Stat Biosci (2018) 10:491–505
https://doi.org/10.1007/s12561-017-9197-9

Robust Rare-Variant Association Tests for Quantitative Traits in General Pedigrees

Yunxuan Jiang[1] · **Karen N. Conneely**[2] · **Michael P. Epstein**[2]

Received: 28 July 2016 / Accepted: 18 May 2017 / Published online: 5 June 2017

Abstract Next-generation sequencing technology has propelled the development of statistical methods to identify rare polygenetic variation associated with complex traits. The majority of these statistical methods are designed for case–control or population-based studies, with few methods that are applicable to family-based studies. Moreover, existing methods for family-based studies mainly focus on trios or nuclear families; there are far fewer existing methods available for analyzing larger pedigrees of arbitrary size and structure. To fill this gap, we propose a method for rare-variant analysis in large pedigree studies that can utilize information from all available relatives. Our approach is based on a kernel machine regression (KMR) framework, which has the advantages of high power, as well as fast and easy calculation of p-values using the asymptotic distribution. Our method is also robust to population stratification due to integration of a QTDT framework (Abecasis et al., Eur J Hum Genet 8(7):545–551, 2000b) with the KMR framework. In our method, we first calculate the expected genotype (between-family component) of a non-founder using all founders' information and then calculate the deviates (within-family component) of observed genotype from the expectation, where the deviates are robust to population stratification by design. The test statistic, which is constructed using within-family component, is thus robust to population stratification. We illustrate and evaluate our method using simulated data and sequence data from Genetic Analysis Workshop 18.

Karen N. Conneely and Michael P. Epstein have contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s12561-017-9197-9) contains supplementary material, which is available to authorized users.

✉ Michael P. Epstein
mpepste@emory.edu

[1] Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA, USA

[2] Department of Human Genetics, Emory University School of Medicine, Emory University, 615 Michael Street, Suite 301, Atlanta, GA 30322, USA

Keywords Rare variant · Pedigree · Quantitative trait · Population stratification

1 Introduction

Next-generation sequencing (NGS) studies of complex human traits and diseases are becoming commonplace for investigating the role of rare polymorphic variation in such phenotypes. Many analytic methods have been developed for the analysis of such rare variants with a particular emphasis on techniques that first aggregate information on rare variants within a gene of interest and then contrast this aggregated genetic information with the phenotypic outcome. The majority of such aggregation-based methods [16,18,22,24,37,38] focus on population-based designs or case–control designs. However, family-based study designs are gaining traction in NGS projects since they provide inherent benefits over the traditional population-based designs. In particular, families ascertained based on multiple relatives with a particular phenotype tend to enrich the sample for rare causal variants compared to a general population, thereby making such variants easier to detect [39].

The appeal of family-based NGS studies has led to the development of a few analytic methods tailored for rare-variant analysis in such designs. Such methods [6,13,15,29] generally apply a modeling framework that accounts for the relatedness of familial samples through appropriate modeling of kinship. However, such methods do not take into account the potential bias of findings due to population stratification. Population stratification is the presence of systematic differences between subpopulations both in the allele frequencies of the rare variants under study as well as in the distribution of phenotype. Failure to model these differences will lead to inflated false positive rate and decreased power to detect real associations. For rare variants, the issue of population stratification is more severe than for common variants, as rare variants are more likely to be young mutations which are more population-specific [11]. It has been shown that inclusion of self-reported ethnicity as a covariate is not sufficient to adjust for population stratification [31]. Similarly, standard methods to adjust for population stratification for common variants may not be as effective an adjustment for rare variants. In particular, genomic control can lead to very conservative results for rare variants [14]. Although principal components works well for spatially distinctive populations, the procedure fails for spatially non-distinctive populations [23].

With these concerns in mind, Jiang et al. [15] developed a rare-variant association test for quantitative traits in parent–child trios and nuclear families that, by design, was robust to population stratification. The method was motivated by the QTDT framework [1], which showed that the observed genotype of a familial subject could be partitioned into orthogonal between-family and within-family components. The between-family component can be defined as the expected value of the subject's genotype within the family and can be constructed as the average of the parents' genotype or the average of the siblings' genotype. The within-family component is the deviation of the observed genotype from the between-family component. While the between-family component is sensitive to population stratification, the within-family component is robust to stratification since it is based on a family-specific deviation. Utilizing a kernel machine regression (KMR) framework for multi-marker analysis of familial

quantitative phenotypes [6,15,30], Jiang et al. [15] created a robust rare-variant test by replacing observed sample genotypes in the standard KMR with their corresponding within-family genotypic components. Simulation results demonstrated the approach yielded appropriate type I error even when strong confounding existed within the sample. As with other KMR approaches, Jiang et al. [15] approach derived p-values analytically using Davies' [9] method, thereby allowing easy application to large-scale sequencing studies.

While the work of Jiang et al. [15] provides a powerful approach that is robust in the presence of population stratification, the method's design limited its application only to nuclear families and parent–child trios. However, many sequencing studies have emerged that utilize phenotype and genotype data collected on multiplex pedigrees that are larger and contain more distant relationships than those in nuclear families. Examples of such studies include the Epi4K study of epilepsy Epi4K Consortium [10]. and the Genetic Analysis Workshop (GAW18) study of blood pressure. Large pedigrees have unique features that make them ideal for mapping traits associated with rare variants. Compared to nuclear families or trios, rare variants are further enriched in large pedigrees [34]. It has been shown that large pedigree studies have increased power compared to smaller families with the same total number of samples, especially for rare-variant sequencing data [32,35,36]. In addition to improved power, analysis of large pedigrees can provide evidence for both co-segregation and association, while population-based studies can only provide evidence for association [17,26,34]. Further, the study of large pedigrees provides a cost-effective strategy for rare-variant analysis as it enables *in silico* imputation of rare-variant genotypes in non-sequenced subjects using information from sequenced relatives coupled to knowledge of inheritance flow [7,34]. With a large pedigree-based study design, researchers can also combine sequencing-based association studies with linkage analyses [26]. Recent research has identified rare variants associated with several diseases or traits such as hyperkalemic hypertension [21], spinocerebellar ataxias [33], hypolipidemia [25], and lithium responsive bipolar disorder [8] by combining association and linkage approaches.

