

Leveraging Population Information in Family-based Rare Variant Association Analyses of Quantitative Traits

by

Yu Jiang

Department of Statistical Science
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Edwin S Iversen

Andrew S Allen

Thesis submitted in partial fulfillment of the requirements for the degree of
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2015

ABSTRACT

Leveraging Population Information in Family-based Rare
Variant Association Analyses of Quantitative Traits

by

Yu Jiang

Department of Statistical Science
Duke University

Date: _____

Approved:

Sayan Mukherjee, Supervisor

Edwin S Iversen

Andrew S Allen

An abstract of a thesis submitted in partial fulfillment of the requirements for
Master of Science in the Department of Statistical Science
in the Graduate School of Duke University
2015

Copyright © 2015 by Yu Jiang
All rights reserved except the rights granted by the
Creative Commons Attribution-Noncommercial Licence

Abstract

Confounding due to population substructure is always a concern in genetic association studies. While methods have been proposed to adjust for population stratification in the context of common variation, it is unclear how well these approaches will work when interrogating rare variation. Family-based association tests can be constructed that are robust to population stratification. For example, when considering a quantitative trait, a linear model can be used that decomposes genetic effects into between and within-family components and a test of the within-family component is robust to population stratification. However, this within-family test ignores between-family information potentially leading to a loss of power. Here, we propose a family-based two-stage rare-variant test for quantitative traits. We first construct a weight for each variant within a gene, or other genetic unit, based on score tests of between-family effect parameters. These weights are then used to combine variants using score tests of within-family effect parameters. Since the between-family and within-family tests are orthogonal under the null hypothesis, this two-stage approach can increase power while still maintaining validity. Using simulation, we show that this two-stage test can significantly improve power while correctly maintaining type I error. We further show that the two-stage approach maintains the robustness to population stratification of the within-family test and we illustrate this using simulations reflecting samples comprised of continental and closely related subpopulations.

Contents

Abstract	iii
List of Tables	vi
List of Figures	vii
List of Abbreviations and Symbols	viii
Acknowledgements	ix
1 Introduction	1
2 Methods	5
2.1 General framework	5
2.1.1 Efficient score at single locus.	5
2.1.2 Combining variants in a gene.	7
2.2 Simulation strategy	8
2.2.1 Homogenous population	8
2.2.2 Continental population stratification.	9
2.2.3 Closely related subpopulations.	10
3 Results	11
3.1 Type I error rates and Power comparison	11
3.2 Test under population stratification	12
3.2.1 Continental population stratification.	12
3.2.2 Closely related subpopulations.	13

3.3 Gene screening with between family component	16
4 Discussion	19
Bibliography	21

List of Tables

3.1	Simulation results based on 1000 replicates at nominal level 0.05 in homogenous populations	12
3.2	Simulations When correlation between MAF and effect size are destroyed, $geneProp = 0.01$	12
3.3	Simulations under different sample size	13
3.4	Simulation results based on 1000 replicates at nominal level 0.05 under population stratification (Continental populations, with $F_{st} = 0.1$, $nfam = 2000$, $protect = 0$, $geneProp = 0.05$)	17
3.5	Simulation results based on 1000 replicates at nominal level 0.05 under population stratification (European populations)	18
3.6	Type I error rates when screening with between-family components	18

List of Figures

3.1	QQ-plot for analysis when population stratification exist	14
3.2	QQ-plot for closely related population	15
3.3	Power comparison when screening with between-family components .	17

List of Abbreviations and Symbols

Symbols

\mathcal{I}_{ab}	Information matrix respect to parameter a, b
\mathcal{S}_a	Efficient score statistics respect to parameter a
X	covariants used in the analysis
y	offspring phenotype

Abbreviations

GWAS	Genome wide association study.
TDT	Transmission disequilibrium test.
SKAT	Sequence kernel association test, developed by Wu et al.
QTL	Quantitative trait locus
MAF	Minor allele frequency

Acknowledgements

It is my great pleasure to acknowledge my advisor Dr. Sayan Mukherjee and committee members: Dr. Edwin Iverson and Dr. Andrew Allen for their help on finishing this thesis.

1

Introduction

In the past decades, Genome wide association studies (GWAS) have been successfully identify many phenotype associated genetic variants. These variants have often informed on phenotype related biologic processes. However, this variants can only explain a small proportion of genetic heritability. For example, twins studies estimates that the narrow sense heritability of human height is 0.7-0.9 (Macgregor et al., 2006; Silventoinen et al., 2003; Visscher, 2008). However, height associated variants identified from GWAS can only explain about 5% of the phenotypic variance (Visscher, 2008; Gudbjartsson et al., 2008; Weedon et al., 2008; Lettre et al., 2008). Since only common tagSNPs ($MAF \geq 5\%$) are presented in GWAS analysis, rare variants are proposed to account for this missing heritability (Cirulli and Goldstein, 2010; Gibson, 2012). As next-generation sequencing technologies become economical, many research groups are transitioning to whole genome or whole exome sequencing as their primary approach to measuring genetic variation, which allows the possibility to comprehensively detect rare variation.

