A Permutation based Mixed Effect Model in Rare Variants Association Study

Faculty Advisor: Dr. Jianping Sun

Luke Vilaseca

Senior Honors Project
Disciplinary Honors College

Department of Mathematics and Statistics
University of North Carolina at Greensboro
United States
November 24, 2020

1 Background

Deoxyribonucleic Acid, DNA, is the hereditary foundation of all living organisms with the primary location being in the nucleus of the organism's cells. Each cell in the human body contains the same DNA which is found on chromosomes. Chromosomes are the storage of human's genetic blueprints that make up everything in the body. This is done through cell division. Cell division replaces old cells with new ones with the new ones containing the same DNA as its parent cell. This is ensured by the chromosomes that ensure the DNA is accurately copied and distributed. On rare occasions, however, the DNA can be incorrectly copied.

Every human possesses a total of twenty-four chromosomes with each containing specific "instructions" on developing the human body. From human to human, 99% of the three billion bases that make up the DNA are the same. Therefore, it is apparent to scientists to see what each chromosome is responsible for in a human being. More specifically, they are able to identify the chromosome that certain diseases can stem. For example, chromosome 1 is linked to the following diseases being encoded in human beings: gaucher, prostate cancer, alzheimer, and glaucoma.

The structure of DNA is two strands, one sugar and one phosphate molecule, that spirals and is connected by multiple amino acid bases. There are four amino acid bases, guanine, adenine, cytosine, and thymine, notated as G, A, C, and T. The amino acid bases link to a phosphate group of the strand and form the basic structural unit of DNA known as nucleotide. These amino acid bases connect to one another by forming in two combinations or what is known as base pairs: adenine pairs with thymine and guanine pairs with cytosine. The way that these base pairs organize is known as genetic code. This order of the bases provides specific information for either creation or maintenance of an organism. The way genetic code works is similar to how letters are used to form different words. As previously mentioned, there are more than three billion DNA base pairs in humans. This complete set of DNA is known as a genome. The information in a human's genome is everything needed to create and maintain people.

Although these are the general rules of DNA, alterations of the most common DNA nucleotide sequences are possible. This is known as genetic variation. These alterations can be described as benign, pathogenic, or unknown when comparing the DNA sequence to that of the Genome Reference Consortium's DNA sequence. The way a person's DNA sequence can acquire this variation is due to one of the following five reasons: single nucleotide polymorphism, insertions, deletions, substitutions, or structural variants. The main DNA sequence variation that will be of note is single nucleotide polymorphism, or SNPs. SNPs are positions in a person's genome where if compared to the same location of another person's genome, then the nucleotide differ. For instance, one person may have a guanine nucleotide where the general population has a cytosine nucleotide. To extend this further, say that 95% of the general population has this cytosine nucleotide at this particular point. Then a total of 5% of individuals have this DNA variation of a guanine nucleotide in the same location. With roughly ten million Some of these SNPs attribute to physical differences among the population and others can affect how an individual develops diseases.

The overall impact of genetic variations on the traits of an individual is measurable. This is what

is known as trait heritability, h^2. Trait heritability is a statistical concept that takes a given trait and tells the overall significance the genetic variation contributes to it. The range of trait heritability is from zero to one with zero meaning that most of the variability in a trait is at the cost of environmental factors and little influence from genetics. On the opposite side, being closer to one indicates that genetics play a substantial role in a trait's variability. This range is continuous and, therefore, can result in complex traits having a heritability of 0.5. This indicates that the trait is influenced by both environmental factors and genetics.

The Genome Wide Association Study, or GWAS, used this concept of heritability and began an observational study on genetic variants to uncover if certain variants in an individual are associated with a trait. One of the primary focuses of this study was on the relationship between single-nucleotide polymorphisms and major human diseases. Ultimately, GWAS was able to validate a relationship between human diseases and SNPs by focusing on the association of common variants with that of major diseases. A common genetic variant is has a minor allele frequency greater than 5%. GWAS was only able to conduct single variant tests on common genetic variants due to the astronomical prices of genetic testing back in the day. Had they been able to conduct tests with more that ten or so genomes at the time, they might have been able to focus on the role of rare gene variants in human diseases. However, this was not the case and the major role rare genetic variations played would not be discovered until years after. Therefore, GWAS focused on common genetic variants and concluded that SNPs were linked to that of human diseases. Another limitation to GWAS was with their heritability. The role of the genetic variance played in a varying trait is a central part of GWAS observational study. However, the heritability was often much less than 0.1. This means that the variation of the trait was caused more by environmental factors than genetic variation.