Given the obvious value of extended pedigrees, it would be useful to develop a robust family-based association test of rare variants for such designs that are also computationally efficient. While the method of Jiang et al. [15] is both robust and fast, it is also only limited to trios and nuclear families and therefore cannot be applied to studies such as GAW18 that possesses sequence data for 20 Mexican American families with an average pedigree size of 70 (see sample pedigree in Supplementary Fig. S1). Therefore, in this paper, we propose an expansion of Jiang et al. [15] framework to allow robust and efficient analysis of multiplex families of arbitrary size and structure. To do so, we employ a non-trivial modification of the QTDT framework for use in extended pedigrees developed by Abecasis et al. [2] that uses information from all genotyped family members to construct a more informative between-family genotypic component. We then derive the within-family component for each genotype and integrate this information within the KMR framework of Schifano et al. [30] to obtain a rare-variant test that is robust to population stratification. In the following sections, we will first introduce our study setting, followed by how we use the QTDT framework to decompose genotype information to obtain a robust within-family component. We then show how to integrate this information within a KMR framework

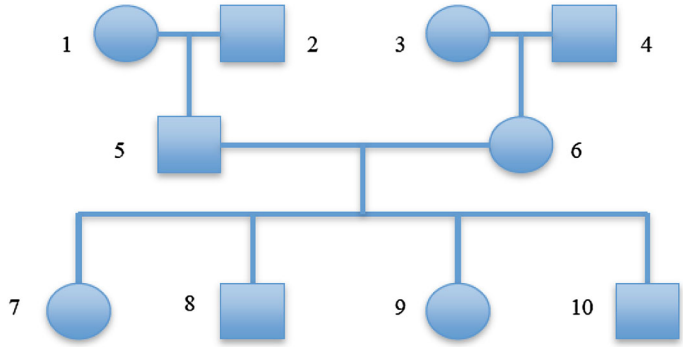

Fig. 1 Example of pedigree structure

to yield our robust test. We will also describe how we can improve the power of our robust test by pre-screening potential trait-influencing genes using genotype and phenotypic information from founders across families. Such founder information is orthogonal to the within-family information used in our proposed test. We then evaluate our method using both simulation studies and sequencing data from a study of systolic and diastolic blood pressure (SBP and DBP) provided by the GAW18.

2 Materials and Methods

2.1 Study Design and Notation

We assume a family-based study consisting of N families, where each family consists of a large pedigree. While we use Fig. 1 as an example here to show the structure of the large pedigree, our method can be applied to any family structure and can accommodate any family size unlike the original framework of Jiang et al. [15]. Suppose there are s rare variants in a gene of interest, and let $\boldsymbol{G}_{ij}$, a $s \times 1$ vector, represent the genotypes of the s rare variants for the jth ($j = 1, 2, \ldots, n_i$) individual in the ith ($i = 1, 2, \ldots, N$) family. We assume an additive model, and let components in $\boldsymbol{G}_{ij}$ take the value of 0, 1, 2, indicating the number of copies of minor alleles at each site. If an individual is not genotyped, then we leave $\boldsymbol{G}_{ij}$ undefined. Let $\boldsymbol{X}_{ij}$, a $c \times 1$ vector, denote the covariates, and denote Y_{ij} as the value of the quantitative outcome for the jth individual in the ith family. For non-founders (defined as individuals with ancestors included in the pedigree, e.g., individuals 5–10 in Fig. 1), let M_{ij} and F_{ij} be the index of mother and father of j individual in the ith family, respectively. For founders (defined as individuals with no ancestors in the pedigree, e.g., individuals 1–4 in Fig. 1), we leave M_{ij} and F_{ij} undefined.

2.2 KMR Framework for Pedigree Data

We create our robust rare-variant association test for a quantitative trait based on the KMR test of Schifano et al. [30] and Chen et al. [6] for association testing of a group of genetic variants with a continuous phenotype allowing for related individuals. As

shown by these authors, the KMR test can be implemented in a linear mixed modeling framework with mean and variance defined through the model:

$$Y_{ij} = \boldsymbol{X}_{ij}^{\mathrm{T}}\alpha + h\left(\boldsymbol{G}_{ij}\right) + f_{ij} + \varepsilon_{ij}, \tag{1}$$

where α is a $c \times 1$ vector of coefficients for $\boldsymbol{X}_{ij}$, f_{ij} is the random effect to account for within-family correlation, and ε_{ij} is the random error term. We further assume that the random effects within a family, $\boldsymbol{f}_i = (f_{i1}, f_{i2}, f_{i3}, \ldots, f_{in_i})^{\mathrm{T}}$, follow a multivariate normal distribution $\boldsymbol{f}_i \sim \mathrm{MVN}(0, 2\Phi_i\sigma^2_{\mathrm{pg}})$. Here Φ_i is the kinship matrix for the ith family (elements in Φ_i represent the pairwise kinship coefficients between relatives in the ith family) and σ^2_{pg} represents the variance due to the shared polygenic effect. We also assume that the random environmental effect ε_{ij} is independent among subjects within and between families and follows a normal distribution with mean 0 and variance σ^2_e.