One of the concerns of rare variants analysis is confounding due to population

stratifications. Population stratification refers to the situation where the allele frequencies vary across subpopulations. When the phenotype trait also varies by subpopulation, perhaps due to non-genetic factors such as environment, Association studies can be confounded, resulting in increased type I error rate. Many methods, such as genomic control (Reich and Goldstein, 2001; Zheng et al., 2006; Devlin and Roeder, 1999), principle component analysis (Price et al., 2006), have been proposed to correct population stratifications. These methods work well for common variants. Compared to common variants, rare variants occur more recently so they are more likely to be population specific and more likely to have local substructure, resulting in a systematic different and stronger stratification impact than common variants. As shown by Mathieson and McVean (Mathieson and McVean, 2012), these approaches to correcting population stratification for common variants may fail in the context of rare variation.

Transmission Disequilibrium Test (TDT) proposed by Spielman et al (Spielman et al., 1993) based on case-parent trios are robust to population stratification, making it a great choice for test of rare variation. However, TDT is designed for dichotomous phenotype. To extend this robust test to the analysis of quantitative traits, several different versions of quantitative TDT are proposed (Abecasis et al., 2000; Allison, 1997; Zhu and Elston, 2001; Sun et al., 2000). Among these methods, Abecasis et al, constructed the test by decomposing the offspring's genotype into two orthogonal component: between-family and within-family component and they show that the within-family component is free of confounding due to population stratification. However, this method is a single locus test and as such it is only appropriate as a test of common variants for two reasons. First, the asymptotic approximation will only be valid when there are an appropriate number of heterozygous parents. Second, even if one used an exact test to maintain the type I error rate, the test will still be under-

powered. Therefore, this method cannot be directly applied to the analysis of rare variants. In the analysis of rare variation in case-control studies, a common strategy is to accumulate information across variants located in the same gene or pathway to improve the test power, for example, the burden test proposed by Morris and Zeggini (Morris and Zeggini, 2010) and the kernel test SKAT (Wu et al., 2011). Both De et al. (De et al., 2013) and Ionita-Laza et al. (Ionita-Laza et al., 2013) utilized the same concept in case-control studies to constructed the rare variants test based on the within-family component. After computing the score statistics respecting to the coefficient of each variants, they accumulated this score statistics with different methods: De et al., combined the score statistics of different loci for each family first while Ionita-Laza et al, combined the score statistics of different families at each locus first.

However, both methods discard the between family information, leading to the loss of power. Jiang et al (Jiang et al., 2014a) proposed an method to screen the possible causal genes with the between family genotype before the test based on within-family component. They show that this method can increase the statistical power while maintain the validity of test comparing to methods using only within family information. Besides screening at the gene level, the between family information can also be used to screen at variants level. Jiang et al. (Jiang et al., 2014b) proposed a weighting scheme in the TDT of case-parent analysis, by comparing the frequency of each variants at population controls and parents. This weighting scheme upweighted variants that have significant different frequencies in parents and offsprings, which are more likely to disease associated variants. Since the weight scheme and the TDT are orthogonal, this method can also improve the statistical power while maintaining the test validity. Considering the orthogonality between the between-family component and within-family component, we adopted a similar weighting scheme for the quantitative trait test but using the between family genotype to construct the

weights for each variant instead of information from external source of controls.

In this thesis, we presented a two-stage test for quantitative traits with the orthogonally decomposed genotype in offspring-parent trios. In chapter 2, we describe our method in details. In the first stage, we constructed linear model for offspring's phenotype and the between-family component of their genotype. For each variant, we computed a score statistics respected to their regression coefficients. In the second stage we constructed the linear model based on the within-family components and also computed the score statistics for each variant. Either De's method or Ionita-Laza's method can be used to combine the score statistics for variants across a gene. During the combining step, the scores computed in the first stage are used as the weights for each variant.

As shown in chapter 3, this weighting scheme is orthogonal to the test constructed by within-family component, thus the type I error rate will be correctly maintained. The simulation results also show that this weighting scheme can significantly improve the statistical power when the proportion of causal variants is small. When confounding factors due to population stratification exist, the test constructed from the within-family component can always maintain the correct type I error rate no matter whether principal component analysis can correct the confounding or not. Due to the orthogonality of the weights, our weighting scheme will not influence this robustness. In chapter 4, we discuss the simulation results and also proposed several future aspects to improve current method.

2

Methods

2.1 General framework

2.1.1 *Efficient score at single locus.*

We used the method in Abecasis et al (2000) to decompose each variant for each trio: offspring's genotype are decamped to between family component (b) and within family component (w). For i -th family at locus j , $b_{ij} = \frac{g_{ijF} + g_{ijM}}{2}$ and $w_{ij} = g_{ijO} - b_{ij}$, where $g_{ijO,M,F}$ represent the genotype of offspring, mother and father's genotype of i -th family at locus j respectively.

At the first stage, we constructed a linear regression model using offspring's genotype and the between-family component.

$$y_i = \alpha + \beta_b b_{ij} + \gamma^T X_i + \epsilon$$

where y_i is the phenotype of offspring in i -th family, and X_i is a vector of covariates. We assume each offspring are independent of each other and $\epsilon \sim N(0, \sigma^2)$.

The the contribution of each trio to the efficient score statistics respect to β_b ,

$$S_{b,ij} = S_{\beta_b, i}^0 - I_{\beta_b \eta}^0 (I_{\eta \eta}^0)^{-1} S_{\eta, i}^0 \quad (2.1)$$

where $\eta = [\alpha, \gamma^T]^T$ represent all covariants. Note, since σ^2 is cancelled in the following test, we do not take σ^2 as a nuisance parameter. $S_{\beta_b, i}^0$ and $S_{\eta, i}^0$ are the first derivative of the logarithm of the likelihood function respect to β_b and η under the null hypothesis, and $I_{\beta_b \eta}$ and $I_{\eta \eta}^0$ are components of the information matrix under null hypothesis. This information matrix are estimated empirically.