In the past decade, the cost to sequence a genome has dropped from almost one-hundred million dollars to one-thousand dollars. This has allowed for studies of the genome to branch out and focus on more niche genetic topics. One such topic that grew in popularity was that of rare genetic variations. A rare genetic variation is one that has a minor allele frequency below 5%. This popularity was due to two main factors: sample sizes of genetic testing could be larger because studies no longer required large funds to operate them and because rare genetic variations are the major contributors to an individual's genetic vulnerability to major diseases. Thus, this gave way to the Rare Variant Association Studies. However, there are certain challenges and limitations that face the Rare Variant Association Studies. First being that single variant tests of the association have very low power due to the fact that a small number of people carry a given rare variant allele. In addition, a given genetic variation in the same gene can have one of the following outcomes: no effect, different effects, or even opposite effect. With these given biological challenges, there is a need for a robust and accurate statistical model that groups the variants and does a region-based test. Thus, this began the development of methods: the Burden test and the Sequence Kernel Association Test. However, these statistical tests did not test for multiple traits at the same time.

2 Existing Methods

There is a growing literature on analyzing the association of a set of rare variants, and most of them can be partitioned into three categories: the burden type of tests, the variance component type of tests, and the omnibus tests.

A natural approach of set-based rare variants association test is to create a new variable that counts the number of risk alleles an individual carries for the set and test whether this new variable is associated with phenotypic variation. These tests, sometimes called burden tests (Morgenthaler and Thilly, 2007; Li and Leal, 2008), are motivated by population genetics evolutionary theory that most rare missense alleles are deleterious, and the effect is therefore generally considered one-sided (Kryukov et al., 2007). This type of tests have simple model(model (1)) and it is obviously that this collapsed variable can be incorporated into a regression model to account for potential confounders (Morris and Zeggini, 2010).

$$Y = \alpha + \beta \sum_{j=1}^{p} G_j + \epsilon \tag{1}$$

In addition, this simple count of risk alleles can also be extended to a weighted sum, where the weight may be the frequency of rare allele in the unaffected subjects (Madsen and Browning, 2009). However, it is also easy to see that burden tests have big limitations on strong assumption about the same direction/magitude of genetic effects, and tend to lose power when only a small proportion of rare variants in the set is truly causal.

Another approach to testing the association of a set of rare variants is to use the mixed effect model framework. In this framework, the effects of variants are assumed to be independently and identically distributed with a mean 0 and variance τ^2. That is, in model (2), assuming β_i's are independently following $N(\mu, \tau^2)$ distribution. To test whether a set of variants is associated with the phenotype, it is equivalent to test whether the variance component $\tau^2 = 0$. Among this variance component type of tests, the sequence kernel association test (SKAT) (Wu et al., 2011) is the most popular one (model (2)).

$$Y = \alpha + \beta_1 G_1 + \beta_2 G_2 + \ldots + \beta_p G_p + \epsilon, \tag{2}$$

where the effect for the j^{th} variant

$$\beta_j \overset{iid}{\sim} (\mu, \tau^2). \tag{3}$$

SKAT is powerful when there is a large number of non-causal variants, and/or both protective and deleterious variants are present. However, it may suffer from inflated type I error under some scenarios (Konigorski et al., 2017; Zhang et al., 2019).

Burden type of tests are more powerful when a large fraction of variants are causal and effects are in the same direction; while the variance component type of tests are more powerful when a small fraction of variants are causal, or the effects have mixed directions. However, both scenarios can happen in real world when scanning the genome. Hence, researchers began to develop omnibus

tests, i.e. combining burden tests and variance component tests, to achieve more robust power under various scenarios. The ways of combining two types of tests include combining two p-values (Derkach A et al., 2013), or combining two test statistics (Lee et al., 2012) obtained from burden and variance component tests.

By assuming the variant effects following an independently and identically distribution with mean μ and variance τ^2, where μ depends on the characteristics of genetic variants, a Mixed effects Score Test (MiST) was proposed by Sun et al., 2013, and was shown to be a combination of burden test and SKAT(model (3)).

