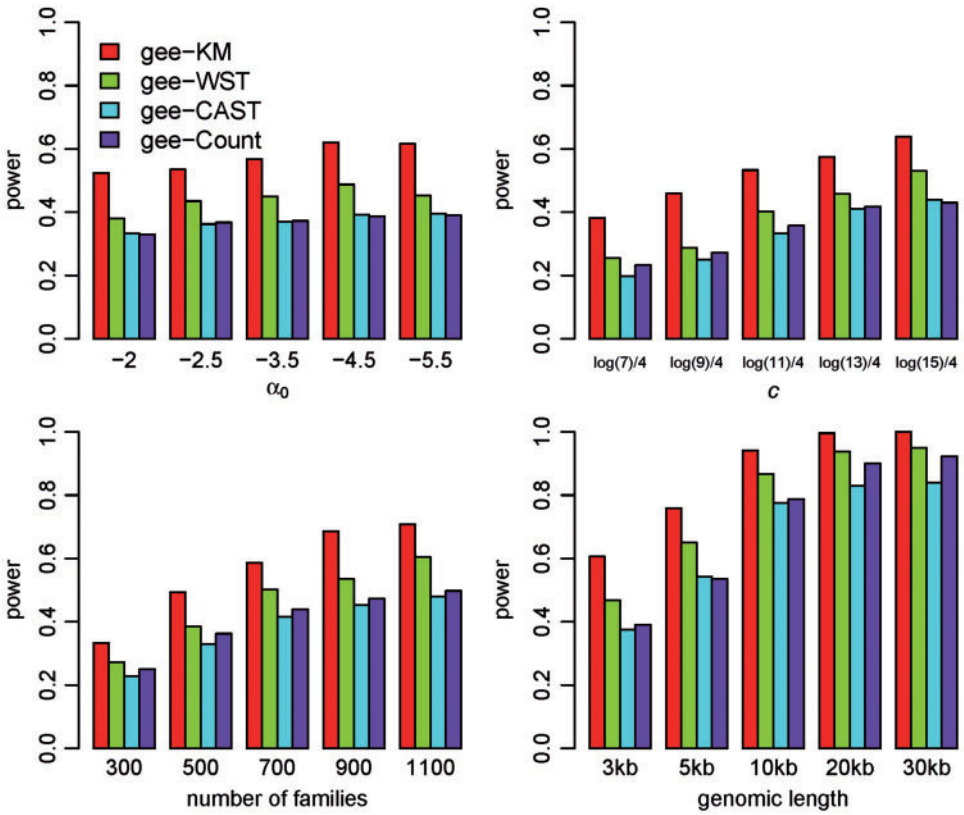

Figure 1. PCA ancestry plot of the Autism WES data. Black circles indicate the ASD samples (including Middle East and US samples), and colored circles indicate samples from the HGDP populations including groups from Africa, Europe, Middle East, Asia and America. The principal component projection plots in C and D illustrate the genetic mixture component of the ASD cohort. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

The protocol includes key steps and scripts to run from merging PLINK (http://pngu.mgh.harvard.edu/purcell/plink/) format files with HapMap data from four ethnic populations, extracting the pruned SNP data, to conducting Principal Component Analysis (PCA) on the merged data. The projected eigenvectors of related members F_t can then be calculated by $F_t = ZP_r$, where P_r is the genotype loading matrix from the training samples, i.e. the merged reference and unrelated individuals.

In the analysis of autism data described below, we used the HGDP data as the reference panel, which includes 938 unrelated individuals from 53 worldwide populations and shares 13 688 autosomal SNPs with our autism data. Based on this shared set of SNPs, we first constructed a reference map using PCA on the HGDP data. We then placed one autism sample into the reference map by applying PCA on the combined data of the autism sample and the HGDP, followed by a Procrustes analysis [23] to transform the new PCA map to match the reference map. This procedure was repeated with the other autism samples, one at a time until all samples were placed into the reference PCA map (Figure 1). For the detailed PCA analysis protocol, one is referred to [24].