Plugging the estimated information matrix into equation 2.1,

$$S_{b, ij} = \frac{b_{ij}}{\sigma^2} (y_i - \mu_i) - I_{\beta_b \eta}^0 (I_{\eta \eta}^0)^{-1} S_{\eta, i}^0$$

where μ_i is the fitted phenotype value under the null hypothesis $\beta_b = 0$. The efficient score for each locus can be computed by summing up each family's contribution $S_{bj} = \sum_{i=1}^n S_{b, ij}$. This efficient score S_{bj} will be used as the weights of locus j when combining all variants in a gene.

At the second stage, the test are constructed based on the within-family component:

$$y_i = \alpha + \beta_w w_{ij} + \gamma^T X_i + \epsilon$$

This model is the same as Ionita-laza and De et al. The null hypothesis is $\beta_w = 0$. Similar to the computation in the first step, we computed each family's contribution to the efficient score statistics respect to β_w . Since the within-family component is orthogonal to the between-family component and all other covariates, $S_{w, ij}$ has a simple form,

$$S_{w, ij} = \frac{w_{ij}}{\sigma^2} (y_i - \mu_i)$$

This efficient score will be used to constructed the test for the association between gene and phenotype.

2.1.2 Combining variants in a gene.

Test for single locus are often underpowered when applying to rare variants. A common strategy to improve the test power is to accumulate variants in a gene (or a genomic region). Two accumulation techniques are possible under our test framework. Assume there are total m variants in a gene, we first using the method adopted by Ionita-laza et al., i.e.,

$$T_{SKAT} = \sum_{j=1}^m c_j^2 \left(\sum_{i=1}^n S_{w,ij} \right)^2$$

where c_j is the weight of locus j , in the test developed by Ionita-laza et al., $c_j = \text{Beta}(\hat{f}_j; 1, 25)$ with \hat{f}_j being the estimated MAF of variant j based on parental genotypes, such that rare variants can be up-weighted.

In our test, we use the efficient score computed in the first step as the weights, thus,

$$T_{SKAT} = \sum_{j=1}^m S_{b,j}^2 \left(\sum_{i=1}^n S_{w,ij} \right)^2$$

This statistics can be represented by quadratic form,

$$T_{SKAT} = \mathbf{1}_n^T S_w \Lambda S_w^T \mathbf{1}$$

where S_w is a $n \times m$ matrix with $S_w[i, j] = S_{w,ij}$ and Λ is a diagonal matrix with $\Lambda[j, j] = S_{b,j}^2$ the efficient score based on within-family component is asymptotically distributed as a normal distribution with mean 0 under the null hypothesis, so T_{SKAT} can be approximated to a mixture of χ^2 distribution, i.e. $T_{SKAT} \sim \sum_{k=1}^m \lambda_k \chi^2(1)$, where λ_m is the eigenvalue of $V = \text{cov}(S_w) \Lambda$, Davie's method (Davies, 1980) are used to approximate the p-value from this mixture of χ^2 distribution.

We can also construct the test with the method in De et al.,

$$W_{burden} = \sum_{i=1}^n \sum_{j=1}^m c_j S_{w,ij}$$

It is reasonable to assume that the efficient score of each family are independently identically distributed with mean 0, thus, the variance of W can be estimated empirically.

$$T_{burden} = \frac{\left[\sum_{i=1}^n \sum_{j=1}^m c_j S_{w,ij} \right]^2}{\sum_{i=1}^n \left[\sum_{j=1}^m c_j S_{w,ij} \right]^2}$$

similarly to the kernel test, the weights can be 1 such that every variants are equally weighted or weighted by the MAF, e.g., $c_j = \text{Beta}(\hat{f}_j; 1, 25)$. In our case, we still use the efficient score computed through the between-family components, then

$$T_{burden} = \frac{\left[\sum_{i=1}^n \sum_{j=1}^m S_{b,j} S_{w,ij} \right]^2}{\sum_{i=1}^n \left[\sum_{j=1}^m S_{b,j} S_{w,ij} \right]^2}$$

According to the central limit theorem, T_{burden} asymptotically distributed as χ^2 distribution with degree-of-freedom 1.

2.2 Simulation strategy

2.2.1 Homogenous population

We perform a simulation study to evaluate the type I error rates and to compare the power with other weighting schemes. In the simulation, we generated a 10000 30kbp haplotype pool with COSI, which mimic the genomic distribution of European population. We randomly select 5 subregions with a total length 1.5kb to represent the captured region in exome sequencing. During each simulation, we first randomly