$$Y_i = \alpha_0 + X\alpha_1 + G\beta + \epsilon, \tag{4}$$

with

$$\beta_j = z_j\pi + \delta_j, \tag{5}$$

where

$$\delta_j \overset{iid}{\sim} (0, \tau^2). \tag{6}$$

Here z_j is a vector contains q characteristics of the j^{th} variant. These characteristics can be any genetic score of a particular variant, or an indicator to denote the property of a variant, such as missense or nonsense mutation. By adding a second level of model to the variant effect β_j, unlike SKAT method, MiST assumes the variant effects are random but with a mean depending on the variant characteristics. Hence, to test whether the variant effect β is zero, it is equivalent to test the mean, $z\pi$, and variance component, τ^2, are zeros simultaneous. That is, the null hypothesis in MiST method is

$$H_0 : \pi = 0 \; and \; \tau^2 = 0, \tag{7}$$

which can be further divided into two sub null hypotheses:

$$H_0^1 : \tau^2 = 0 \; and \; H_0^2 : \pi = 0 \mid \tau^2 = 0 \tag{8}$$

This two step hypothesis has the advantage of not requiring estimations of π and τ^2 under H_1, which is typically a difficult issue. Similar to other omnibus tests, MiST is valid, robust and powerful under different scenarios of genome scan. However, because it includes the variance component type of tests, it is expected also having similar limitations as SKAT, such as the inflated type I error. Some preliminary simulation studies have shown that when the continuous phenotypic trait departures from normal distribution, then MiST tends to have inflated type I error, especially when test significance is stringent. Hence, in this proposal, we propose to develop a modified MiST method based on computation permutations to control the potential type I error inflation.

3 Proposed Method

One pros of the MiST is that it can calculate p-value analytically so that to reduce the computation cost, which, otherwise, could be huge for an implement genome-widely. However, this analytical

p-value is obtained based on several approximations related with normal distribution. Hence, when the traits are not closed to normal distributions, the approximation could get worse and result the inflated type I error.

Therefore, there is a need to improve MiST so that it can control type I error well. One possible way to solve the issue with type I error is by using a permutation based computing method. Suppose there is $n = 1000$ individuals and for each of these individuals it is possible to observe the trait, Y_i, covariate X_i, and the genotype G_i, for $i = 1, 2, \ldots, 1000$. This is demonstrated in Table 1.

Table 1: Original Data Set

Y_1	X_1	G_1
Y_2	X_2	G_2
Y_3	X_3	G_3
$\vdots$	$\vdots$	$\vdots$
Y_{1000}	X_{1000}	G_{1000}

If there is no relationship between Y and G, then there should not exist the relationship between Y and G either after their nature order was broken. For example, in Table 2, where we break the original order between (Y_i, X_i) and G_i, then we should expect that there is no relationship between Y and G under this new order, if we assuming no such association exist under the original order. The reason that Y and X are kept together is due to the fact that there may exist an association. Therefore, this association requires Y and X to be together.

Table 2: Permutated Data Set

Y_{103}	X_{103}	G_1
Y_{72}	X_{72}	G_2
Y_{546}	X_{546}	G_3
$\vdots$	$\vdots$	$\vdots$
Y_{389}	X_{389}	G_{1000}

If the original data sets contains n individuals, then it is possible to obtain $n!$ permutated data sets. For each of these permutated data sets, MiST calculates two tests statistic, S_τ^j and S_π^j, to test the two hypotheses proposed in model (4). Suppose the corresponding test statistics calculated under the original data set are denoted as, S_τ^0 and S_π^0. Using the definition of p-values, the permutated p-values, p_τ and p_π, for two hypotheses are calculated as the proportions of $S_\tau^j \geq S_\tau^0$ and $S_\pi^j \geq S_\pi^0$, where $j = 1, 2, ..., n!$.

$$p - value = \frac{\#\ of\ S_j \geq S_0}{n!}, j = 1, 2, \ldots, n!$$

In addition, in order to test two hypotheses simultaneously, we need to combine two permutated p-values, p_τ and p_π, in to an overall p-value. Here, we proposed to use Fisher's combination method. Let

$$q = -2 \times \{\ln p_\tau + \ln p_\pi\}, \tag{9}$$

then $q \sim \chi^2(4)$ approximately. Therefore, $p_{overall}$ can be calculated as

$$p_{overall} = P(\chi^2(4) > q)\,. \tag{10}$$

In reality, since the sample size n is usually large, it is impossible to do the full permutation, i.e. $n!$ permutated data sets. Hence, we can choose to generate P permutated data sets for each of N available data sets. Consequently, for each of the N data sets, a permutated overall p-value, p_i or $p_{overall}$, for $i = 1, 2, \ldots, N$, is obtained. The empirical type I error is then calculated the following way

$$P(type\ I\ error) = \frac{\#\ of\ p_i < \alpha}{N}, i = 1, 2, \ldots, N \tag{11}$$

That is, the final empirical type I error rate is the number of p-values less than the test significance level, α, divided by N, the number of data sets.

