The efficiency of single SNP and SNP-set analysis in genome-wide association studies

Songklanakarin J. Sci. Technol.
43 (1), 243-251, Jan. - Feb. 2021

http://www.sjst.psu.ac.th

Original Article

The efficiency of single SNP and SNP-set analysis in genome-wide association studies

Sirikanlaya Sookkhee[1], Pianpool Kirdwichai[1*], and M. Fazil Baksh[2]

[1] *Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bang Sue, Bangkok, 10800 Thailand*

[2] *Department of Mathematics and Statistics, School of Mathematical, Physical and Computational Sciences, University of Reading, Reading, Berkshire, RG6 6FN United Kingdom*

Received: 9 May 2019; Revised: 2 December 2019; Accepted: 15 December 2019

Abstract

The objective of this research is to compare and identify effective methods for the identification of gene loci associated with a disease outcome in the analysis of genome-wide data. We evaluate three methods, namely single SNP analysis, Sequence Kernel Association Test (SKAT) and the recently proposed Generalized Higher Criticism (GHC). The simulated data used in this research were constructed from a control data set in a study of Crohn's disease. True positive (TP) and false positive rate (FP) were evaluated under different genetic models for disease with significant thresholds adjusted for multiple hypothesis testing based on the permutation method. The findings are mixed with all three methods giving similar TP rates under some disease models and different rates for other models. Overall, GHC is shown to be preferable in terms of error rates but it is disadvantageous in terms of computational efficiency.

Keywords: single SNP analysis, Sequence Kernel Association Test, Generalized Higher Criticism, permutation test

1. Introduction

Advances in medical science and the promise of innovative treatments such as gene therapy depend on accurate and effective statistical analysis methods for identifying target genetic region(s) for treatment. Single SNP analysis is a traditional method used to analyze the association between SNPs and disease by analyzing each SNP loci while logistic regression and chi-square test (Clarke *et al*., 2011; Lewis, 2002) are simple and effective ways to conduct the single SNP tests for association. Although the single SNP analysis has proven useful in many studies, this method faces the problem of type I error rate inflation and methods for correcting this, such as Bonferroni, are often overly conservative. An alternative to single SNP analysis, which groups variants into SNP-sets and jointly tests for an effect with disease outcome, is motivated by the possibility that there could be many variants, each with small individual effects and in LD with causal variant(s) (Wu *et al*., 2010). This approach is claimed to be more effective (Kirdwichai & Baksh, 2019; Schaid, Rowland, Tines, Jacobson, & Poland, 2001; Zhao *et al*., 2012) in identifying disease loci.

Wu *et al*. (2010) proposed a logistic kernel machine model to analyze the association between SNP-set (grouping by gene, gen networks, pathways or haplotype block) with disease outcome and claimed that grouping into SNP-sets can lead to improving the power of the test. Wu *et al*. (2011) utilize this concept and proposed SKAT to analyze rare variant association with complex traits. SKAT is a supervised, flexible, computationally efficient regression method to test for association between common or rare variants and disease (Iuliana-Laza, Lee, Makarov, Buxbaum, & Lin, 2013; Lee *et al*., 2012). A crucial feature in implementing SKAT is assigning appropriate weights as this can have an impact on the power of the test procedure (Lee, Miripolsky, & Wu, 2017).

*Corresponding author
Email address: pianpool.k@sci.kmutnb.ac.th

More recently the Generalized Higher Criticism (GHC) was proposed for testing multiple SNPs simultaneously in genome-wide association studies (Barnett, Mukherjee, & Lin, 2017). This approach extends a technique called higher criticism (HC) and compares the observed significant findings in the single SNP analysis with the expected number under the null while accounting for the correlation between the hypothesis test statistics. It should be noted that HC has been applied in high-dimensional, low correlation signal detection settings whereas SNP-sets frequently have a low number of true positive associations which can be highly correlated. The GHC method is flexible to the correlation structure and is computationally efficient, producing a p-value without the need for simulation of the null distribution. SKAT and GHC are implemented in the R packages SKAT (Lee, 2017) and GHC (Barnett, 2015)

Crohn's disease is an autoimmune disease that is suspected to be genetically linked with several genes being identified across the genome as being associated with chromosome 16. It's not clear which genes are responsible for this disease. Therefore, the simulation study in this paper uses genotype data on Chromosome 16 from 1,504 individuals in the 1958 British Birth Cohort (Wellcome Trust Case and Control Consortium, 2007). False positive (FP) and true positive (TP) rates were evaluated for the different methods with significant thresholds adjusted for multiple hypothesis testing using a permutation method. Computational efficiency is also evaluated. Additionally, the methods are applied to real data from a study (Wellcome Trust Case and Control Consortium, 2007) on Crohn's disease and the findings are compared.