picked four haplotypes from the pool to generate parents haplotype, and then randomly selected one haplotype from each parent to form offspring’s genotype, assuming no crossover within a gene. The offspring’s phenotype are then simulated based on the simulation scheme in SeqSIMLA (Chung and Shih, 2013). Specifically, we fixed the proportion of genotypic variance explained by the gene (V_g) and then randomly picked up some as QTLs from variants with MAF less than 0.03 in the capture regions. We assume each variant explain the same proportion of genetic variance from V_g unless specifically stated. With the additive model in SeqSIMLA we compute the genotypic value $a_k = \sqrt{\frac{V_k}{2p_k(1-p_k)}}$ for each causal variant where V_k is the variance explained by causal variant k, $V_k = \frac{V_g}{N_c}$, N_c is the total number of causal variants in the gene. Therefore, the phenotype of the offspring $y = \mu + \sum_{k=1}^K a_k(g_{kO} - 1) + P + E$, where μ is a user specific mean in populations, we set $\mu = 0$; P follows a normal distribution with mean 0 and variance being the polygenic variance V_{poly} and E follows a normal distribution with mean 0 and variance being environment variance, $V_E = V_{tot} - V_g - V_{poly}$. In most simulation setting, we set $V_g = 0.05V_{tot}$ and $V_{poly} = 0.15V_g$. Repeating this process until 2000 trios are generated in each simulation set.

2.2.2 Continental population stratification.

To induce confounders due to population stratification, we simulated two haplotype pools with COSI, mimicking the genome feature of European and African populations, and then generated samples from these two population pools separately with the same model described above but allowing different phenotypic mean (μ) in two populations. Besides variants that locate in the gene in the analysis, we generated 100 common variants to be used to perform PCA. These common variants are generated from the Balding-Nichols model (Balding and Nichols, 1995). Specifically, we

generated the ancestry allele frequency (p) for each variants from a uniform distribution with $\min = 0.1$ and $\max = 0.3$, and then the allele frequency in two populations (p_1, p_2) are generated through a beta distribution ($\frac{1-F_{st}}{F_{st}}p, \frac{1-F_{st}}{F_{st}}(1-p)$). For samples generated from population 1, we simulated the genotype for the common variants from a binomial distribution with mean p_1 , similarly, generated genotype for samples in population 2 with mean p_2 . During these simulation, we assume all these 100 common variants are independent of each other. In the analysis step, these variants are used to perform the PCA and then the top 10 components are included in the analysis as covariates.

2.2.3 Closely related subpopulations.

We also examined the performance of all these methods when subtle population structure exist. We first generated a haplotype pool with a total length 10 Mbp, calibrating to mimic the European populations with COSI. We randomly picked 5 variants with minor allele frequency greater than 0.4 from all genetic variants in the pool and then divided the haplotype pool to 32 subgroups based on the possible combinations of these 5 variants. To mimic the exome sequencing studies, we randomly picked a 30 Kbp region to represent the gene used in simulation analysis for this 10Mbp haplotype and then randomly picked 5 subregions with total length 1.5 Kbp from this 30 Kbp region to represents the captured regions in exome sequencing. Besides these variants, we also picked the common variants used in PCA by selecting one variants from every 100 common variants which has a minor allele frequency in $[0.1, 0.3]$. Following the same procedure described in 2.2.1 generated samples from each subgroup, during simulation, different phenotypic mean (μ) were used in different subgroups. We generated a total 2000 trios in each simulation dataset and in the analysis the top 10 principal components are included as covariates.

3.1 Type I error rates and Power comparison

To evaluate the performance of the weighting scheme, we compared the power of our weighting scheme to others. Table 3.1 and Table 3.2 summarized the the type I error rates and power of all methods under different simulation scenarios. The weighting scheme did not impact the validity of the test, all type I error rates were maintained in the correct level. As can be seen from Table 3.1, when the proportion of causal variants is small (less than 40%), our weighting scheme can significantly improve the test power, for example, when 10% of variants are causal variants in the gene, our weighting scheme can increase about 50% of the power for the burden test comparing the test with uniform weights,i.e., increased from 0.254 to 0.388 and about 40% comparing to that weighted by MAF. Similar performance was also observed in the kernel test. Moreover, our weighting scheme is not based on any assumption on the relation between MAF and effect size of the variants, thus, as shown in Table 3.2, the power increase can still be preserved when we changed the simulation scheme, but weighting based on MAF can lead of loss of power.

Table 3.1: Simulation results based on 1000 replicates at nominal level 0.05 in homogenous populations

causal	nfam	varP	geb	ges	bu	bm	bs	su	sm	ss
0	2000	0	0.045	0.047	0.042	0.036	0.043	0.038	0.049	0.042
0.1	2000	0.048	0.308	0.31	0.254	0.273	0.388	0.272	0.308	0.365
0.2	2000	0.050	0.502	0.482	0.39	0.466	0.59	0.382	0.463	0.531
0.3	2000	0.050	0.718	0.626	0.561	0.637	0.698	0.453	0.526	0.577
0.4	2000	0.050	0.813	0.674	0.68	0.77	0.737	0.473	0.553	0.554

Table 3.2: Simulations When correlation between MAF and effect size are destroyed, geneProp = 0.01

causal	nfam	geb	ges	bu	bm	bs	su	sm	ss
0	2000	0.043	0.048	0.042	0.036	0.041	0.038	0.049	0.043
0.1	2000	0.216	0.246	0.138	0.136	0.27	0.191	0.226	0.282
0.2	2000	0.348	0.432	0.248	0.234	0.447	0.303	0.376	0.461
0.3	2000	0.538	0.631	0.351	0.332	0.586	0.482	0.542	0.615
0.4	2000	0.634	0.726	0.449	0.42	0.657	0.548	0.598	0.663

Table 3.3 presents the correlation between sample size and the increase of power. As can be seen, with the increase of the sample size, we can observe more significant power increase by comparing the power of test under our weighting scheme to that uniformly weighted. This is most likely because that the larger sample size can help more accurately estimate the weights for each variants at the first stage.

