HHS Public Access

Author manuscript

Methods Mol Biol. Author manuscript; available in PMC 2019 January 29.

Published in final edited form as:

Methods Mol Biol. 2017 ; 1666: 527–538. doi:10.1007/978-1-4939-7274-6_26.

Detecting Multiethnic Rare Variants

Weiwei Ouyang[1], **Xiaofeng Zhu**[2], and **Huaizhen Qin**[3,4]

[1]Department of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 1610, New Orleans, LA, 70112, USA.

[2]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, 44106, USA.

[3]Department of Global Biostatistics and Data Science, Tulane University School of Public Health and Tropical Medicine, 1440 Canal Street, Suite 1610, New Orleans, LA, 70112, USA. hqin2@tulane.edu.

[4]Department of Population and Quantitative Health Sciences, Case Western Reserve University School of Medicine, Cleveland, OH, 44106, USA.

Abstract

Genome-wide association studies have identified many common genetic variants which are associated with certain diseases. The identified common variants, however, explain only a small portion of the heritability of a complex disease phenotype. The missing heritability motivated researchers to test the hypothesis that rare variants influence common diseases. Next-generation sequencing technologies have made the studies of rare variants practicable. Quite a few statistical tests have been developed for exploiting the cumulative effect of a set of rare variants on a phenotype. The best-known sequence kernel association tests (SKATs) were developed for rare variants analysis of homogeneous genomes. In this chapter, we illustrate applications of the SKATs and offer several caveats regarding them. In particular, we address how to modify the SKATs to integrate local allele ancestries and calibrate the cryptic relatedness and population structure of admixed genomes.

Keywords

Next-generation sequencing; Common disease–rare variants hypothesis; Linear mixed-effect models; Unrelated individuals; Sib pair designs; Family designs; Homogeneous population; Admixed population; Global ancestry; Local ancestry; Cryptic relatedness; Population structure

1 Introduction

1.1 Background

Initially, genome-wide association studies (GWAS) aimed to localize common genetic risk factors for complex common diseases. It was believed that a large number of genotyped samples can provide sufficient power to detect common variants which have modest effects

[3]Notes

on a phenotype. Hundreds of GWAS have been performed for mapping genetic variants of common diseases, such as hypertension, bipolar disease, coronary artery disease, diabetes, and cancer [1–3]. Such studies have successfully identified thousands of genes which are significantly associated with hundreds of traits [4] (https://www.ebi.ac.uk/gwas/). Genome-wide association studies have greatly advanced our understanding of genetic mechanisms of many common diseases.

For a common disease phenotype, however, the significant common variants identified by GWAS account for only a small fraction of the heritability observed in family studies [5]. For example, height is known to be a heritable trait with estimated heritability around 0.8 from family and twin studies, which implies about 80% of the trait variation is attributable to genetic factors. Multiple GWAS on height [3, 6–9] identified hundreds of significant common variants, which together explain only 27.4% of height variation. The missing heritability may be potentially accounted for by many rare variants [5, 10–13]. With the publication of the 1000 Genomes Project [14], we entered the era of next-generation sequencing studies. Deep sequencing technologies have been providing more comprehensive and accurate descriptions of rare variants. By directly testing rare variants in candidate genes, next-generation sequencing studies have identified many rare variant associations for a range of common diseases, e.g., type I diabetes, sterol absorption, plasma levels of LDL-C and blood pressure [15–18].

Quite a few sequence association tests have been developed for exploiting the cumulative effect of a set of rare variants on a phenotype. Prominent population-based sequence association tests include the weighted sum test [19], the C-alpha score test [20], the Estimated REgression Coefficient (EREC) test [21], the variable threshold (VT) test [22], the Sequence Kernel Association Test (SKAT) [23], the SKAT-O [24], and the Smoothed Functional Principal Component Analysis (SFPCA) method [25]. As a combination of SKAT and burden test, SKAT_O cannot always outperform the burden test or SKAT (*see* Note 1). Population-based tests can be invalid and suboptimal in the presence of familial relatedness. Rare variants may arise from recent mutations in pedigrees [26–28]. Several family-based sequence association tests were developed, e.g., the sibpair and odds ratio weighted sum tests [29, 30], the famSFPCA [31], and the famSKAT [32]. These family-based association tests require clear relatedness information or adopt a conventional kinship estimate of cryptic relatedness, e.g., KING-Robust [33]. As detailed below, SKAT and famSKAT assume linear mixed-effect models to integrate variant sets and compute significance analytically.