2. Methods

2.1 Single SNP analysis

Logistic regression (Bush & Moore, 2012) is an extension of linear regression where the outcome of a linear model is transformed using a logistic function that predicts the probability of having case status given a genotype class. Suppose that the possible genotypes at a particular locus are CC, CT, and TT and suppose that C is the rarer of the two alleles C and T. The additive genetic model then corresponds to TT = 0, CT = 1, and CC = 2, respectively.

Let P_{ij} be the probability of disease for this individual. The logistic regression model is

$$\text{logit}(P_{ij}) = \alpha_0 + \boldsymbol{\alpha}'\mathbf{C}_i + \boldsymbol{\beta}_j T_{ij} \quad (1)$$

where a_0 is an intercept term, α' is a vector of regression coefficients for the q covariates, $\mathbf{C}_i = (C_{i1}, C_{i2}, \ldots, C_{iq})$ is the covariate matrix, $\boldsymbol{\beta}_j$ is a vector of regression coefficients for p observed variants, and T_{ij} in a genotype for the j^{th} SNP $(j = 1,2,\ldots,m)$ on the i^{th} individual. $(i = 1,2,\ldots,n)$ Here, $\boldsymbol{\beta}_j = 0$ corresponds to the null hypothesis of lack of association between the j^{th} SNP and disease. In this research, the additive model is used in the simulations and assumed in the analysis of the WTCCC data, but the single SNP analysis is done using a logistic regression model. Logistic regression is a more efficient, but asymptotically equivalent, alternative to the Pearson's c^2 test for analyzing this type of study data.

2.2 Sequence Kernel Association Test (SKAT)

A key feature of SKAT is that it tests for the joint effects of multiple variants in a region of the genome on disease. Regions can be defined by genes, haplotype block, principal component analysis, or sliding window etc. For each region, SKAT analytically calculates a p-value for association (Lettre, Lange, & Hirschhorn, 2007; Lewis, 2002; Wu *et al*., 2010). The theory underlying the test procedure can be viewed within the kernel machine regression framework. Consider the semiparametric logistic regression model for the i-th subject in a study (i=1,..., n)

$$logitP(y_i = 1) = \alpha_0 + \alpha' \boldsymbol{C}_i + h(\boldsymbol{T}_i), \quad (2)$$

where y_i is a binary disease outcome taking values 0 (disease-free) or 1 (disease), a_0 is an intercept term, α' is the vector of regression coefficients for the covariates $\boldsymbol{C}_i$ and $\boldsymbol{T}_i$ are the observed variants and is related to disease through a nonparametric function $h(\cdot)$, which is assumed to lie in a

functional space generated by a positive semidefinite kernel function $K(\cdot,\cdot)$. Under this model, the null hypothesis of no association between disease and gene region $H_0 : h(\boldsymbol{T}) = 0$ can be tested by assuming the $n \times 1$ vector $\boldsymbol{h} = [h(\boldsymbol{T}_1),\ldots, h(\boldsymbol{T}_n)]'$ for the genetic effects of the n subjects follow a distribution with mean 0 and covariance $t\boldsymbol{K}$, where t is a variance component that indexes the effect of the variants.