Simulations

We carried out a series of simulation studies to evaluate the performance of the proposed method. All simulations were based on a genomic region of length 1 Mb based on a coalescent model using the software COSI [25]. A total of 10 000 haplotypes were generated in this genomic region. We first generate the genotypes of each pedigree founder by randomly sampling with replacement two haplotypes from the 10 000 haplotype pool. The genotypes of each offspring were generated using allele-dropping algorithm, i.e. the parental haplotypes are transmitted to an offspring with equal chance. We randomly picked a subregion of 3 kb as our test region in each simulation.

The phenotype mean for each individual was simulated from: $\text{logit}(\mu_{ij}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \beta_1 g_1 + \cdots + \beta_s g_s$, where the intercept α_0 was chosen to set the prevalence to 0.01, X_1 and X_2 are two covariates and $(g_1, \ldots, g_s)$ are the genotypes of the causal variants selected. Assuming rarer variants have larger effects, the magnitude of β_k was set to be $c|\log_{10}MAF_k|$. Unless otherwise stated, we set $c = 1n(13/4)$ and 10% of the rare variants are causal, which gives maximum odds ratio = 13 for variants with MAF $= 10^{-4}$. There are mainly two strategies to simulate correlated binary phenotypes. The 1st strategy is to simulate correlated binary random variables directly given the mean vector and the correlation matrix [10]. We implemented the described method in Park *et al.* [26] and included it in our gskat package. A caveat to this strategy is that the means and correlations for the binary variables must satisfy a set of constraints. Thus, some correlation matrices formed directly based on the kinship matrix (where the correlations between parents often set at zero) may not be working. The 2nd strategy, which is more realistic, is to simulate the correlation structure implicitly through introducing a latent residual variable with imposed correlation structure: (1) generate matrix normal variable data matrix; (2) induce block correlation by multiplying the Cholesky decomposition of a correlation matrix by the original data matrix generated in the previous step, or directly generate correlated multivariate normal variables using existing packages; (3) get residual random variable by taking the logit

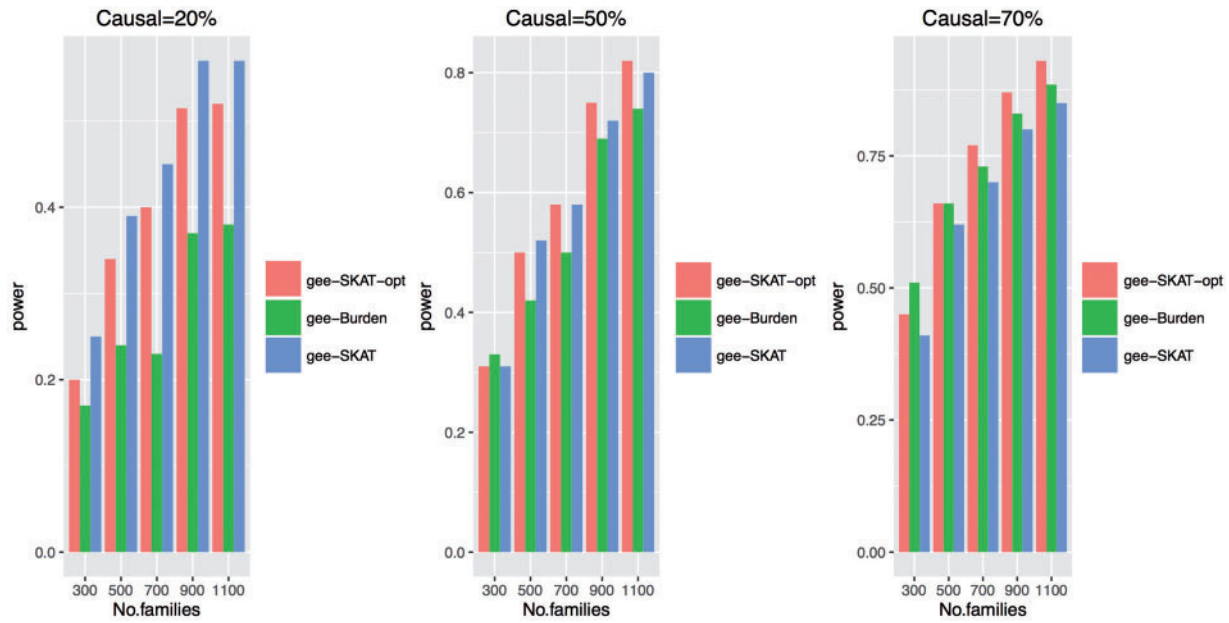

Figure 2. Power comparisons of the GEE-KM and different SNP set tests using simulated data. This shows that power estimates of GEE-KM, GEE-WST, GEE-CAST and GEE-Count. The plots consider different simulation settings by varying prevalence, effect size, sample size factor and genomic length of the gene region. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