4 Simulation Studies

To investigate the performance of proposed permutation test method, we did multiple simulation studies under various scenarios, by running MiST package through R studio. Particularly, our simulation studies focus on the type I error rate by comparing empirical false discover rate under different distributions for model error.

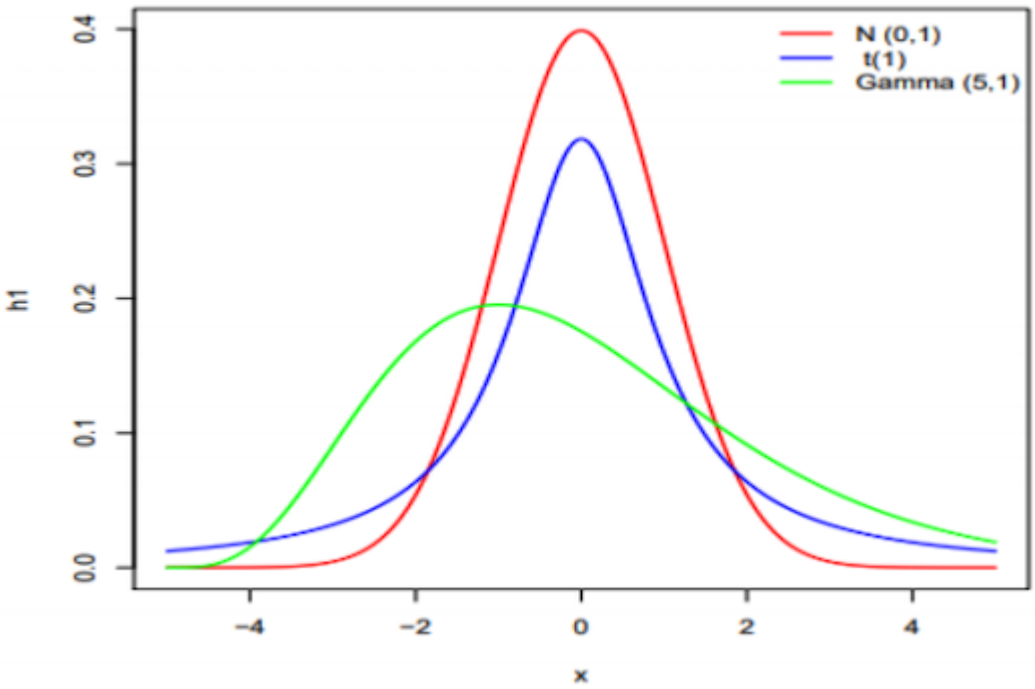
In the simulations, we generated the data sets by using $t(3)$, and $Gamma(5, 1) - 5$, as model errors, respectively, and compared the emprical type I error rate when significance level, alpha equals 0.05 and 0.01. The reason we choose these two distributions for the model error is because we want to check the performance of the proposed permutation method when the error doesn't follow a normal distribution, which is an ideal scenario. Particularly, t-distribution is as symmetric as normal but having longer tails than normal, and Gamma distribution is skewed compared with normal (Figure 1).

Since the focus of the simulations is on the inflated type I error, the data sets generated must be done so under the null model. This means that Y is a response generated from the model (12)

$$Y_{n\times 1} = X_{n\times m}\alpha_{m\times 1} + \epsilon_{n\times 1} \tag{12}$$

In the simulation study the sample size n equals 1000, and we consider two covariates in the null model. Hence, in the null model (12), the covariate matrix X contains 3 columns, where the first column represents for intercept and the remaining two columns represent for two covariate variables. The second column of X is generated from the $Uniform(10, 50)$ and represents a continuous

Figure 1: Density Curve Plot



covariate. The third column of X is generated from $Bernoulli(0.5)$ distribution and accounts for a binary covariate. In addition, the coefficient alpha in the null model is equal to a vector of $(0.5, 1, 1)$ for the intercept and the effects of two covariate variables. For the model error, ϵ, it is generated from one of two distributions. The first one is the t-distribution, $t(3)$, which is symmetric with a heavier tail compared to the normal distribution. The second distribution is $Gamma(5, 1) - 5$. Gamma is skewed and by minus 5, it has a mean of 0, similar to the t and the ideal scenario, standard normal distribution. Taking all this information, Y is generated by plugging in X, α, and ϵ into the null model (12).