1.2 The Sequence Kernel Association Test (SKAT)

SKAT [23] appears to be a most prevailing method for analyzing unrelated genomes. For individual i, let y_i be the trait value, g_{ij} be the copy number of the minor allele at the jth SNP, and x_{ik} be the value of the kth covariate, $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{iL})'$ and $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{im})'$. For a quantitative trait, SKAT assumes

[1] SKAT_O [24] was developed to combine the SKAT and burden tests. However, it is not uniformly optimal and stable compared to either SKAT or the burden test. The computation of the *P*-value from the SKAT-O statistic depends on a small number of grids and thus is inefficient and inaccurate. In the presence of familial correlation, SKAT and SKAT-O cannot control type I error rates.

$$y_i = \alpha_0 + x_i'\alpha + g_i'\beta + \varepsilon_i,$$

where, and $\varepsilon_i \sim N(0; \sigma^2)$ is random error. For a binary trait (i.e., $y_i = 1$ for a case, and 0 for a control), SKAT assumes.

$$\text{logit}P(y_i = 1) = \alpha_0 + x_i'\alpha + g_i'\beta.$$

SKAT allows arbitrary sizes and directions of variant effects. For both equations, α_0 is an intercept term, $\boldsymbol{\alpha}=(\alpha_1,\ldots\alpha_m)'$ is the column vector of regression coefficients for the m covariates, and $\boldsymbol{\beta}=(\beta_1,\ldots,\beta_L)'$ is the column vector of regression coefficients for the L test variants. Let H_0:$\boldsymbol{\beta}=0$ be the null hypothesis of no association between the test variants and trait. SKAT assumes that each β_j follows an arbitrary distribution with mean 0 and variance $w_j\tau$, where $\tau(\geq 0)$ is a variance component and $w_j(\leq 0)$ is a prespecified weight for variant j. H_0: $\boldsymbol{\beta}=0$ is equivalent to H_0:$\tau=0$. SKAT utilizes the variance-component score statistic

$$Q = (y - \hat{\mu})'K(y - \hat{\mu}),$$

where $\boldsymbol{K}=\boldsymbol{GWG}'$, $\boldsymbol{G}[\boldsymbol{g}_1,\ldots,\boldsymbol{g}_N]'$, $\boldsymbol{W}=\text{diag}(w_1,\ldots,w_L)$, $\hat{\mu}_i = \hat{\alpha}_0 + \boldsymbol{x}_i'\hat{\boldsymbol{\alpha}}$ for a dichotomous trait, and $\hat{\mu}_i = \text{logit}^{-1}\left(\hat{\alpha}_0 + \boldsymbol{x}_i'\hat{\boldsymbol{\alpha}}\right)$ for a dichotomous trait, $\hat{\alpha}_0$ and $\hat{\alpha}$ are hypothesis by regressing y on the covariates only. The default weight is $\sqrt{w_j} = Beta\left(\text{MAF}_j; 1, 25\right)$ where MAF_j is the frequency of the minor allele at the jth SNP, which can be evaluated from the entire sample. The default weighting scheme is not always optimal (*see* Note 2). When analyzing admixed genomes, for example, ancestries of variants should be incorporated for better weighting schemes (*see* Note 3). For a binary trait, statistic Q collapses to the C-alpha test statistic T [20] when all the w_j are set to be 1 and all covariates are excluded.

Under the null hypothesis, Q follows a mixture of chi-square distributions. To be specific,

$$Q \sim \sum_{j=1}^{L} \lambda_j \chi_{1,j}^2,$$

2. In SKAT and famSKAT, prior information on the rare variants can be integrated by using “weights” and “sqrtweights”, respectively. SKAT is equivalent to the C-alpha test when setting “weights = rep(1, L)”, where L is the number of SNPs in the test gene. This weighting scheme is not optimal for many scenarios. The default weighting scheme in many cases is not optimal either. In real data analysis, it is not always the case that the rarer a variant is, the more important it is.