3.2 Test under population stratification

3.2.1 Continental population stratification.

We evaluated the performance of our weighting scheme when population stratification exist. In this scenario, we also evaluated the performance of population based methods. As can be see from Figure 3.1, the type I error rates of population based methods are inflated while that of our method can still be maintained without PCA.

Table 3.3: Simulations under different sample size

nfam	causal	varP	geb	ges	bu	bm	bs	su	sm	ss
500	0.2	0.048	0.224	0.17	0.141	0.142	0.145	0.111	0.1	0.119
1000	0.2	0.049	0.341	0.342	0.249	0.263	0.344	0.225	0.237	0.295
2000	0.2	0.050	0.502	0.482	0.39	0.466	0.59	0.382	0.463	0.531
3000	0.2	0.050	0.595	0.562	0.489	0.572	0.703	0.461	0.582	0.661
5000	0.2	0.050	0.739	0.665	0.615	0.74	0.834	0.561	0.722	0.791

After PCA adjustment, the type I error rates of population can be controlled under this simulation setting. Table 3.4 show the power of different methods when population stratification exist. The statistical power of the weighting scheme is reduced when the estimation of the weights was biased without PCA, but this disadvantage is eliminated after adjusted by PCA. As can be seen, after PCA adjustment, our weighting scheme still have the highest power among three weighting scheme for both burden test and kernel test.

3.2.2 Closely related subpopulations.

As Mathieson and McVean (Mathieson and McVean, 2012) pointed out when the population structure are more subtle, the PCA failed to correct the inflated type I error rates in the population based test. As can be seen from Figure 3.2, both population based burden test and kernel test show an obvious inflated type I error rate even when adjusted by PCA. However, tests based on the within-family component and weighted by between-family components maintain the correct type I error rate whether PCA adjustment used. In this simulation setting, Our weighting scheme still maintain a reasonable power even though the efficient score computed in the first stage may be biased by the unadjustable population stratification as can be seen from Table 3.5