Within Eq. (1) above, $h(\boldsymbol{G}_{ij})$ is a function of $\boldsymbol{G}_{ij}$ defined through a positive semidefinite kernel function $k(\cdot, \cdot)$. Following Liu et al. [20] and Kwee et al. [16], $h(\boldsymbol{G}_{ij})$ can be represented as $\sum_{i'}\sum_{j'}\vartheta_{i'j'}k(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'})$, where $\vartheta_{i'j'}$ are unknown parameters. It is worth noting that the kernel function, $k(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'})$, measures the genetic similarity between subject j in family i and subject j' in family i' and contrasts this similarity to phenotypic similarity between the two subjects. It has been shown that appropriate choice of the kernel can increase the power [37]. Frequently used kernels include the identity by state (IBS) kernel or the linear kernel. The IBS kernel, which takes the form $k(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'}) = \sum_{l=1}^{s}(2 - |G_{ijl} - G_{i'j'l}|)$, measures the genetic similarity as the number of alleles that share by state. It assumes a nonlinear effect of each rare variant and can thus enable the study of epistatic effects. The linear kernel, on the other hand, assumes a linear relationship between the trait and the variants. The kernel takes the form $k(\boldsymbol{G}_{ij}, G_{i'j'}) = \sum_{l=1}^{s}(G_{ijl}G_{i'j'l})$. Additionally, we can include prior knowledge of variants that are possibly causal in the gene by assigning each variant a weight. If prior knowledge is not available, weights can also be calculated as a function of minor allele frequency (MAF; under the logic that the rarer the allele, the more likely it is selected against and therefore the more likely it is to be pathogenic). Wu et al. [37] suggest calculating the weights based on a beta distribution, which assigns greater weight to less frequent variants. For a given weight, we can create weighted kernels such as the weighted linear kernel $k(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'}) = \sum_{l=1}^{s}w_l(G_{ijl}G_{i'j'l})$, where w_l denotes a normalized weight for variant l in the gene.

It can be easily shown that the estimator of h takes the same form as in the linear mixed model with h as a random effect [20,30]:

$$y = X\alpha + h + f + e, \tag{2}$$

where $\boldsymbol{\alpha}$ is a $c \times 1$ vector of coefficients for fixed effect $\mathbf{X}$, $\boldsymbol{h}$ is an $\sum_{i=1}^{N} n_i \times 1$ vector of random effects that follow an arbitrary distribution with mean 0 and variance $\tau\boldsymbol{K}$, where K is the genetic similarity matrix with element $\langle ij, i'j'\rangle$ equal to $k(\boldsymbol{G}_{ij}, \boldsymbol{G}_{i'j'})$; $\boldsymbol{f} = (\boldsymbol{f}_1^{\mathrm{T}}, \boldsymbol{f}_2^{\mathrm{T}}, \ldots, \boldsymbol{f}_N^{\mathrm{T}})^{\mathrm{T}} \sim N(0, 2\sigma^2_{\mathrm{pg}}\boldsymbol{\Phi})$, where $\boldsymbol{\Phi}$ is a block diagonal matrix with Φ_i on the diagonal. Finally, $e = (\boldsymbol{e}_1^{\mathrm{T}}, \boldsymbol{e}_2^{\mathrm{T}}, \ldots, \boldsymbol{e}_N^{\mathrm{T}})^{\mathrm{T}} \sim N(0, \sigma^2_e\mathbf{I})$. Thus, the test of whether genotype is associated with the outcome is equivalent to

testing whether the random component h equals 0 or not. We adopted the variance component score test, which is the locally most powerful test [19]. As $\boldsymbol{h}$ has the variance of $\tau \boldsymbol{K}$, the test of whether $\boldsymbol{h} = 0$ is equivalent to testing whether $\tau = 0$. The null hypothesis is H_0: $\tau = 0$, and the test statistic takes the form:

$$Q = \frac{1}{2}\left(Y - X\hat{\alpha}_0\right)\hat{V}_0^{-1} K \hat{V}_0^{-1}\left(Y - X\hat{\alpha}_0\right), \tag{3}$$

where all parameters are estimated under the null hypothesis. $\widehat{V_0} = 2\widehat{\sigma^2_{\text{pg}}}\boldsymbol{\Phi} + \widehat{\sigma^2_e} I$ denotes the sample variance/covariance matrix estimated under the null. To obtain the null distribution of Q, we define a projection matrix $P = \hat{V}_0^{-1} - \hat{V}_0^{-1} X (X^{\mathrm{T}} \hat{V}_0^{-1} X)^{-1} X^{\mathrm{T}} \hat{V}_0^{-1}$ such that $P \widehat{V_0} P = P$. Thus, under the null, we have

$$Q = \frac{1}{2} Y^{\mathrm{T}} P K P Y = \sum_{i=1}^{N} \lambda_i \chi^2_{1i}, \tag{4}$$

where λ_i are eigenvalues of $\frac{1}{2} D \hat{V}_0^{-1/2} K \hat{V}_0^{-1/2} D$, here $D = I - \hat{V}_0^{-1/2} X (X^{\mathrm{T}} \hat{V}_0^{-1/2} X)^{-1} X^{\mathrm{T}} \hat{V}_0^{-1/2}$. As χ^2_{1i} are independently and identically distributed random variables, Q is distributed as an asymptotic mixture of χ^2 distributions, and the p-values can be calculated using the Davies method [9].

2.3 QTDT Framework for General Pedigrees

In the presence of population stratification, association testing of $\boldsymbol{G}_{ij}$ with Y_{ij} in models (1) and (2) may lead to spurious association due to the underlying differences in allele frequencies of the subpopulations. However, for family studies, family members can be used as internal controls, where an expected genotype can be constructed using the family members' information. Tests based on the within-family component (deviation of observed genotype from expected within family) will not be influenced by population structure, even in the most extreme case, where each of the N pedigrees is drawn from a different population. Here, we leverage the work of Abecasis et al. [1] and present the method to calculate transmission scores for individuals in general pedigrees.