The semiparametric logistic regression model in equation (1) is equivalent (Wu *et al*., 2011) to

$$logitP(y_i = 1) = \alpha_0 + \alpha'\boldsymbol{C}_i + \beta'\boldsymbol{T}_i, \quad (3)$$

where $\boldsymbol{\beta}' = (\beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients for the p observed variants in the gene region with each β_j following an arbitrary distribution with mean of zero and a variance of $w_j t$, for variance component t and where w_j is a pre-specified weight for variant j. The null hypothesis $H_0 : \boldsymbol{\beta} = \boldsymbol{0}$ is equivalent to the hypothesis $H_0 : t = 0$, which may be tested with a variance-component score test statistic

$$S = (\boldsymbol{Y} - \hat{\mu})'\boldsymbol{K}(\boldsymbol{Y} - \hat{\mu}) \quad (4)$$

where $\boldsymbol{K} = \boldsymbol{TWT}'$, $\hat{\boldsymbol{\mu}}$ is the predicted mean of $\boldsymbol{Y}$= $(y_1,\ldots y_n)'$ under H_0 that is $\hat{\mu} = logit^{-1}(\hat{\alpha}_0 + C\hat{\alpha}), \hat{\boldsymbol{a}}_0$ and $\hat{\boldsymbol{\alpha}}$ are estimated under the null model by regressing $\boldsymbol{y}$ on the covariate $\boldsymbol{C}$ and $\boldsymbol{T} = [T_1, \ldots, T_n]$ is an $n \times p$ matrix with elements variant j of individual i, and $\boldsymbol{W} = diag\left(w_1, \ldots, w_p\right)$ contains the weights of the p variants. SKAT uses the variance-component score test statistic to test the null of no genetic effect but exploits the semiparametric regression approach in computing $\boldsymbol{K}$. The form for $\boldsymbol{K}$ used by SKAT is an $n \times n$ symmetric matrix with elements $\boldsymbol{K}(\boldsymbol{T}_i, \boldsymbol{T}_{i'})$ that measures genetic similarity between the i-th and $i' - th$ subjects in the study. The weighted linear kernel $\boldsymbol{K}(\boldsymbol{T}_i, \boldsymbol{T}_{i'}) = \sum_{j=1}^{p} w_{ij} T_i T_{i'}$ is used in this study.

Choosing an appropriate weight is very important in SKAT because a good choice of weights can improve the power of the test. Weight functions can be specified in the SKAT package in R. In this paper we select four weights based on the Beta density function $Beta(x_j : a,b)$

$$\sqrt{w_j} = \frac{x_j^{a-1}(1- x_j)^{b-1}}{B(a,b)}, 0 < x_j < 1;\ a,b > 0, \quad (5)$$

where B denotes the beta function, a and b are prespecified scale and shape parameters and x_j is the estimated minor allele frequency (MAF) for SNP j using all cases and controls. First selected is the default weight in the SKAT package $Beta(x_j : 1,25)$ which, by choosing a small and b large, substantially up-regulates rare variants and down-regulates common variants (Wu *et al*., 2011). Second is the Madsen and Browning weight, defined as $Beta(x_j : 0.5,0.5)$, which corresponds to $\sqrt{w_j} = 1/\sqrt{MAF_j(1- MAF_j)}$; that is w_j the inverse of the variance of the genotype marker j. The Madsen and Browning weight can pick up signals from both common and rare variants but is thought to suffer from low power. The third option $w_j = 1/\sqrt{q_j}$, which we call inverse mean, is equivalent to $Beta(x_j : 0.5,1)$. The final option considered in this paper is $Beta(x_j : 10,10)$, which gives the appearance of a symmetrical distribution similar to the normal distribution. These weight functions are illustrated in Figure 1. (Sookkhee, Baksh, & Kirdwichai, 2018).

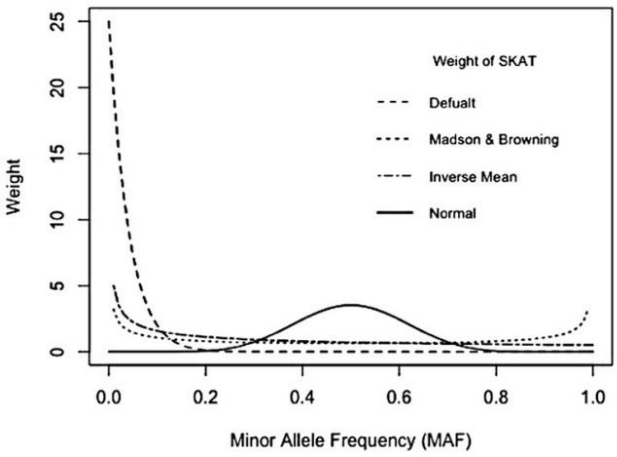


Figure 1. Weight functions use in SKAT analysis

2.3 Generalized Higher Criticism (GHC)

While SKAT uses the observed variants in a gene region to construct a joint score test statistic for association with disease outcome, GHC (Barnett, 2018) tests gene regions for association by using single variant statistics and their correlation matrix to construct a new test statistic and its distribution**.** Consider the parametrization of $P(y_i = 1)$ for the j-th variant in a set of *p* variants,