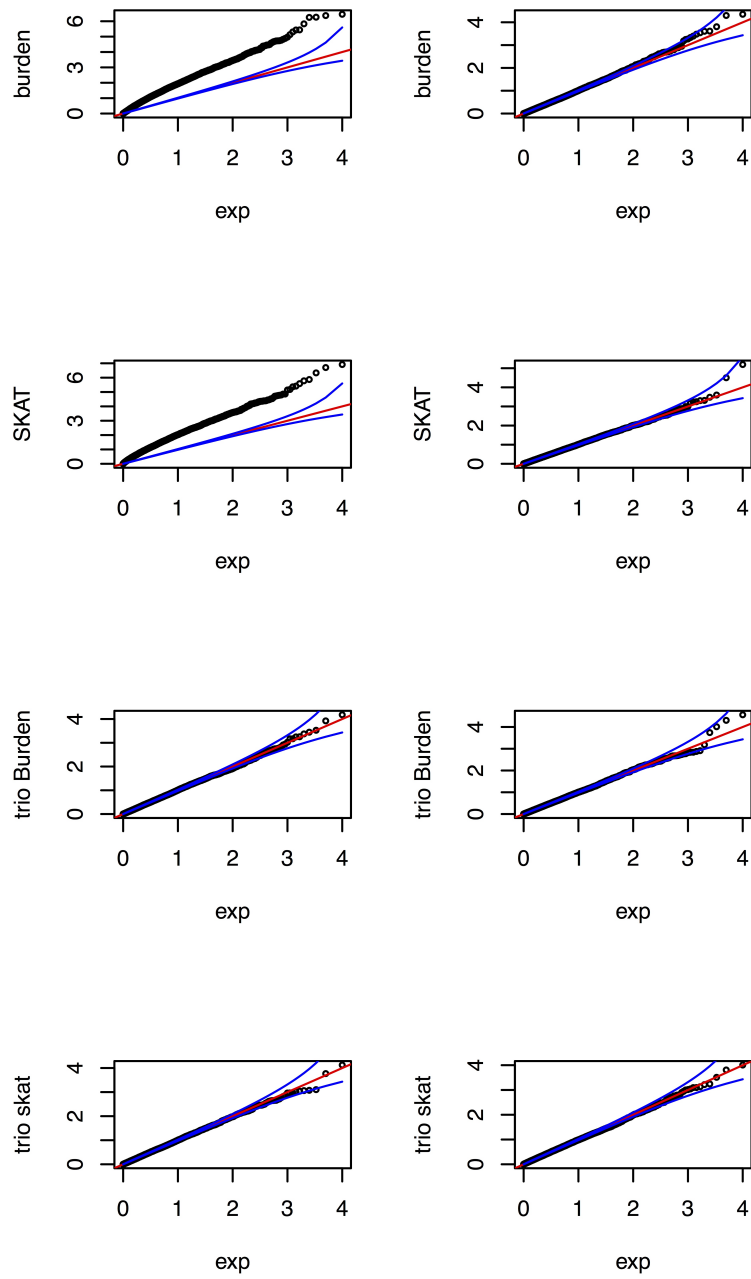


FIGURE 3.1: QQ-plot for analysis when population stratification exist

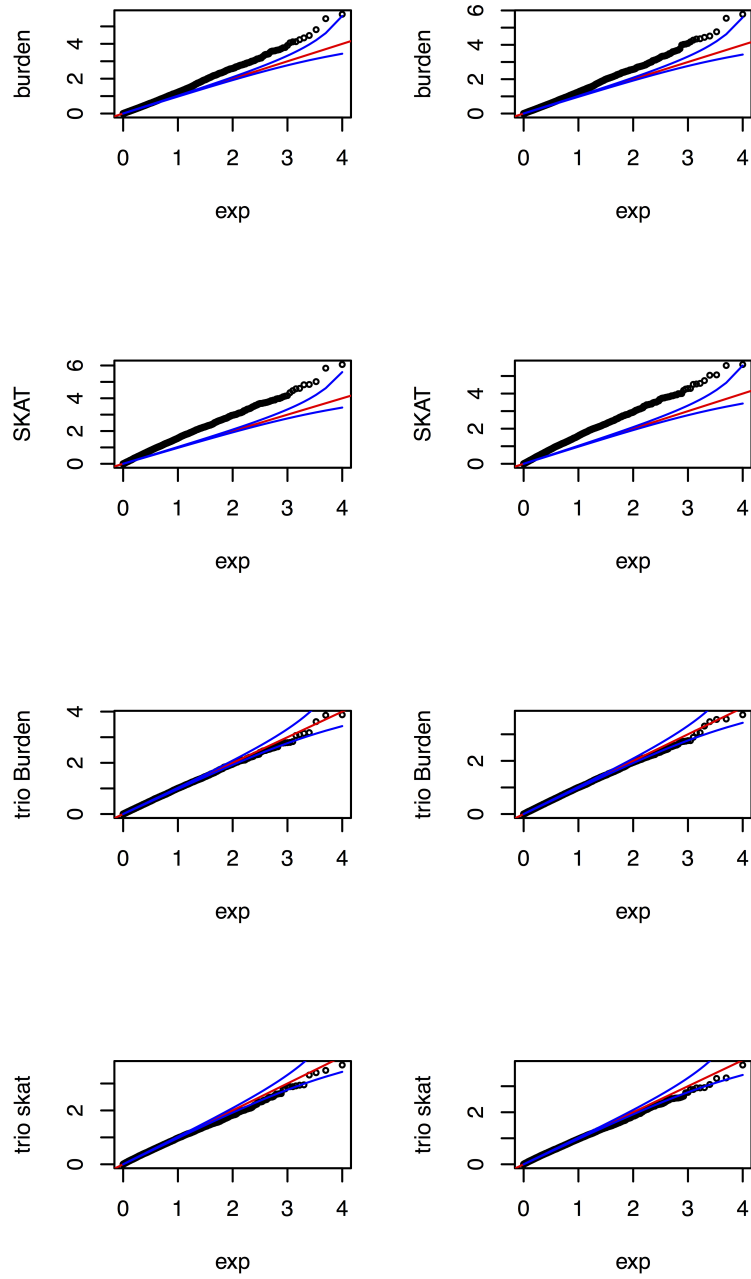


FIGURE 3.2: QQ-plot for closely related population

3.3 Gene screening with between family component

We evaluate if the screening procedure proposed by Jiang et al. (Jiang et al., 2014a) can be integrated to our analysis. In this setting, we generate 10 genes and randomly selected one as causal genes in the power comparison. Bonferroni correction are used in the multiple gene analysis. In this analysis, the between-family component are used twice, we first evaluate if the type I error rate can be maintained. As can be seen from Table 3.6, the type I error rate of test after screening are well controlled and the type I error rate is preserved after Bonferroni adjustment.

We then compare the power of test with screening to that without screening. As can be seen From Figure 3.3, the test power can be improve from two ways. First, comparing the test without screening, i.e., bs and bm in the plot, test weighted by between-family component have a higher power across different causal proportion. Second, comparing the test which including both screening procedure and weighting scheme, bs-screen, to those only included screening procedure (gm-screen) or only included weighting scheme (bs), we can found there is also significant power increase.