When using the MiST package, it is paramount to provide the genotype matrix, $G_{n\times p}$ as well as the matrix of the characteristic, $Z_{p\times v}$, where p is the number of genetic variants, and v is the number of variant characteristics. Therefore, G and Z must be generated before simulations. In the simulation study, we designed to use a real world data set from the Dallas Heart Study (Victor et al., 2004) as the genotype matrix G, in order to mimic the realistic complexity among genetic variants. Specifically in the simulation study, the genotype data of 50 rare variants from gene ANGPTL5 from 1000 individuals was used as the genotype matrix G. For simplicity, the variant characteristic matrix Z was set to be a vector of length 50 with all the elements as 1's. Finally, we generated $N = 1000$ data sets, and for each data set, run $P = 2000$ permutations. The simulation results are summarized in Table 3.

Table 3: Empirical Type I Error Rates

	Permutation method		MiST	
	$t(3)$	$Gamma(5,1)-5$	$t(3)$	$Gamma(5,1)-5$
$\alpha = 0.01$	0.057	0.061	0.058	0.057
$\alpha = 0.01$	0.015	0.013	0.017	0.013

In addition, in Table 3, we also list the empirical type I error rate by using the original MiST package directly on the same $N = 1000$ data sets without permutations. From the comparison between the proposed permutation method and the original MiST method, we can see that the permutation method improves the type I error rate when the model error follows a t-distribution, epically when the significant level is small. When $\alpha = 0.01$, the type I error rate of the permutation method for $t(3)$ is 0.015, which is smaller than 0.017, the one obtained by using MiST directly. There is also slightly decrease of type I error rate for $t(3)$, when $\alpha = 0.05$ by using the permutated method. However such improvement is so obvious when model error follows a gamma distribution, and further research is needed for the potential reason.

5 Discussion

To sum up, a modified MiST, where the p-value is obtained empirically by permutating test statistics to avoid approximation related with normal distribution, is used in this project. To eliminate potential large computation burden from permutations, we can propose to use a two-step permutation. That is, MiST permutates limited times in the first step for the purpose of scanning, and then only does extensive permutations for the candidate regions obtain from the first step. For the simulations, large simulation studies are used to check the performance of the proposed modified MiST method, especially in the view of controlling type I error under various scenarios, and the simulation studies show that the proposed permutation method helps to improve the control of type I error under some circumstance, especially when the significance level alpha is small.

Since in the genetic association study, people usually use more stringent significance level than 0.05 or 0.01 due the large number of genetic variants, it is worth to conduct further investigations on the performance of permutated method when alpha is as small as 10^{-3} or 10^{-4}.

In addition, the current permutation method is time consuming because of the needs for large numbers of permutations. For example, in order to run the simulation for one distribution, it would take about two days. Hence, possible future works could modify the current permutation testing in order to decrease the computation burden.

6 Bibliography

1. Derkach A, Lawless JF, Sun L. Robust and powerful tests for rare variants using Fisher's method to combine evidence of association from two or more complementary tests. *Genetic epidemiology* 2013; 37 (1), 110-121.

2. Konigorski S,Yilmaz YE, and Pischon T. Comparison of single-marker and multimarker tests in rare variant association studies of quantitative traits. *PLoS ONE*. 2017; 12(5): e0178504.

3. Kryukov GV, Pennacchio LA, Sunyaev SR. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet.* 2007; 80:727–739.

4. Lee S, Wu MC, and Lin X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*. 2012; 13, 762-775.

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

6. Madsen BE, Browning SR. Methods for detecting association with rare variants for common diseases: Application to analysis of sequence data. *Am J Hum Genet.* 2009; 83:311–321.

7. Morgenthaler S, Thilly WG. A strategy to discover genes that carry multi-allelic or monoallelic risk for common diseases: A cohort allelic sums test (CAST). *Mutat Res.* 2007; 615:28–56.

8. Morris AP, Zeggini E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidem.* 2010; 34:188–193.

9. Sun J, Zheng Y, Hsu L. A unified mixed-effects model for rare-variant association in sequencing studies. *Genet Epidem.* 2013; 37(4): 334–344.

10. Victor R.G., Haley R.W., Willett D.L., et al. The Dallas Heart Study: a population-based probability sample for the multidisciplinary study of ethnic differences in cardiovascular health. The American Journal of Cardiology. Vol. 93, Issue 12, (2004), 1473-1480.

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

12. Zhang X, Basile AO, Pendergrass SA, and Ritchie MD. Real world scenarios in rare variant association analysis: the impact of imbalance and sample size on the power in silico. *BMC Bioinformatics* 2019; 20: Article number: 46.