3. High-order information, e.g., the dispersion effects of variants, could be used for computing SNP-wise weights. Such a weighting scheme would be particularly informative when many true causal variants are in LD and/or there are latent interaction effects, i.e., G×G and G×E interactions. A joint location-scale test was proved under certain scenarios to improve statistical power to detect associated SNPs, genes, and pathways [45]. An alternative way to integrate the variance heterogeneity is to utilize it in the weighting scheme of SKAT and its extensions. Such a weighting scheme should be particularly useful for admixed genomes due to the large range of admixture LD. In addition, the weight of local ancestry can also incorporate prior information of ancestral prevalence. For example, if the disease has a higher prevalence in Africans, then a larger weight would be given for lower frequency of local European ancestry. It deserves formal efforts to figure out novel, effective weighting schemes for rare variants analysis of admixed genomes.

Where the $\chi^2_{1,j}$ are independent χ^2_1 variables, and the λ_j are the eigenvalues of the matrix $\boldsymbol{P}_0^{1/2}\boldsymbol{K}\boldsymbol{P}_0^{1/2}$, $\boldsymbol{P}_0 = \boldsymbol{V} - \boldsymbol{V}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{V}\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{V}$ and $X=[1,(x_1,\ldots,x_n)']$ is an $n\times(m+1)$ covariate matrix. For quantitative phenotypes, $\boldsymbol{V} = \hat{\sigma}_0^2\boldsymbol{I}, \hat{\sigma}_0^2$ is the is the estimator of σ^2 under the null hypothesis. For binary phenotypes, $\boldsymbol{V} = diag\left(\hat{\mu}_1(1-\hat{\mu}_1),\ldots,\hat{\mu}_n(1-\hat{\mu}_n)\right)$. Therefore, the P-value of Q can be closely approximated with the computationally efficient Davies' method [34]. The SKAT package depends on R version 2.13.0 or above. In Subheading 2.1, we illustrate how to install and run this package to scan homogeneous genomes.

1.3 The Family-Based SKAT

When analyzing family data, SKAT has inflated type I error if the relatedness between family members is ignored. To calibrate familial correlation, Chen et al. [32] extended SKAT to the family-based SKAT (famSKAT) for rare variant association analysis with quantitative traits in family data. Compared to SKAT, famSKAT has a different form of test statistic and null distribution, but is equivalent to SKAT when there is no familial correlation. In famSKAT, the vector of a quantitative trait is assumed to follow a linear mixed effects model

$$y = X\alpha + G\beta + \delta + \varepsilon,$$

where $\boldsymbol{X}$ is an $n\times(m+1)$ covariate matrix, $\boldsymbol{\alpha}$ is a $(m+1)\times1$ vector consisting of fixed effects parameters (an intercept and m coefficients for covariates), $\boldsymbol{G}$ is an $n\times L$ genotype matrix for L rare genetic variants of interest, $\boldsymbol{\beta}$ is an $L\times1$ vector for the random effects of rare variants, $\boldsymbol{\delta}$ is an $n\times1$ vector for the random effects of familial correlation, and $\boldsymbol{\varepsilon}$ is an $n\times1$ vector for the error. The vector of errors $\boldsymbol{\varepsilon}$ and the random effects $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ are assumed to be normally distributed and uncorrelated with each other. To be specific, we assume

$$\beta \sim \mathcal{N}(0,\tau W), \quad \delta \sim \mathcal{N}\left(0,\sigma_G^2\Phi\right), \quad \varepsilon \sim \mathcal{N}\left(0,\sigma_E^2 I\right),$$

where $\boldsymbol{W}$ is the prespecified diagonal weight matrix for the rare variants, $\boldsymbol{\Phi}$ is twice.

the kinship matrix of size $n\times n$ obtained from family information only, $\boldsymbol{I}$ is the identity.

matrix of size $n \times n$, and $\left(\tau,\sigma_G^2,\sigma_E^2\right)$ are corresponding variance component parameters. Under these assumptions, testing H_0:$\boldsymbol{\tau}$=0 versus H_1:$\boldsymbol{\tau}$>0 is equivalent to testing H_0:$\boldsymbol{\beta}$=0 versus H_1:$\boldsymbol{\beta}$ ≠ 0. The famSKAT statistic is

$$Q = \left(y - X\hat{\beta}\right)'\hat{\Sigma}^{-1}\mathrm{GW}G'\hat{\Sigma}^{-1}\left(y - X\hat{\beta}\right),$$