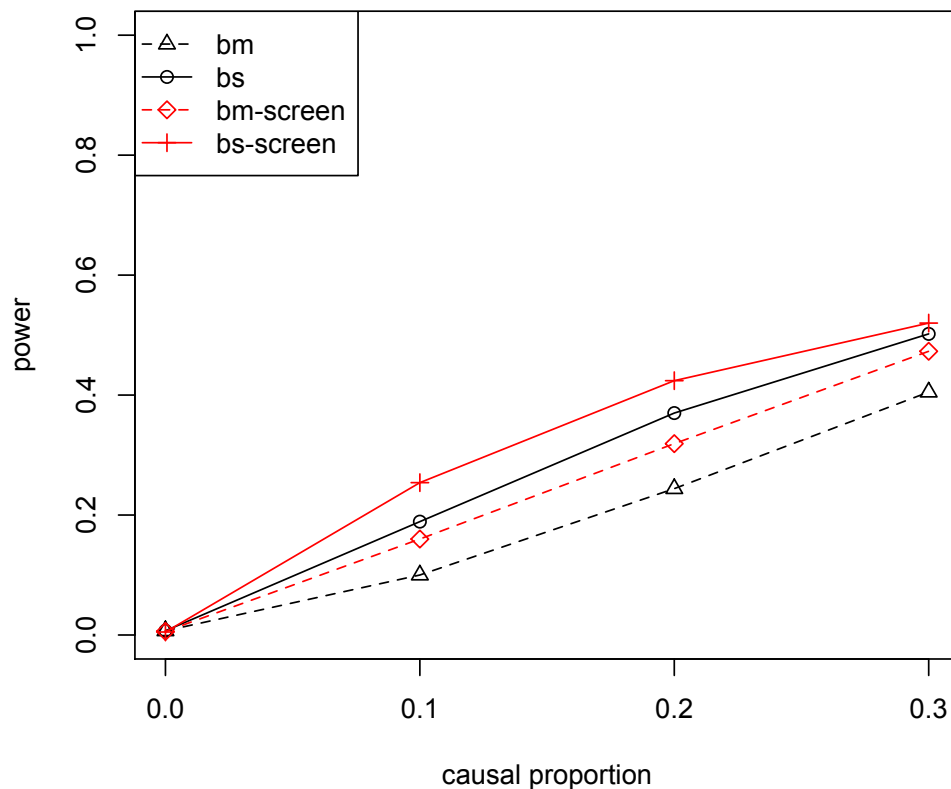


FIGURE 3.3: Power comparison when screening with between-family components

Table 3.4: Simulation results based on 1000 replicates at nominal level 0.05 under population stratification (Continental populations, with $F_{st} = 0.1$, $nfam = 2000$, $protect = 0$, $geneProp = 0.05$)

PCA	causal	burden	skat1	skat	bu	bm	bs	su	sm	ss
F	0.1	0.238	0.728	0.843	0.272	0.304	0.418	0.344	0.413	0.414
	0.2	0.37	0.775	0.896	0.435	0.488	0.423	0.381	0.447	0.341
	0.3	0.526	0.798	0.916	0.564	0.656	0.412	0.381	0.467	0.291
T	0.1	0.436	0.627	0.755	0.277	0.305	0.558	0.34	0.407	0.553
	0.2	0.654	0.707	0.868	0.434	0.488	0.63	0.369	0.447	0.509
	0.3	0.819	0.75	0.905	0.565	0.642	0.664	0.387	0.465	0.446

Table 3.5: Simulation results based on 1000 replicates at nominal level 0.05 under population stratification (European populations)

causal	burden	skat1	skat	bu	bm	bs	su	sm	ss
0.1	0.594	0.582	0.797	0.294	0.31	0.422	0.242	0.324	0.385
0.2	0.813	0.68	0.893	0.462	0.521	0.554	0.305	0.407	0.439
0.3	0.903	0.765	0.933	0.592	0.678	0.613	0.325	0.443	0.432
0.1	0.581	0.583	0.791	0.283	0.301	0.416	0.244	0.316	0.385
0.2	0.799	0.668	0.892	0.454	0.512	0.556	0.311	0.406	0.435
0.3	0.889	0.764	0.937	0.586	0.666	0.609	0.325	0.444	0.427

Table 3.6: Type I error rates when screening with between-family components

screen	$\alpha = 0.05$				$\alpha = 0.01$			
	bs	bm	ss	sm	bs	bm	ss	sm
100	0.044	0.046	0.022	0.034	0.007	0.007	0.003	0.006
50	0.042	0.04	0.025	0.038	0.009	0.005	0.005	0.007
40	0.041	0.045	0.03	0.033	0.01	0.008	0.004	0.005
30	0.037	0.055	0.035	0.029	0.007	0.008	0.006	0.005
20	0.043	0.047	0.033	0.043	0.007	0.011	0.005	0.006
10	0.052	0.048	0.046	0.045	0.008	0.012	0.007	0.007

4

Discussion

In this thesis, we proposed a novel method to robustly test the QTL in nuclear families. For each family, the genotype is decomposed to two parts: the within-family component and the between-family component. The best is constructed based on the within-family component, thus, the test is to robust to population stratification. Instead of discarding the between-family component, we use the between-family component to build a weights for each variants when combining their score statistics. Comparing to the conditional test, we fully exploited the genotype information. When the proportion of causal variants are small, our methods show that this weighting scheme can significantly improve the test power.

In general, the tests based within-family components have a lower power than the population based tests, even when weighted by between-family components. However, when population stratification exists, population based methods cannot maintain the right type I error rate. Our method is always robust to population stratification no matter the population stratification is caused by continental difference populations or closely related populations.

We also show that our method can be combined with the screening procedure developed by Jiang et al (Jiang et al., 2014a). When combining with the screening procedure, the between-family components have been used twice, however, this did not influence the validity of our methods due to the fact that the final test is constructed from within-family components. combining screening procedures can also improve the test power when the screening rate is appropriate for both our methods or methods without weighting scheme.

Currently, we presented the method only in offspring-parent trios, but this methods can be easily extended to other nuclear family structure, for example, parental genotype are unknown but sibling genotypes are given. As Abecasis et al proposed (Abecasis et al., 2000), the between family component will be the average of sibling's genotype, i.e., $b = \frac{\sum_j g_j}{n_s}$ and the within-family component is defined the same as in our method. In such situation, at the first stage, the between-family component will be used only once for each family to compute the weights for each locus and then the within-family components will be taken as indecent for siblings to construct the test statistics at the second stage.