The QTDT framework [1] for general pedigrees decomposes a genotype into a between-family component (which is sensitive to population stratification) and a within-family component (which is robust to population stratification). For relative j in family i, let $\boldsymbol{B}_{ij}$ and $\boldsymbol{W}_{\boldsymbol{ij}}$ denote vectors of between-family and within-family genotype components for the s rare-variant genotypes in $\boldsymbol{G}_{\boldsymbol{ij}}$. Assuming all parents in the pedigree are genotyped, the between-family component for founders (with no ancestors included in the pedigree) will be equal to their observed genotypes, while the between-family component for non-founders at each rare-variant genotype is equal to the average genotype of the between-family components of that individual's parents: such that $\boldsymbol{B}_{ij} = \frac{\boldsymbol{B}_{M_{ij}} + \boldsymbol{B}_{F_{ij}}}{2}$. Using the pedigree in Fig. 1 as an example,

suppose all the individuals in the pedigree are genotyped. Suppressing the family index for ease of presentation, the between-family components for founders 1, 2, 3, and 4 are $\boldsymbol{B}_1 = \boldsymbol{G}_1$, $\boldsymbol{B}_2 = \boldsymbol{G}_2$, $\boldsymbol{B}_3 = \boldsymbol{G}_3$, $\boldsymbol{B}_4 = \boldsymbol{G}_4$, respectively. For the non-founders in the second generation, the between-family component for individual 5 is $\boldsymbol{B}_5 = \frac{\boldsymbol{B}_1+\boldsymbol{B}_2}{2}$, and between-family component for 6 is $\boldsymbol{B}_6 = \frac{\boldsymbol{B}_3+\boldsymbol{B}_4}{2}$. For the non-founders in the third generation, the between-family components for individuals 7–10 are $\frac{\boldsymbol{B}_5+\boldsymbol{B}_6}{2} = \frac{\boldsymbol{B}_1+\boldsymbol{B}_2+\boldsymbol{B}_3+\boldsymbol{B}_4}{4}$. It can be seen that, in the situation where all founders are genotyped, the between-family component of any non-founder is calculated as follows:

$$\boldsymbol{B}_{ij} = \sum_{f \in F} 2\varphi_{ijf}\boldsymbol{G}_{if}, \tag{5}$$

where in the ith family, f is the index of founders, $\boldsymbol{G}_{if}$ is the rare-variant genotype vector of the founder, φ_{ijf} is the kinship coefficient between individual j and founder f, and F is the set of all the genotyped founders.

In the situation where the parents' genotypes are missing, the between-family component $\boldsymbol{B}_{ij}$ is equal to the average of the genotypes for all sibling of relative j. For example in Fig. 1, if individuals 5 and 6 are not genotyped, then the between-family component for individuals 7–10 is $\frac{\boldsymbol{G}_7+\boldsymbol{G}_8+\boldsymbol{G}_9+\boldsymbol{G}_{10}}{4}$. The average of genotypes of siblings in the family is the sufficient statistic for the between-family component [1]. We note that, when applied to parent–child trios and nuclear families, the proposed method for calculating the between-family component we describe here is then equivalent to the forms of the between-family component outlined in the work of Jiang et al. [15].

The within-family genotype vector for the s rare-variant genotypes $\boldsymbol{W}_{ij}$ is then calculated as the difference between the observed genotype vector and the between-family genotype vector:

$$\boldsymbol{W}_{ij} = \boldsymbol{G}_{ij} - \boldsymbol{B}_{ij}. \tag{6}$$

Positive values within $\boldsymbol{W}_{ij}$ indicate excess transmission of the minor (reference) allele, while negative values of $\boldsymbol{W}_{ij}$ indicate excess transmission of the major allele. As discussed above, the within-family component is not influenced by population substructure; thus, the test on the within-family component is robust to population stratification.

As discussed before, directly testing based on the observed rare-variant genotypes in models (1) and (2) will lead to spurious association in the presence of population stratification. For our robust test, we follow the same approach as in our earlier work [15] and simply calculate $\boldsymbol{W}_{ij}$ as described above, replace $\boldsymbol{G}_{ij}$ with $\boldsymbol{W}_{ij}$ in Eqs. (1) and (2), and construct our score statistic Q in (3) using $\boldsymbol{W}_{ij}$.

2.4 Screening Methods

Although the within-family component has the advantage of robustness to population stratification, constructing tests based only on the within-family genotypic component while ignoring the between-family component reduces power. However, if founders' phenotype and genotype data are available, we can borrow the idea of Purcell et al. [27] to implement a screening procedure to potentially increase power. Specifically,

we use the founders' phenotype and genotype information in the first stage to identify those regions showing strongest signals of association. We can perform such testing using standard burden or variance component tests for unrelated subjects. We then implement a second stage where we test only the top regions from the first stage using our proposed test in (3) based on the within-family genotypic component; the number of top regions in the second stage can take a value between 1 and the total number of regions. In this project, we assume 10–50% of the regions enter the second stage. By pre-screening in this manner, we reduce the multiple testing burden for our robust test, thereby increasing power. As the within-family component and the between-family component are orthogonal to each other by design [1], population stratification that can invalidate the first-stage analysis using founders will not invalidate the within-family component test.

2.5 Simulation Studies

We evaluate type I error rate and power of our method using simulated sequencing data generated by *cosi* [28], which has high resemblance with empirical data. To simulate large pedigrees, we first use *cosi* to simulate 5000 haplotypes of European ancestry and 5000 haplotypes of African ancestry. We then randomly draw and pair haplotypes within each population and randomly select one haplotype from each parent to pass down to offspring. Our simulated pedigree has the same structure as Fig. 1. We assume that there are 10 non-overlapping genes or regions of interest, each 30 kb long. We show the empirical distribution of rare variants in these regions across simulated datasets in Supplementary Fig. S2.