transformation on the CDF of the normal variables. Note that, although the scheme of generating binary variables obtained from a truncated liability model is more straightforward, it is not recommended in the simulations for testing models with logistic parametrization. For each test, we simulate 700 nuclear families ascertained with at least one disease individual in each family. Each simulation was replicated 0.5×10^7 times for type I error evaluation. Power evaluation was based on 1000 replicates, assuming the type I error rate is 0.05. We compare the GEE-based burden test against the original Kernel test based on the simulated data.

Type I error and statistical power

The type I error control of the proposed method is tested based on the data sets simulated under the null, where the mean vector of the null logistic regression model is computed as $\text{logit}(\mu_{ij}) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2$. We use two resampling approaches to obtain empirical *P*-values for each test: permutation and perturbation. The permutation scheme is valid here because in this simulation, we assume all families have the same pedigree structure. The results are presented in the quantile-quantile plots (Q-Q) in Supplementary Figure S1. It indicates that both tests are valid, in which the type I error can be well controlled across a wide range of *P*-value thresholds. We do observe that Satterthwaite's method provides a better approximation for large *P*-values when compared with Liu's method and our method. However, for inference in GWAS, the accuracy in the tail is apparently more important. We compare the power of the kernel association test with three types of burden tests in simulations under a broad range of scenarios (by varying the values of α_0, c , the sample size and the genomic length of the tested region). The results, presented in Figure 2, show that the kernel test outperforms the burden tests across all simulation configurations considered. It also shows that the weighted sum test (WST) method tends to outperform the other burden tests. As expected, the power of all tests increases with sample size and effect size. It is found that the kernel test is substantially more powerful than burden tests at intermediate levels of the sample and effect size parameters. It is known that burden test tends to outperform SKAT-type test when a large proportion of variants in a SNP set are truly causal and influence the phenotype in the same direction and with similar magnitudes [16, 17]. The optimal test introduced in this article is more robust to the effect size structure than the previously developed GEE-SKAT [10]. As shown in Figure 3, the optimal tests adaptively choose the weight between burden and SKAT based on data, and achieve a power either larger or close to the most powerful test under all scenarios. When there is a high proportion (50% and 70%) of casual variants in a region or gene, the optimal test and burden test are more powerful than the GEE-KM test.

Application to whole-exome sequencing data

We illustrate the application of the proposed method using data from an ongoing exome sequencing study to identify genes associated with autism spectrum disorders (ASDs). Whole-exome sequencing was performed on whole blood DNA samples from a total of 831 individuals selected from a larger ASD cohort. Study populations consisted both of nuclear families and case-control sample sets from populations with ethnic groups in the Middle East and the United States. Most of the families were small nuclear families (including 192 trios) with at least one affected offspring. After variant calling and quality control procedures, a total of 534 692 SNVs were identified across the tested samples, about 88% of which have an MAF <0.05 and about 81% have an MAF <0.01.

To adjust for population stratification, we need to apply principle component analysis based on outside population reference panels. Note that there are many consanguineous families included in this study. Therefore, we are unable to use principle component estimates derived from the parents (assuming they are unrelated) selected from each family and then project the rest of the family members (11). We have outlined the key steps above.