$$logitP(y_i = 1) = \alpha_0 + \alpha' \boldsymbol{C}_i + \beta_j T_{i,j}, \quad (6)$$

where β_j is the effect of the *j-th* variant and $T_{i,j}$ is the observed *j-th* variant on the *i-th* subject, and the other terms are as in the previous section. The GHC approach exploits the fact that while $\boldsymbol{p}$ might be large in the test of the global null $H_0: \beta = \mathbf{0}$, in a genetic construct variants are likely to be correlated and generally only a small subset of variants are signals for association. In other words, a sparse set of the $\beta' = (\beta_1, \beta_2, \ldots, \beta_p)$ are not zero. GHC aims to account for both sparse signals and correlation among SNPs when combining individual marker test statistics.

Let $\boldsymbol{T_j'} = (T_{1,j} \quad \cdots \quad T_{n,j})$ be the vector of observed variants at the *j-th* marker, $\boldsymbol{Y}$= $(y_1,\ldots y_n)'x$ be the observed disease status and $\hat{\mu} = (\widehat{\mu_1}, \quad \ldots \quad, \widehat{\mu_n})'$ be the predicted mean for $\boldsymbol{Y}$ under the assumption of no genetic effect. The score test statistic for β_j under the global null is

$$Z_j = \frac{\boldsymbol{T_j'}(\boldsymbol{Y} - \hat{\mu})}{\sqrt{\boldsymbol{T_j' P T_j}}}, \quad (7)$$

where $P = W - WC(C'WC)^{-1}C'w$ and $W = diag\{\widehat{\mu_1}(1 - \widehat{\mu_1}) \ldots, \widehat{\mu_n}(1 - \widehat{\mu_R})\}$. These individual variant test statistics are asymptotically jointly distributed as $\boldsymbol{Z} \sim MVN(0, \sum)$, where the $(i,k)^{th}$ component of $\sum$ is estimated by

$$\hat{\sigma}_{jk} = \frac{\boldsymbol{T_j' P T_k}}{\sqrt{\boldsymbol{T_j' P T_j}}\sqrt{\boldsymbol{T_k' P T_k}}}. \quad (8)$$

Define $S(t)$ by

$$S(t) = \sum_{j=1}^{p} \mathbf{1}_{\{|z_j| \geq t\}}. \quad (9)$$

The generalized higher criticism test statistic is defined as

$$T_{GHC} = \sup_{t \geq t_0} \{\frac{S(t) - 2p(1 - \phi(t))}{\sqrt{\widehat{var}(S(t)}}\}, \quad (10)$$

where$\phi(t)$ is the standard normal distribution function and $\widehat{var}(S(t))$ is calculated accounting for the correlation between the Z'_jS. The p-value

$$P(GHC) \geq T_{GHC} \quad (11)$$

is also calculated accounting for this correlation. The algorithm behind this approach is implemented in the R package GHC.

2.3 Multiple hypothesis testing

For *m* independent tests and the rejection level α for each test, the probability of falsely rejecting at least one true null hypothesis, otherwise known at the family-wise error rate (FWER), increases in such a way that for even a moderate number of tests we will almost surely incorrectly reject at least one true null hypothesis. In fact, the probability at least one type I error in *p* tests is $1 - (1 - \alpha)^m$. As shown in Figure 2, the probability at least one type I error gets close to 1 even for a small number of tests.

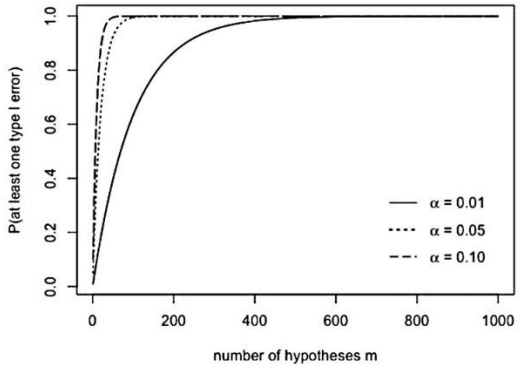


Figure 2. Probability of at least one false positive finding for different number of hypotheses m and significance level 0.01. 0.05 and 0.10.