For each family, we simulate phenotype data from a multivariate normal distribution, whose mean and variance vary according to different scenarios. For type I error rate simulations, all 10 regions are null, while for power simulations we randomly select 1 region of the 10 to harbor causal variation. Rare variants are defined as variants with MAF smaller than 3%. To simulate population substructure, we simulate the outcome for the null model as follows: $Y_{ij} = \gamma I_{\text{African,ij}} + f_{ij} + e_{ij}$, where γ is the mean trait difference between European and African, and $I_{\text{African,ij}}$ is the indicator variable, which is 1 for African individuals and 0 for European individuals. For the power simulations, we let either 5 or 15% of the rare variants in the causal region influence phenotype. Within each family, we simulate the random effects $\boldsymbol{f}_{ij}$ through $f_i \sim \text{MVN}(0,\ 0.56 \times 2\Phi_i)$. e_{ij} is the random error and follows a standard normal distribution. For each causal variant, we define the effect size as $\beta = c \times |\log_{10} \text{MAF}|$, where c is a pre-defined constant. Thus, the outcome is simulated as $Y_{ij} - \gamma I_{\text{African},ij} + \beta_{ij} \times G_{ij} + f_{ij} + e_{ij}$. We perform 5000 simulations to evaluate type I error rate. For power simulation, we also perform 5000 simulations and calculate power as the proportion of simulations with the causal region correctly identified. Unless otherwise noted, we applied a linear genotype kernel for analysis.

2.6 GAW18 Data

The GAW18 provides whole genome sequence data for extended pedigrees and phenotypes such as SBP and DBP. The dataset was drawn from the T2D-GENES Consortium Project 2; a family-based study that aims to identify low-frequency variants that increase the risk of type-2 diabetes. The original dataset contains whole genome sequences for the odd-numbered chromosomes only (chromosomes 1, 3, 5,…,21) for 464 individuals from 20 Mexican American families. The dataset we used in this project contains 959 individuals of which 464 of them were directly sequenced by Complete Genomics, Inc., while the remaining 495 had sequence data imputed from array-based genotype data by the T2D-GENES Consortium. In addition to SBP and DBP, the dataset also includes information on age, gender, current use of antihypertensive medicine, and current smoking status. We include these phenotypes as covariates in our model. Detailed information about the dataset can be found at Almasy et al. [4].

After standard data cleaning procedure removed subjects with missing SBP or DBP measurements, our final dataset contained 855 individuals. Genes were annotated using information from the 1000 Genome Project (http://www.1000genomes.org/). We tested all genes in the 11 odd-numbered chromosomes, where each gene was tested individually. For each gene, we calculated the empirical frequency of the variants within the gene and only performed tests on the rare variants, where a rare variant was defined as having a MAF less than 3%. For perspective, we show the empirical distribution of rare variants within genes in the GAW18 project in Supplementary Fig. S3. We constructed the test statistics using within-family components as defined above.

3 Results

3.1 Type I Error

We first performed null simulations to show that population stratification can lead to inflated type I error rate for sequencing studies of large pedigrees. Figure 2 summarizes the empirical type I error rates of a study with 25 European pedigrees and 75 African pedigrees, each with the same size and family structure as shown in Fig. 1. We first set the mean trait difference (γ) between European and African to be 1 (Fig. 2, left) and further increased it to 2 (Fig. 2, right). Both figures show that in the presence of population stratification, test statistics constructed on observed genotype have inflated type I error rates (leftmost bars in each panel of in Fig. 2). As population structure becomes more extreme, the inflation becomes more severe (Fig. 2, right). We then performed tests based on the robust test based on our two-stage screening procedure using founders' genotypes and phenotypes. Figure 2 shows that testing on the within-family component combined with the screening method leads to appropriate control of the type I error rate in the presence of population stratification.

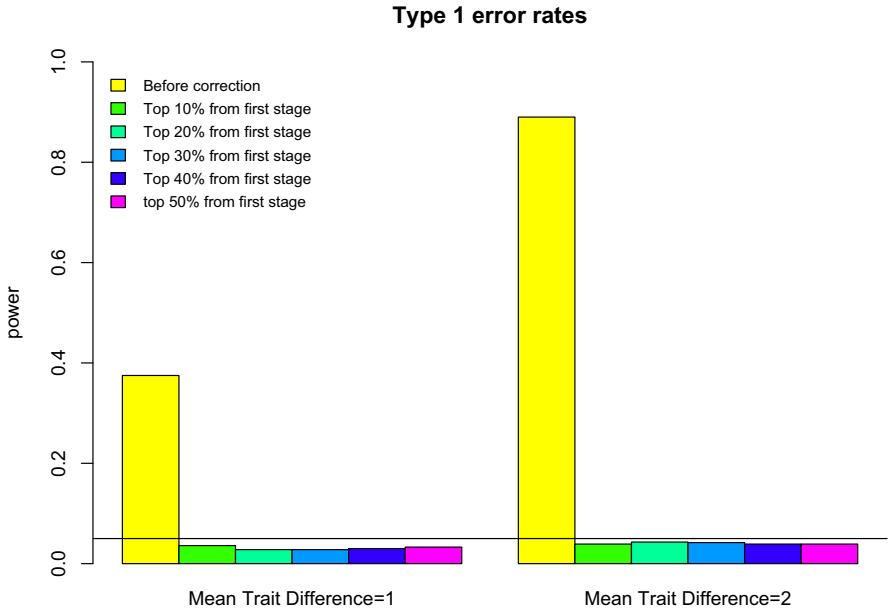

Fig. 2 Type I error rates. *Left* mean trait difference between European and African is 1. *Right* mean trait difference between European and African is two 10–30 kb regions are simulated. *Yellow bar* type I error rate tested on observed genotype. *Others* type I error rate tested on within-family component, with different number of genes at second stage. *Black line* $y = 0.05$