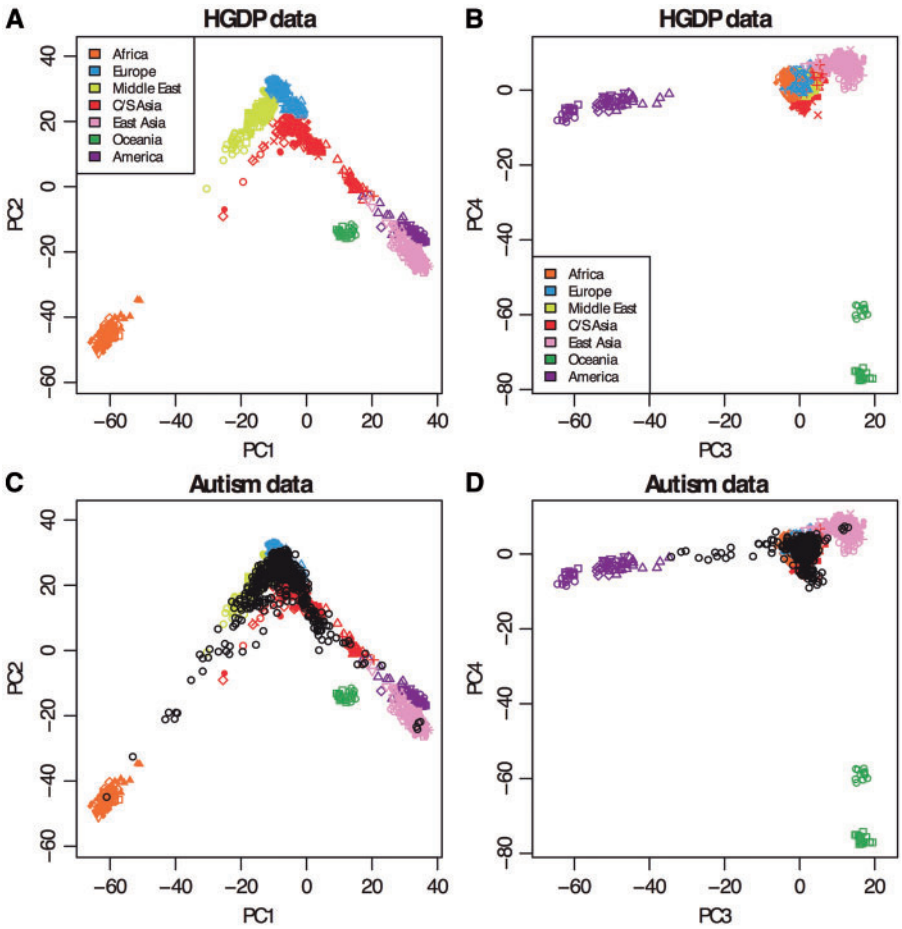

Figure 3. Power comparisons of the optimal method, burden and GEE-SKAT. From left to right, the plots consider settings in which 20% of rare variants were causal, 50% of rare variants were causal and 70% of rare variants were causal, respectively. The effect size factor c is scaled down with larger percentages of casual variants. Total family sizes considered were 300, 500, 700, 900 and 1100, respectively. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

A total of 19 629 genes were tested in the genome-wide gene-level analysis using the proposed method. The Q-Q plots of the *P*-values from the genome-wide scan is shown in Figure 4. In our top list of 50 genes ranked on *P*-values, five genes (RAPGEF4, SLC1A1, SLC6A4, PXN and LEP) have been identified in association with autism by previous GWAS (www.gene.sfari.org).

We note that this preliminary analysis is mainly used to demonstrate the utility of gene-based association test with family samples and to show that our analysis strategy is well calibrated in terms of type I error. Clearly, the candidate genes need to be further validated in additional samples, and even by using additional analysis methods. There are many aspects that are worth further exploration and discussion in building analysis pipelines, including the selection of the correct inheritance model, the choice of SNP weights (we simply used MAF weights in this exemplary analysis) and whether to exclude synonymous SNPs from each gene.