Where $\hat{\Sigma} = \hat{\sigma}_G^2\boldsymbol{\Phi} + \hat{\sigma}_E^2\boldsymbol{I}, \hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\hat{\Sigma}^{-1}\boldsymbol{y}$, and $\left(\hat{\sigma}_G^2,\hat{\sigma}_E^2\right)$ are maximum likelihood estimators of $\left(\hat{\sigma}_G^2,\hat{\sigma}_E^2\right)$ the null linear mixed effects model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\alpha} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}$. Under the null hypothesis,

$$Q \sim \sum_{j=1}^{L} \lambda_j z_j^2,$$

where the z_j^2 are independent χ_1^2 variables, and the λ_j are the eigenvalues of the matrix $W^{1/2} G' \hat{\Sigma}^{-1} \left(\hat{\Sigma} - X \left(X' \hat{\Sigma}^{-1} X \right)^{-1} X' \right) \hat{\Sigma}^{-1} G W^{1/2}$. Therefore, the *P*-value of Q can be closely approximated with the computationally efficient Davies' method [34]. In Subheading 2.2, we illustrate how to install and run this package to scan homogeneous genomes.

1.4 Rare Variants Analysis of Admixed Genomes

Sequence association tests were originally designed for gene-based association analysis of homogeneous genomes. These methods do not explicitly model the particular information resources or confounders of admixed genomes. Current admixed genomes are formed as various mosaics of two or more ancestral genomes. For example, genomes of African Americans often have ancestral genomic segments from Europeans and West Africans. On average, admixture regions extend over several megabases in current admixed genomes [35]. Most genetic variants have different frequencies in different ancestral populations and, thus, variants in an identical admixture region are associated with their ancestral origins—local ancestries [36]. Genetic data on admixed individuals offer distinctive advantages for localizing admixture blocks that harbor causal variants which exhibit different frequencies between ancestral populations [37]. Population structure [38] and related-ness [39] are two essential confounders in genetic association analysis of admixed genomes. Population structure is due to differences in genetic ancestry among samples; and cryptic relatedness is due to distant relatedness among samples with no known family relationships. Rare variants can show a stratification that is systematically different from, and typically stronger than, common variants [40]. Accounting for population structure is more challenging when family structure or cryptic relatedness is also present [41].

Local ancestry captures the cumulative effect on the phenotype of causal variants in the entire ancestral block. The local ancestry weighted dosage test [42] was specifically designed for identifying rare variant associations in admixed populations. However, this test only allows for unrelated subjects and a binary disease. It can be invalid and suboptimal in the presence of cryptic relatedness. Rare variant association methods have not been explicitly optimized for admixed genomes. Therefore, in Subheading 2.3, we illustrate how to modify famSKAT to perform gene-based association analysis of admixed genomes.

2 Methods

2.1 The SKAT R Package

This package aggregates individual score statistics of SNPs in a set and efficiently computes the set-level *P*-value. It requires R version 2.13.0 or above. For installation, just run *install.packages("SKAT")* and then *require(SKAT).* In this section, we illustrate how to run this package to scan homogeneous genomes. For such a purpose, we perform gene-wise association analysis on the data of genotypes of chromosome 22 and primary phenotypes

from the Genetics of Alcoholism (COGA) study. After routine quality control, the trimmed dataset comprises drinking symptoms and genotypes of 14,720 SNPs on 991 unrelated whites.