3.2 Power

We next examined power of the proposed robust test. For power simulations, we assume the mean trait different between European and African is 0.25. For each simulation, we randomly drew 25 European pedigrees and 75 African pedigrees from the haplotype pools. We varied the percentage of rare causal variants in the causal region from 5% (Fig. 3a) to 15% (Fig. 3b). We also assumed different effect sizes ($\beta = c \times |\log_{10} \text{MAF}|$) for the causal variants by letting c take the values 0.4, 0.5, and 0.6. Figure 3 shows that power increases as the percentage of causal variants in a region increases and as the effect size increases. We next investigated whether the two-stage screening approach using founder information improves power over a within-family analysis that ignores screening. As shown in Fig. 3, screening on the top 10–50% of hits can yield noticeable improvements in power over the naïve strategy. In addition to applying the linear genotype kernel, we also considered a weighted genotype linear kernel for screening and analysis (with weights based on MAFs using the weight function of Wu et al. [37]). Results, which we show in Supplementary Fig. S4, show similar results to the linear genotype kernel. With screening, we observed slight improvement of the weighted linear kernel over the unweighted linear kernel, particularly when larger effect sizes were assumed.

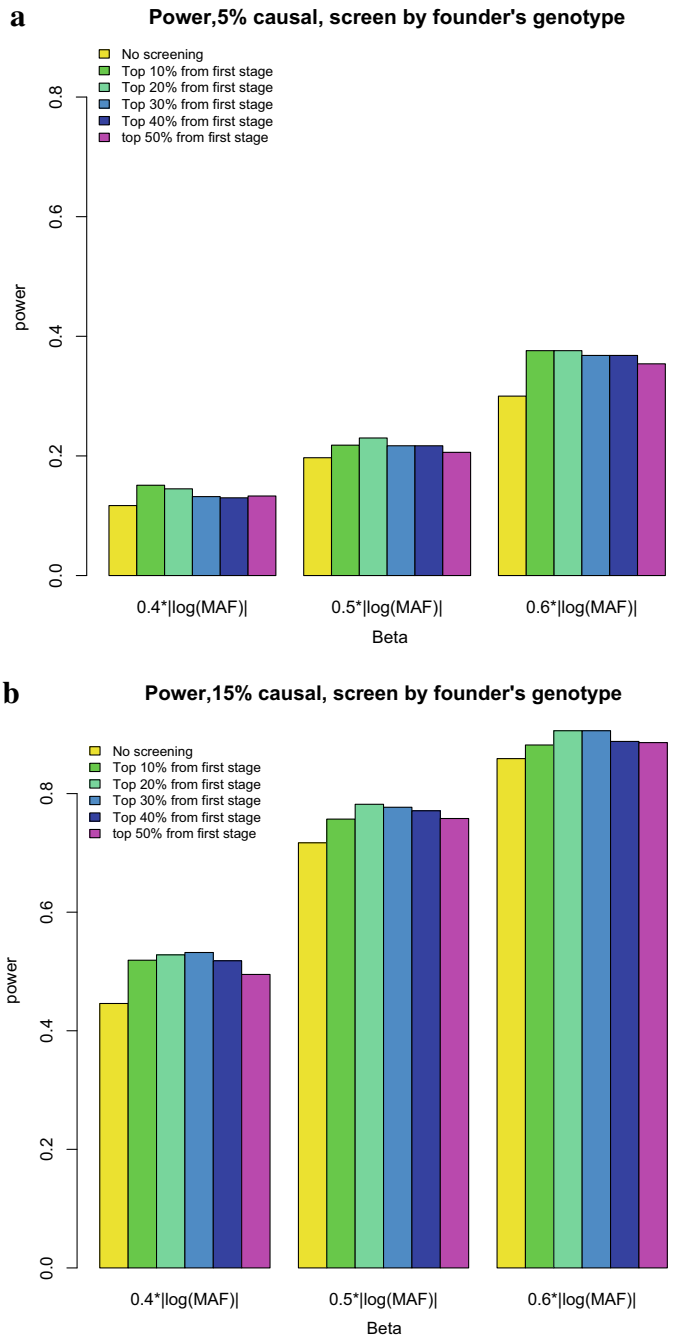

Fig. 3 Power to detect rare-variant association in large pedigrees. **a** 5% of rare variants in the causal region are causal variants. **b** 15% of rare variants in the causal region are causal variants. *Yellow bars* power without screening. *Others* power with screening. Mean trait different between European and African is 0.25. 10–50% regions entered second stage

3.3 Application to GAW18 Dataset

We used GAW18 data to test for association between DBP/SBP and genes on odd numbered chromosomes. Within each gene, we calculated empirical frequencies of variants and only tested on variants with frequencies smaller than 3%. GAW18 provides longitudinal phenotype information, where SBP and DBP were measured in up to four follow-ups for each subject. We used the baseline measurement to test for association. We also controlled for age, gender, current usage of antihypertensive medicine, and current smoking status in our model. The pedigrees are relatively large in the dataset. The median number of individuals in a pedigree is 37 (min 22, max 74). Among the participants, 20.2% of them smoke, 9.4% took medicine, and 57.7% of them are female.