Discussion

The burden test and sequence (SNP-set) kernel association test statistic are two most popular methods for rare variants association test, which have been routinely used in data analysis workflows for projects using exome chips, whole-exome sequencing or whole-genome sequencing. Where recent efforts to extend these statistics to related samples focus on continuous traits, our proposed approach addresses the family-based burden and SKAT statistics for both dichotomous and continuous traits. It can also be extended to consider a mixture of discrete and continuous traits. We have created an efficient statistical strategy to tackle this challenge by combining the KM and the generalized estimating equation methods. This new development is particularly important in the association studies of many complex diseases (such as psychiatric disorders) studies where binary phenotypic values are more often considered. The method is shown to be a powerful and computationally efficient test for rare variants, and more importantly it allows for both continuous and binary phenotypes. Both simulations and real data analysis demonstrate the proposed test can control type I error well under various scenarios. By leveraging the strength of the KM framework, the KM-based methods improve power by capturing the combined effects of multiple genetic variants and by providing a statistic with adaptively estimated degrees for freedom. They are also able to incorporate nonadditive and higher-order effects by expanding the feature space implicitly. By using the proposed strategy for correcting for the population stratification in the family data, our method allows for further power improvement of association tests through pooling samples from different ethnic groups.

Similar to the methods based on mixed-effect and generalzied linear models [13], our method allows easy modeling of covariates and, if interest exists, interactions among environmental and genetic factors [27]. It can also incorporate data on different combinations of related and unrelated individuals. However, the marginal model GEE method offers additional advantages over

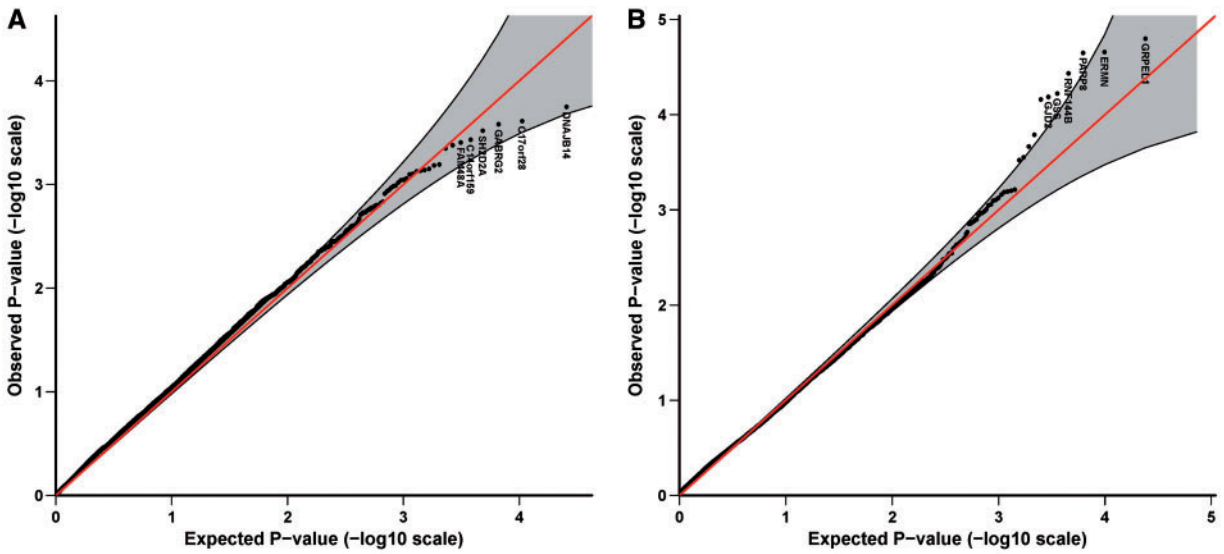

Figure 4. Q-Q plots for autism WES data. Results are shown for the burden (**A**) and GEE-SKAT (**B**) tests. The x axis represents $-\log_{10}$ expected *P*-values, and the y axis represents $-\log_{10}$ observed *P*-values. A colour version of this figure is available at BIB online: https://academic.oup.com/bib.

the mixed model method. First, the generalized mixed model is computationally more demanding and tends to yield less stable results in its implementation, making it less appealing for the analysis of binary phenotypes. Second, the GEE model is robust to the misspecification of correlation matrices among family members. Like most GEE-based approaches, the better the working correlation is specified, the more powerful is the test. The proposed quadratic score test will lose efficiency when the working correlation matrix is incorrectly assigned, especially when the family size is not constant [28, 29]. For the selection of working correlation structures, one may use the expected relatedness (kinship matrix) or realized relatedness calculated base on whole-genome marker data. However, we recommend simply using the identified working correlation matrix, especially for the initial exploratory analysis or when the family relationship is unknown or complex. We are concerned that the currently widely used empirical or realized kinship matrix calculated based on all SNP markers may not represent the true structure of dependence between the family members. But the misuse of the kinship matrix may have a greater impact on GLMM methods where more assumptions are required. The selection of the optimal working correlation matrix in the context of GEE-KM quadratic score test is worth further investigation. The choice can be possibly based on criteria such as the quasi-likelihood under the independence model criterion (QIC) [30] and the correlation information criterion (CIC) [31] or can be implicitly done based on the quadratic inference function (QIF) [32].