Step 1. Formatting the data.—Four plain input files are needed as detailed below. The first file is Geno_chr22.txt, containing a 991×14,720 white-spaced matrix of genotypic scores. The PLINK commands *--recode* and --recodeA can generate *Gen- oChr22.txt* from PLINK format COGA data:

```
./plink --file COGA --chr 22 --recode --out COGAChr22 --noweb
./plink --file COGAChr22 --recodeA --out GenoChr22.txt --noweb
```

Geno_chr22.txt is derived by removing the first six columns of GenoChr22.txt (FID, IID, PAT, MAT, SEX, PHENOTYPE). At each SNP, the default reference allele is the minor allele; missing genotypes are recorded as *NA*. This step recodes genotypic scores as shown below:

0	0	0	1	0	…
2	1	2	2	2	…
0	NA	2	1	2	…
…	…	…	…	…	…

The second file, Gene_chr22.txt, comprises basic information on chromosome-wide genes. Except for the header name line, each row is for one gene, including gene name, start position, end position and chromosome index:

Gene	Start	End	Chr
ACO2	40195074	40254938	22
ACR	49523517	49530592	22
ADM2	49266877	49271731	22
……			

The third file, SNP_chr22.txt, comprises basic information of the SNPs on chromosome 22:

Chr	rs_ID	Pos
22	rs2334386	14430353
22	rs2334336	14519442
22	rs12163493	14809328
……		

The last file, Pheno_Cov.txt, comprises individual trait values. The trait values are the residuals of drinking symptoms after adjusting for sex, age, and the first ten principal components (PCs). In this file, the trait values are organized one person per row:

−1.31541709
−0.095934426
−0.717596737
……

The four 4 plain files are loaded into the R platform and saved into COGA_Chr22.RDATA by running the following R command line:

```
save.image("COGA_Chr22.RData");
```

Step 2. Running the SKAT package.—The COGA_Chr22.RDATA needs to be loaded into R for gene-wise association analyses. The SKAT outputs the corresponding statistics, estimated parameters and gene-wise *P*-values (which are usually of particular interest). Running the following R command lines saves gene-wise *P*-values of the genes on chromosome 22 to array SKAT_p:

```
load("COGA_Chr22.RData");
SKAT_p<-array()
for (i in 1:dim(Gene_chr22)[1]) {
ID<-which((Geno_chr22$Pos<=Gene_chr22$End[i])&(Geno_chr22
$Pos>=Gene_chr22$Start[i]))
Z<-as.matrix(Geno_chr22[,ID])
pheno<-Pheno_Cov
obj<-SKAT_Null_Model(pheno~1, out_type="C");
SKAT_p[i]<-SKAT(Z, obj)$p.value
}
```

2.2 The famSKAT R Package

The famSKAT R code can be downloaded from https://cdn1.sphharvard.edu/wp-content/uploads/sites/1589/2014/07/famSKAT_v1.8_04052013.txt. In addition to the inputs of SKAT, the kinship matrix needs to be calculated in order to run famSKAT. This matrix can be generated by using the makekinship(.) function in the R package *kinship.* The input of this makekinship(.) function is a plain file, i.e., Fam.txt. This file comprises the family ID, individual ID, paternal ID and maternal ID of each subject, which can be extracted from the first four columns of the ped file in PLINK format, e.g., COGA.ped. Gene-wise *P*-values for chromosome 22 can be derived by running the following R code.

```
install.packages("kinship");
require(kinship);
kin_matrix<-makekinship(Fam[,1], Fam[,2], Fam[,3], Fam[,4])
```

```
famSKAT_p<-array()
for (i in 1:dim(Gene_chr22)[1]){
ID<-
which((Geno_chr22$Pos<=Gene_chr22$End[i])&(Geno_chr22 $Pos>=Gene_chr22$Start[
i]))
Z<-as.matrix(Geno_chr22[,ID])
pheno<-Pheno_Cov
famSKAT_p[i]<-famSKAT(phenotype=pheno,genotypes=Z,id=Fam[,2],
fullkins¼kin_matrix)$pvalue
}
```

2.3 Admixed Sequence Association Analysis

As in Chapter 21, we use the data on African Americans in the Maywood cohort study for the purpose of illustration.