Bibliography

- Abecasis, G., Cardon, L., and Cookson, W. (2000), “A general test of association for quantitative traits in nuclear families,” *The American Journal of Human Genetics*, 66, 279–292.
- Allison, D. B. (1997), “Transmission-disequilibrium tests for quantitative traits.” *American journal of human genetics*, 60, 676.
- Balding, D. J. and Nichols, R. A. (1995), “A method for quantifying differentiation between populations at multi-allelic loci and its implications for investigating identity and paternity,” in *Human Identification: The Use of DNA Markers*, pp. 3–12, Springer.
- Chung, R.-H. and Shih, C.-C. (2013), “SeqSIMLA: a sequence and phenotype simulation tool for complex disease studies,” *BMC bioinformatics*, 14, 199.
- Cirulli, E. T. and Goldstein, D. B. (2010), “Uncovering the roles of rare variants in common disease through whole-genome sequencing,” *Nature Reviews Genetics*, 11, 415–425.
- Davies, R. B. (1980), “Algorithm AS 155: The distribution of a linear combination of χ^2 random variables,” *Applied Statistics*, pp. 323–333.
- De, G., Yip, W.-K., Ionita-Laza, I., and Laird, N. (2013), “Rare variant analysis for family-based design,” *PLoS One*, 8, e48495.
- Devlin, B. and Roeder, K. (1999), “Genomic control for association studies,” *Biometrics*, 55, 997–1004.
- Gibson, G. (2012), “Rare and common variants: twenty arguments,” *Nature Reviews Genetics*, 13, 135–145.
- Gudbjartsson, D. F., Walters, G. B., Thorleifsson, G., Stefansson, H., Halldorsson, B. V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S., et al. (2008), “Many sequence variants affecting diversity of adult human height,” *Nature genetics*, 40, 609–615.

- Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., and Lin, X. (2013), “Family-based association tests for sequence data, and comparisons with population-based association tests,” *European Journal of Human Genetics*, 21, 1158–1162.
- Jiang, Y., Conneely, K. N., and Epstein, M. P. (2014a), “Flexible and Robust Methods for Rare-Variant Testing of Quantitative Traits in Trios and Nuclear Families,” *Genetic epidemiology*, 38, 542–551.
- Jiang, Y., Satten, G. A., Han, Y., Epstein, M. P., Heinzen, E. L., Goldstein, D. B., and Allen, A. S. (2014b), “Utilizing population controls in rare-variant case-parent association tests,” *The American Journal of Human Genetics*, 94, 845–853.
- Lette, G., Jackson, A. U., Gieger, C., Schumacher, F. R., Berndt, S. I., Sanna, S., Eyheramendy, S., Voight, B. F., Butler, J. L., Guiducci, C., et al. (2008), “Identification of ten loci associated with height highlights new biological pathways in human growth,” *Nature genetics*, 40, 584–591.
- Macgregor, S., Cornes, B. K., Martin, N. G., and Visscher, P. M. (2006), “Bias, precision and heritability of self-reported and clinically measured height in Australian twins,” *Human genetics*, 120, 571–580.
- Mathieson, I. and McVean, G. (2012), “Differential confounding of rare and common variants in spatially structured populations,” *Nature genetics*, 44, 243–246.
- Morris, A. P. and Zeggini, E. (2010), “An evaluation of statistical approaches to rare variant analysis in genetic association studies,” *Genetic epidemiology*, 34, 188.
- Price, A. L., Patterson, N. J., Plenge, R. M., Weinblatt, M. E., Shadick, N. A., and Reich, D. (2006), “Principal components analysis corrects for stratification in genome-wide association studies,” *Nature genetics*, 38, 904–909.
- Reich, D. E. and Goldstein, D. B. (2001), “Detecting association in a case-control study while correcting for population stratification,” *Genetic epidemiology*, 20, 4–16.
- Silventoinen, K., Sammalisto, S., Perola, M., Boomsma, D. I., Cornes, B. K., Davis, C., Dunkel, L., De Lange, M., Harris, J. R., Hjelmborg, J. V., et al. (2003), “Heritability of adult body height: a comparative study of twin cohorts in eight countries,” *Twin research*, 6, 399–408.
- Spielman, R. S., McGinnis, R. E., and Ewens, W. J. (1993), “Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (IDDM).” *American journal of human genetics*, 52, 506.
- Sun, F., Flanders, W., Yang, Q., and Zhao, H. (2000), “Transmission/disequilibrium tests for quantitative traits,” *Annals of human genetics*, 64, 555–565.

- Visser, P. M. (2008), "Sizing up human height variation," *Nature genetics*, 40, 489–490.
- Weedon, M. N., Lango, H., Lindgren, C. M., Wallace, C., Evans, D. M., Mangino, M., Freathy, R. M., Perry, J. R., Stevens, S., Hall, A. S., et al. (2008), "Genome-wide association analysis identifies 20 loci that influence adult height," *Nature genetics*, 40, 575–583.
- Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011), "Rare-variant association testing for sequencing data with the sequence kernel association test," *The American Journal of Human Genetics*, 89, 82–93.
- Zheng, G., Freidlin, B., and Gastwirth, J. L. (2006), "Robust genomic control for association studies," *The American Journal of Human Genetics*, 78, 350–356.
- Zhu, X. and Elston, R. C. (2001), "Transmission/disequilibrium tests for quantitative traits," *Genetic epidemiology*, 20, 57–74.