We performed association tests using our robust test. The genome-wide significance level with Bonferroni correction is $\alpha_{\text{Bonferroni}} = 0.05/7034 = 7.1 \times 10^{-6}$. We chose the linear-weighted kernel and used the Davies method to calculate p-values. Following Wu et al. [37], the weight is calculated as $w_j \sim \text{Beta}(\text{MAF}_j, 1, 25)$. The results of testing SBP and DBP are summarized in Fig. 4. As shown in Fig. 4, we did not observe any genes passing the genome-wide significance level (7.1×10^{-6}, based on Bonferroni adjustment for 7034 genes). At the suggestive level (1×10^{-4}), one gene on chromosome 21 is associated with SBP, and one gene on chromosome 7 is associated with DBP. The gene associated with SBP is open reading frame 33 (C21orf33), which is a protein-coding gene and is over-expressed in Down syndrome Yahya-Graison et al. [3]. LSM5 is associated with DBP at the suggestive level. It has been found that human LSM1–7 genes were expressed in Hela cells within cytoplasmic foci Ingelfinger et al. [12], which contains important factors in the degeneration of mRNA. In addition to the Manhattan plots shown in Fig. 4, we also constructed QQ plots of results using both the observed genotypes and the within-family components of the genotypes. We present these QQ plots in Supplementary Fig. S5, which show inflation of SBP (but not DBP) when analyzing observed genotypes. We observed no such inflation when analyzing the within-family component, although results for SBP showed some deflation in p-values.

4 Discussion

In this paper, we presented a framework for rare-variant sequencing studies in large pedigrees. Large pedigrees have several important features that make them ideal for finding traits with associated rare variants. Our previous work for robust and efficient family-based analysis [15] was only applicable to parent-case trios or nuclear families and so, in this work, we expand the work to handle these large pedigrees of arbitrary size and structure such as those in the GAW18 study of blood pressure. Our model, which combines a kernel machine framework for rare-variant analysis with a QTDT framework for general pedigrees, provides a powerful, efficient, and robust way to identify such associations in large pedigree studies. As the test score statistics follows an asymptotically mixed χ^2 distribution, the calculation of p-values is much easier

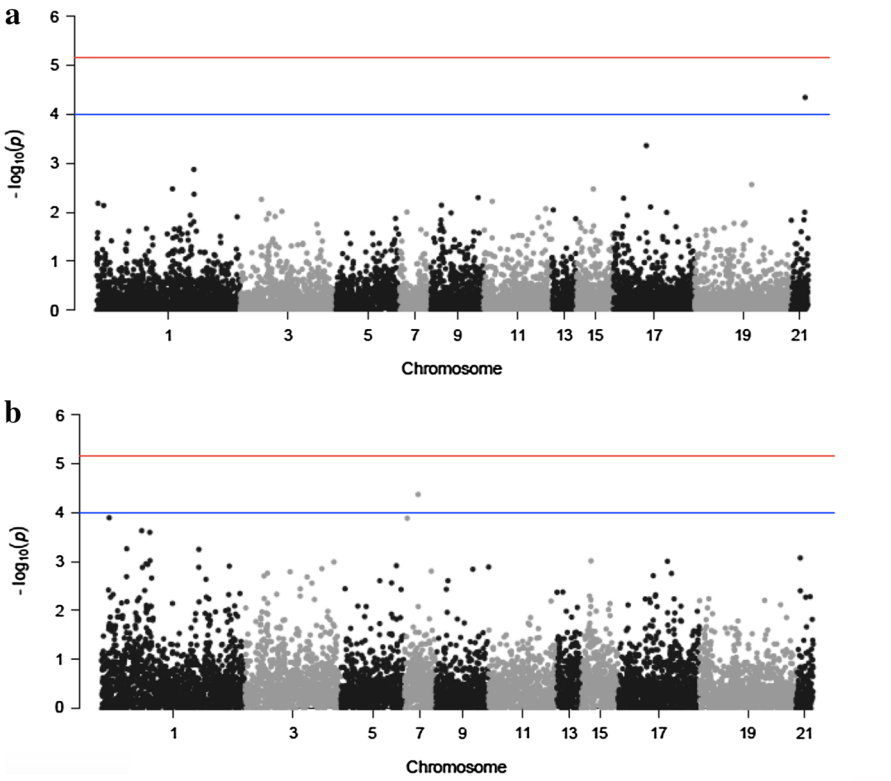

Fig. 4 Manhattan plots for GAW18 analyses. **a** Association analyses between SBP and within-family component of genotypes within genes on odd number of chromosomes. **b** Association analyses between DBP and within-family component of within genes on odd number of chromosomes. *Red line* genome-wide significant level ($p < 7.1 \times 10^{-6}$), *blue line* suggestive significant level ($p < 1 \times 10^{-4}$)

compared to other methods. This feature also makes our model applicable to large-scale genetic studies.

We also applied our method on GAW18 data to identify SBP/DBP-associated rare variants. We tested all the genes on odd numbers of chromosomes. This application gives an example that our method can be easily applied to large-scale data. The analysis of a gene takes 70 s on a 768 processors running Linux OS with 512 GB or RAM.

The data from GAW18 are based on 20 extended Mexican American families. For studies that do not have records of participants' geographic origin or studies whose participants are from different origins, our method provides a robust way to perform the test.

In this project, we assumed that rare variants only associated with a single phenotype. However, there is substantial interest in identifying genetic factors with pleiotropic effects that influence multiple distinct phenotypes. Current methods for family data are not well equipped to investigate the effect of pleiotropy. For example, while analyzing GAW18 data, analyses seeking to identify genes simultaneously associated with both SBP and DBP cannot be performed. However, Broadaway et

al. [5] provide a framework that can test cross-phenotype effects of rare variants. Their method is based on kernel distance-covariance, whose test statistics also asymptotically follow a mixed χ^2 distribution. In contrast to our method presented here, Broadaway et al. focused only on unrelated individuals. In the future, we would like to combine our robust test with the method of Broadaway et al. (2016) to test cross-phenotype effects of rare variants in related individuals.