2.3.1 Ancestry Deconvolution—As detailed in Chapter 21, SNP-wise ancestries of African American are inferred from the genotype data by the ELAI software [43]. For example, the local ancestry scores for Chromosome 22 are saved in the output file AdmWood701_Chr22.ps21.txt. Each row is for one individual. Every two columns are for one SNP. The odd columns comprise CEU local ancestry scores. For each individual, we define the global ancestry (*Global_A*) as the average of genome-wide SNP-wise European ancestry scores. We adopt global ancestry rather than PCs to represent population structure (*see* Note 4). To integrate gene-wide ancestries, we extract SNP-wise European ancestry scores and save them in file *Chr22_Local.txt*. Then, we load the Chr22_Local.txt file into R and round individual scores to 0/1/2 format:

```
Local<-read.table("Chr22_Local.txt")
Local_Chr22<-round(Local)
```

SNPs within an admixture block share identical local ancestry. We assume no ancestry-switch point within a gene. Hence, the following R code can be applied to extract local ancestry information for a test gene:

```
A<-as.matrix(Local_Chr22[, ID1])
C<-duplicated(t(A))
New<-A[, !C]
if (dim(as.matrix(New))[2]==1){
Local_A<-New
}
if (dim(as.matrix(New))[2]!=1){
Count<-array(0,dim=c(dim(as.matrix(New))[2],dim(A)[2]))
```

4. For unrelated individuals, PCs and global ancestry proportions are highly correlated surrogates for population structure. However, the PC analysis does not distinguish population structure from familial correlation. ELAI [43] can accurately infer SNP-wise local ancestries of an admixed individual, using available ancestry haplotypes from 1000 genomes as references. This algorithm is robust to familial correlation. For each individual, we define the global ancestry as half the average of the genome-wide local ancestries.

```
for (j in 1:dim(A)[2]){
for (k in 1: dim(as.matrix(New))[2]){
Count[k,j]<-ifelse(sum(A[,j]==New[,k])==701,1,0)
}
}
ID<-which(rowSums(Count)==max(rowSums(Count)))
Local_A<-New[,ID]
}
```

Here, "ID1" comprises the IDs of SNPs in the test gene. The number "701" in the *ifelse*() function is the sample size of the dataset. Due to possible inference errors, different local ancestry scores may occur within a test gene. When this is the case, we choose the local ancestry shared by most SNPs within the gene to represent the local ancestry of the gene.

2.3.2 Inference of Cryptic Relatedness—For admixed genomes, we adopt the REAP algorithm [44] to infer cryptic relatedness. This algorithm accounts for population structure and ancestry-related assortative mating. Thus it provides more accurate relatedness inference for admixed genomes (*see* Note 5). The REAP package can be freely downloaded from http://faculty.washington.edu/tathornt/software/REAP/download.html.

Four input files are needed to run the REAP package. The first two files are *Data_701_qc*.tped and *Data_701_qc*.tfam format files that can be generated by running the PLINK command:

```
./plink --file Data_701_qc --recode12 --output-missing-genotype 0
--transpose --out Data_701_qc
```

The third file, *Wood_Ancestry.txt*, is the individual ancestry file that comprises four columns. The first two columns are family ID and individual ID. The third column gives the global European ancestry proportions. The fourth column gives the global African ancestry proportions.

The fourth file, *Wood_freq.txt*, has two columns. To generate this file, a reference allele needs to be assigned for each SNP. It can be the minor allele in CEU, for example. The first column of *Wood_freq.txt* contains the frequencies of the reference alleles in European ancestry; and the second column contains the frequencies of the reference alleles in African ancestry.

Kinship can be inferred by running the following command line:

```
./REAP-gData_701_qc.tped-pData_701_qc.tfam-aWood_Ancestry.txt-
f Wood_freq.txt -k 2 -t 0.025 -r 2
```

5.The original famSKAT requires known pedigree information to properly control the type I error rate and thus it cannot be directly applied to account for cryptic relatedness between admixed genomes. Conventional methods such as KING-Robust [33] provide accurate relatedness inference for the data of homogeneous genomes. Such methods, however, cannot appropriately accounts for population structure in admixed genomes. The REAP algorithm [44] appropriately accounts for population structure and ancestry-related assortative mating and thus provides more accurate relatedness inference for admixed genomes.

One of the output files, REAP_Kincoef_matrix.txt, contains the kinship matrix information.