Because the information from parental and all other individuals is fully incorporated, both the GEE and the mixed model method will provide generally more powerful tests than the allelic transmission-based test such as the transmission disequilibrium test (TDT) and family-based association test (FBAT), especially in the current sequencing studies where the sample size is still limited. Compared with similar SKAT-type test developed based on FBAT [33], the GEE-SKAT allows for more flexible family structure. However, GEE-SKAT tests are conducted unconditional on parental genotypes, which are thus not robust to possible population stratification. The principal components of population structure need to be adjusted as covariates in all analyses. In this study, we applied a modified PCA by using the HGDP data as the reference map. This method overcomes the challenge in our data in which the PCA cannot be performed on parents because they are related. It also avoids the shrinkage phenomenon observed in calculating the predicted principal component scores [34]. Therefore, it is also effective when used in combination with other family-based association tests that are based on the generalized mixed model or GEE methods.

We have implemented the proposed method in an R package 'gskat' along with other useful functions for simulations. Although the asymptotic distribution of the GEE-KM is derived and implemented in the package, we recommend using *P*-values that are computed based on the perturbation method proposed in this article. Results from extensive simulations and real data applications show that this resampling method is much less affected by small sample sizes, ascertainment schemes and proportions of cases in collected samples. The genetic background of families that leads to developmental defects is of keen interest to both basic and clinical-oriented researchers. Combining SNVs and copy number variants [35] and other variant types together in a complete test will lead to a better assessment of inherited genetic burden and will be a valuable future extension.

In the same vein of the GEE framework, the method described can be readily applied to the association test with multiple related phenotypes. It is known that—when multiple phenotypes are correlated and measure the same or related underlying trait—a more powerful test can be constructed by jointly testing the common effect of a variant on multiple phenotypes. Similar to the correlation of effect size among SNPs in the family-based test, the correlation of effects of the same variant on different phenotypes will be unknown. Under the null hypothesis, the parameterρ disappears and is not identifiable. Thus, the same technique of calculating *P*-value can be used in the optimal test for multiple phenotypes.

Key Points

- gskat is a set of useful tools for gene prioritizing with mixed family and unrelated samples.
- GEE-KM provides a promising testing framework for large-scale mega-analysis.
- Need to further extend the KM and improve the computational efficiency to incorporate difference types of variants.

Supplementary Data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Acknowledgements

The authors wish to thank three anonymous reviewers for their comments.

Funding

National Institutes of Health (grant no. P20 CA192994, in part).