Acknowledgements This work was supported by NIH Grants GM117946 and HG007508. The Genetic Analysis Workshop 18 (GAW18) is supported by NIH Grant R01 GM031575. The GAW18 whole genome sequence data were provided by the T2D-GENES Consortium, which is supported by NIH Grants U01 DK085524, U01 DK085584, U01 DK085501, U01 DK085526, and U01 DK085545. The other genetic and phenotypic data for GAW18 were provided by the San Antonio Family Heart Study and San Antonio Family Diabetes/Gallbladder Study, which are supported by NIH Grants P01 HL045222, R01 DK047482, and R01 DK053889.

References

1. Abecasis GR et al (2000) A general test of association for quantitative traits in nuclear families. Am J Hum Genet 66(1):279–292
2. Abecasis GR et al (2000) Pedigree tests of transmission disequilibrium. Eur J Hum Genet 8(7):545–551
3. Aït Yahya-Graison E et al (2007) Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. Am J Hum Genet 81(3):475–491
4. Almasy L et al (2014) Data for Genetic Analysis Workshop 18: human whole genome sequence, blood pressure, and simulated phenotypes in extended pedigrees. In: BMC proceedings. BioMed Central
5. Broadaway KA et al (2016) A statistical approach for testing cross-phenotype effects of rare variants. Am J Hum Genet 98(3):525–540
6. Chen H et al (2013) Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37(2):196–204
7. Cheung CY et al (2014) A statistical framework to guide sequencing choices in pedigrees. Am J Hum Genet 94(2):257–267
8. Cruceanu C et al (2013) Family-based exome-sequencing approach identifies rare susceptibility variants for lithium-responsive bipolar disorder 1. Genome 56(10):634–640
9. Davies RB (1980) Algorithm AS 155: the distribution of a linear combination of χ^2 random variables. J R Stat Soc C 29(3):323–333
10. Epi4K Consortium (2012) Epi4K: gene discovery in 4000 genomes. Epilepsia 53(8):1457–1467
11. Gravel S et al (2011) Demographic history and rare allele sharing among human populations. Proc Natl Acad Sci USA 108(29):11983–11988
12. Ingelfinger D et al (2002) The human LSm1-7 proteins colocalize with the mRNA-degrading enzymes Dcp1/2 and Xrnl in distinct cytoplasmic foci. RNA 8(12):1489–1501
13. Jiang D, McPeek MS (2014) Robust rare variant association testing for quantitative traits in samples with related individuals. Genet Epidemiol 38(1):10–20
14. Jiang Y et al (2013) Assessing the impact of population stratification on association studies of rare variation. Hum Hered 76(1):28–35
15. Jiang Y et al (2014) Flexible and robust methods for rare-variant testing of quantitative traits in trios and nuclear families. Genet Epidemiol 38(6):542–551
16. Kwee LC et al (2008) A powerful and flexible multilocus association test for quantitative traits. Am J Hum Genet 82(2):386–397
17. Laird NM, Lange C (2006) Family-based designs in the age of large-scale gene-association studies. Nat Rev Genet 7(5):385–394
18. Lee S et al (2012) Optimal tests for rare variant effects in sequencing association studies. Biostatistics 13(4):762–775
19. Lin X (1997) Variance component testing in generalized linear models with random effects. Biometrika 84:309–326

20. Liu D et al (2007) Semiparametric regression of multidimensional genetic pathway data: least-squares kernel machines and linear mixed models. Biometrics 63(4):1079–1088
21. Louis-Dit-Picard H et al (2012) KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. Nat Genet 44(4):456–460
22. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2):e1000384
23. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44(3):243–246
24. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. Genet Epidemiol 34(2):188–193
25. Musunuru K et al (2010) Exome sequencing, ANGPTL3 mutations, and familial combined hypolipidemia. N Engl J Med 363(23):2220–2227
26. Ott J et al (2015) Genetic linkage analysis in the age of whole-genome sequencing. Nat Rev Genet 16:275–284
27. Purcell S et al (2005) Parental phenotypes in family-based association analysis. Am J Hum Genet 76(2):249–259
28. Schaffner SF et al (2005) Calibrating a coalescent simulation of human genome sequence variation. Genome Res 15(11):1576–1583
29. Schaid DJ et al (2013) Multiple genetic variant association testing by collapsing and kernel methods with pedigree or population structured data. Genet Epidemiol 37(5):409–418
30. Schifano ED et al (2012) SNP set association analysis for familial data. Genet Epidemiol 36(8):797–810
31. Serre D et al (2008) Correction of population stratification in large multi-ethnic association studies. PLoS ONE 3(1):e1382
32. Simpson CL et al (2011) Old lessons learned anew: family-based methods for detecting genes responsible for quantitative and qualitative traits in the Genetic Analysis Workshop 17 mini-exome sequence data. In: BMC proceedings. BioMed Central Ltd
33. Wang JL et al (2010) TGM6 identified as a novel causative gene of spinocerebellar ataxias using exome sequencing. Brain 133(12):3510–3518
34. Wijsman EM (2012) The role of large pedigrees in an era of high-throughput sequencing. Hum Genet 131(10):1555–1563
35. Wijsman EM, Amos CI (1997) Genetic analysis of simulated oligogenic traits in nuclear and extended pedigrees: summary of GAW10 contributions. Genet Epidemiol 14(6):719–735
36. Wilson AF, Ziegler A (2011) Lessons learned from Genetic Analysis Workshop 17: transitioning from genome-wide association studies to whole-genome statistical genetic analysis. Genet Epidemiol 35(S1):S107–S114
37. Wu MC et al (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93
38. Zawistowski M et al (2010) Extending rare-variant testing strategies: analysis of noncoding sequence and imputed genotypes. Am J Hum Genet 87(5):604–617
39. Zöllner S (2012) Sampling strategies for rare variant tests in case–control studies. Eur J Hum Genet 20(10):1085–1091