2.3.3 Running famSKAT—After deriving individual global ancestries, local ancestries, and the kinship matrix, running the following R code yields the *P*-values for genes on chromosome 22:

```
kin_matrix<- read.table("REAP_Kincoef_matrix.txt")
famSKAT_p<-array()
for (i in 1:dim(Gene_chr22)[1]){
ID<-which((Geno_chr22$Pos<=Gene_chr22$End[i])&(Geno_chr22
$Pos>=Gene_chr22$Start[i]))
Z<- as.matrix(cbind(Geno_chr22[,ID], Local_A))
pheno<-Pheno_Cov
famSKAT_p[i]<-famSKAT(phenotype=pheno,genotypes=Z,id=Fam
[,2],fullkins=kin_matrix, covariates=Global_A)$pvalue
}
```

In such an analysis, gene-wide local ancestry is taken as a surrogate for the cumulative effect of the gene-wide variants; the global ancestry is adjusted as a fixed effect, and the relatedness is adjusted as a random effect.

Acknowledgments

This work was funded in part by NIH grant HG003054 to X.Z. and by Tulane's Committee on Research fellowship (600890) and Carol Lavin Bernick Faculty Grant (632119) to H.Q.

References

1. Burton PR, Clayton DG, Cardon LR et al. (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447 (7145):661–678 [PubMed: 17554300]
2. Heid IM, Jackson AU, Randall JC et al. (2010) Meta-analysis identifies 13 new loci associated with waist-hip ratio and reveals sexual dimorphism in the genetic basis of fat distribution. Nat Genet 42(11):949–960 [PubMed: 20935629]
3. Lango Allen H, Estrada K, Lettre G et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. Nature 467(7317):832–838 [PubMed: 20881960]
4. Hindorff LA, Junkins HA, Hall P, et al. (2011) A catalog of published genome-wide association studies http://www.genome.gov/26525384
5. Manolio TA, Collins FS, Cox NJ et al. (2009) Finding the missing heritability of complex diseases. Nature 461(7265):747–753 [PubMed: 19812666]
6. Gudbjartsson DF, Walters GB, Thorleifsson G et al. (2008) Many sequence variants affecting diversity of adult human height. Nat Genet 40 (5):609–615 [PubMed: 18391951]
7. Lettre G, Jackson AU, Gieger C et al. (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. Nat Genet 40(5):584–591 [PubMed: 18391950]
8. Weedon MN, Lango H, Lindgren CM et al. (2008) Genome-wide association analysis identifies 20 loci that influence adult height. Nat Genet 40(5):575–583 [PubMed: 18391952]
9. Wood AR, Esko T, Yang J et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 46(11):1173–1186 [PubMed: 25282103]

10. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69(1):124–137 [PubMed: 11404818]
11. Zuk O, Hechter E, Sunyaev SR et al. (2012) The mystery of missing heritability: genetic interactions create phantom heritability. Proc Natl Acad Sci U S A 109(4):1193–1198 [PubMed: 22223662]
12. Gorlov IP, Gorlova OY, Sunyaev SR et al. (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. Am J Hum Genet 82(1):100–112 [PubMed: 18179889]
13. Cirulli ET, Goldstein DB (2010) Uncovering the roles of rare variants in common disease through whole-genome sequencing. Nat Rev Genet 11(6):415–425 [PubMed: 20479773]
14. Consortium GP (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073 [PubMed: 20981092]
15. Cohen J, Pertsemlidis A, Kotowski IK et al. (2005) Low LDL cholesterol in individuals of African descent resulting from frequent nonsense mutations in PCSK9. Nat Genet 37 (2):161–165 [PubMed: 15654334]
16. Cohen JC, Pertsemlidis A, Fahmi S et al. (2006) Multiple rare variants in NPC1L1 associated with reduced sterol absorption and plasma low-density lipoprotein levels. Proc Natl Acad Sci U S A 103(6):1810–1815 [PubMed: 16449388]
17. Ji W, Foo JN, O'Roak BJ et al. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. Nat Genet 40(5):592–599 [PubMed: 18391953]
18. Nejentsev S, Walker N, Riches D et al. (2009) Rare variants of IFIH1, a gene implicated in antiviral responses, protect against type 1 diabetes. Science 324(5925):387–389 [PubMed: 19264985]
19. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. PLoS Genet 5(2): e1000384 [PubMed: 19214210]
20. Neale BM, Rivas MA, Voight BF et al. (2011) Testing for an unusual distribution of rare variants. PLoS Genet 7(3):e1001322 [PubMed: 21408211]
21. Lin D-Y, Tang Z-Z (2011) A general frame-work for detecting disease associations with rare variants in sequencing studies. Am J Hum Genet 89(3):354–367 [PubMed: 21885029]
22. Price AL, Kryukov GV, de Bakker PI et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. Am J Hum Genet 86(6):832–838 [PubMed: 20471002]
23. Wu MC, Lee S, Cai T et al. (2011) Rare-variant association testing for sequencing data with the sequence kernel association test. Am J Hum Genet 89(1):82–93 [PubMed: 21737059]
24. Lee S, Emond MJ, Bamshad MJ et al. (2012) Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. Am J Hum Genet 91(2):224–237 [PubMed: 22863193]
25. Luo L, Zhu Y, Xiong M (2012) Quantitative trait locus analysis for next-generation sequencing with the functional linear models. J Med Genet 49(8):513–524 [PubMed: 22889854]
26. Lupski JR, Belmont JW, Boerwinkle E et al. (2011) Clan genomics and the complex architecture of human disease. Cell 147(1):32–43 [PubMed: 21962505]
27. Najmabadi H, Hu H, Garshasbi M et al. (2011) Deep sequencing reveals 50 novel genes for recessive cognitive disorders. Nature 478 (7367):57–63 [PubMed: 21937992]
28. Chakravarti A (2011) Genomics is not enough. Science 334(6052):15 [PubMed: 21980079]
29. Feng T, Elston RC, Zhu X (2011) Detecting rare and common variants for complex traits: sibpair and odds ratio weighted sum statistics (SPWSS, ORWSS). Genet Epidemiol 35 (5):398–409 [PubMed: 21594893]
30. Zhu X, Feng T, Li Y et al. (2010) Detecting rare variants for complex traits using family and unrelated data. Genet Epidemiol 34(2):171–187 [PubMed: 19847924]
31. Zhu Y, Xiong M (2012) Family-based association studies for next-generation sequencing. Am J Hum Genet 90(6):1028–1045 [PubMed: 22682329]
32. Chen H, Meigs JB, Dupuis J (2013) Sequence kernel association test for quantitative traits in family samples. Genet Epidemiol 37 (2):196–204 [PubMed: 23280576]
33. Manichaikul A, Mychaleckyj JC, Rich SS et al. (2010) Robust relationship inference in genome-wide association studies. Bioinformatics 26(22):2867–2873 [PubMed: 20926424]

34. Davies RB (1980) The distribution of a linear combination of x2 random variables. Appl Stat 29(3):323–333
35. Smith MW, O'Brien SJ (2005) Mapping by admixture linkage disequilibrium: advances, limitations and guidelines. Nat Rev Genet 6 (8):623–632 [PubMed: 16012528]
36. Qin H, Morris N, Kang SJ et al. (2010) Interrogating local population structure for fine mapping in genome-wide association studies. Bioinformatics 26(23):2961–2968 [PubMed: 20889494]
37. Qin H, Zhu X (2012) Power comparison of admixture mapping and direct association analysis in genome-wide association studies. Genet Epidemiol 36(3):235–243 [PubMed: 22460597]
38. Price AL, Patterson NJ, Plenge RM et al. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38(8):904–909 [PubMed: 16862161]
39. Yu J, Pressoir G, Briggs WH et al. (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38(2):203–208 [PubMed: 16380716]
40. Mathieson I, McVean G (2012) Differential confounding of rare and common variants in spatially structured populations. Nat Genet 44 (3):243–246 [PubMed: 22306651]
41. Price AL, Zaitlen NA, Reich D et al. (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7):459–463 [PubMed: 20548291]
42. Mao X, Li Y, Liu Y et al. (2013) Testing genetic association with rare variants in admixed populations. Genet Epidemiol 37(1):38–47 [PubMed: 23032398]
43. Guan Y (2014) Detecting structure of haplotypes and local ancestry. Genetics 196 (3):625–642 [PubMed: 24388880]
44. Thornton T, Tang H, Hoffmann TJ et al. (2012) Estimating kinship in admixed populations. Am J Hum Genet 91(1):122–138 [PubMed: 22748210]
45. Soave D, Corvol H, Panjwani N et al. (2015) A joint location-scale test improves power to detect associated SNPs, gene sets, and path-ways. Am J Hum Genet 97(1):125–138 [PubMed: 26140448]