References

1. Thomas DC, Yang Z, Yang F. Two-phase and family-based designs for next-generation sequencing studies. *Front Genet* 2013;**4**:276.
2. Witte JS, Gauderman WJ, Thomas DC. Asymptotic bias and efficiency in case-control studies of candidate genes and gene-environment interactions: basic family designs. *Am J Epidemiol* 1999;**149**:693–705.
3. Chen W-M, Abecasis GR. Family-based association tests for genomewide association scans. *Am J Hum Genet* 2007;**81**:913–26.
4. Schifano ED, Epstein MP, Bielak LF, *et al*. SNP set association analysis for familial data. *Genet Epidemiol* 2012;**36**:797–810.
5. Chen H, Meigs JB, Dupuis J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 2013;**37**:196–204.
6. Listgarten J, Lippert C, Kang EY, *et al*. A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 2013;**29**:1526–33.
7. Oualkacha K, Dastani Z, Li R, *et al*. Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol* 2013;**37**:366–76.
8. Lippert C, Xiang J, Horta D, *et al*. Greater power and computational efficiency for kernel-based association testing of sets of genetic variants. *Bioinformatics* 2014;btu504.
9. Liang K-Y, Zeger SL. Longitudinal data analysis using generalized linear models. *Biometrika* 1986;**73**:13–22.
10. Wang X, Lee S, Zhu X, *et al*. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genet Epidemiol* 2013;**37**:778–86.
11. Wu MC, Lee S, Cai T, *et al*. Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 2011;**89**:82–93.
12. Zhang D, Lin X. Hypothesis testing in semiparametric additive mixed models. *Biostatistics* 2003;**4**:57–74.
13. Jiang D, McPeek MS. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol* 2014;**38**:10–20.
14. Lin D, Zeng D. On the relative efficiency of using summary statistics versus individual-level data in meta-analysis. *Biometrika* 2010;**97**:321–32.
15. Liu L, Sabo A, Neale BM, *et al*. Analysis of rare, exonic variation amongst subjects with autism spectrum disorders and population controls. *PLoS Genet* 2013;**9**:e1003443.
16. Lee S, Emond MJ, Bamshad MJ, *et al*. Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 2012;**91**:224–37.
17. Lee S, Wu MC, Lin X. Optimal tests for rare variant effects in sequencing association studies. *Biostatistics* 2012;**13**:762–75.
18. Davidson R, Flachaire E. The wild bootstrap, tamed at last. *J Econ* 2008;**146**:162–9.
19. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. New York: Chapman & Hall/CRC, 1994.
20. Zhu X, Li S, Cooper RS, *et al*. A unified association analysis approach for family and unrelated samples correcting for stratification. *Am J Hum Genet* 2008;**82**:352–65.
21. Li JZ, Absher DM, Tang H, *et al*. Worldwide human relationships inferred from genome-wide patterns of variation. *Science* 2008;**319**:1100–4.
22. Anderson CA, Pettersson FH, Clarke GM, *et al*. Data quality control in genetic case-control association studies. *Nat Protoc* 2010;**5**:1564–73.
23. Wang C, Szpiech ZA, Degnan JH, *et al*. Comparing spatial maps of human population-genetic variation using Procrustes analysis. *Stat Appl Genet Mol Biol* 2010;**9**:13.
24. Wang C, Zhan X, Liang L, *et al*. Improved ancestry estimation for both genotyping and sequencing data using projection procrustes analysis and genotype imputation. *Am J Hum Genet* 2015;**96**:926–37.
25. Schaffner SF, Foo C, Gabriel S, *et al*. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res* 2005;**15**:1576–83.
26. Park CG, Park T, Shin DW. A simple method for generating correlated binary variates. *Am Stat* 1996;**50**:306–10.
27. Lin X, Lee S, Christiani DC, *et al*. Test for interactions between a genetic marker set and environment in generalized linear models. *Biostatistics* 2013;**14**:667–81.
28. Rotnitzky A, Jewell NP. Hypothesis testing of regression parameters in semiparametric generalized linear models for cluster correlated data. *Biometrika* 1990;**77**:485–97.
29. Trégouët D-A, Ducimetiere P, Tiret L. Testing association between candidate-gene markers and phenotype in related individuals, by use of estimating equations. *Am J Hum Genet* 1997;**61**:189–99.
30. Pan W. Akaike's information criterion in generalized estimating equations. *Biometrics* 2001;**57**:120–5.
31. Hin LY, Wang YG. Working correlation structure identification in generalized estimating equations. *Stat Med* 2009;**28**:642–58.
32. Lindsay BG, Qu A. Inference functions and quadratic score tests. *Stat Sci* 2003;394–410.
33. Ionita-Laza I, Lee S, Makarov V, *et al*. Family-based assocation tests for sequence data, and comparisons with population-based association tests. *Eur J Hum Genet* 2014;**21**:1158–62.
34. Lee S, Zou F, Wright FA. Convergence and prediction of principal component scores in high-dimensional settings. *Ann Stat* 2010;**38**:3605.
35. Tzeng J-Y, Magnusson PK, Sullivan PF, *et al*. A new method for detecting associations with rare copy-number variants. *PLoS Genet* 2015;**11**:e1005403.