The simplest method for controlling the FWER is Bonferroni correction. An alternative way to find the appropriate threshold is the permutation method. This method is an empirical method that calculates the p-value by using the observed test statistic value in the permuted distributed of test statistics under the null model. In this research, 10,000 replicates were used for computing the multivariate sampling distribution under the null hypothesis with no gene effect and to establish significance thresholds giving a type I error close to 0.05.

Simulation results of single SNP analysis, SKAT-normal weight and GHC test under the null model of no gene effect in Table 1 confirms that that Bonferroni adjustment leads to a type I error that is much lower than the desired level and therefore Bonferroni correction is conservative and constringent. Considering the Type I error rates based on the new thresholds, shown in Table 1, it's clear that in all cases the nominal type I error of 0.05 is achieved using the Permutation threshold. Therefore, the permutation thresholds are selected for comparing the efficiency of the single SNP analysis, SKAT, and GHC.

2.4 The data and disease model simulation

The genotype data used in this simulation are 13,479 SNPs on Chromosome 16 from 1,504 unaffected individuals in the WTCCC study of Crohn's disease. Using 3,008 haplotypes constructed from the 1,504 genomes, new genotype data were generated and assigned disease status based on 2 disease SNPs both of which have very high MAF's and are highly correlated with other SNPs on their respective genes. The first SNP rs3789038 is located at position 50711672bp in gene HMOX2 and has MAF equal to 0.31. The second, SNP rs3785142 has MAF equal 0.48 and is located at position 50753236bp in gene CYLD (Sookkhee, Baksh, & Kirdwichai, 2017). The model for one disease SNP used to generate disease status is

$$P(\text{diseased} \mid T) = \frac{e^{a_0 + b_1 T}}{1 + e^{a_0 + b_1 T}}, \tag{12}$$

where T is the number of copies of the rare allele of the disease SNP, a_0 is a pre-specified baseline relative risk of disease and b_1 is the gene effect, which in this study was set equal to 0.2. The disease model for two disease SNP is

$$P(\text{diseased} \mid T_1, T_2) = \frac{e^{a_0 + b_1 T_1 + b_2 T_2}}{1 + e^{a_0 + b_1 T_1 + b_2 T_2}} . \tag{13}$$

This model assumes the two disease SNPs act linearly on the logit scale and two situations are investigated. The first is for gene effect $b_1 = 0.1$ and $b_2 = 0.2$ while in the second case the gene effects are fixed at $b_1 = 0.2$ and $b_2 = 0.1$.

The procedure in simulating the data is shown in Figure 3. The 1,504 genotypes are first separated to construct 3,008 haplotypes as the initial data. The haplotypes are next coded to 0 (major allele) and 1 (minor allele) and used to construct pairs of parental genotypes using randomly selected haplotypes. Each pair of parental genotypes are then used to construct the genotype of an individual in the study. Disease status was then determined via the disease probability given by equation 12 in the case of one disease SNP and by equation 13 in the case of two disease SNPs.

Table 1. Achieved type I error of single SNP, SKAT analysis and GHC test under the null model in tests with $a = 0.05$

Method	Bonferroni threshold	Permutation threshold
Single SNP analysis	0.033	0.059
SKAT Default	0.040	0.057
SKAT Madson & Browning	0.031	0.055
SKAT Inverse mean	0.033	0.059
SKAT Normal	0.040	0.055
GHC method	0.039	0.057

3. Simulation Study

In this research, simulation results are presented from single SNP analysis, SKAT, and GHC test for two cases. The simulation result in the first case is one disease SNP and the second is two disease SNPs. A total of 1,500 replicate studies, each consisting 3,000 cases and 3,000 controls are simulated and, for each study, a count is made of the number of SNPs incorrectly identified as significantly associated with disease (false positive rate) and whether the disease SNP is correctly identified (true positive rate) by the single SNP, SKAT and GHC test described above.

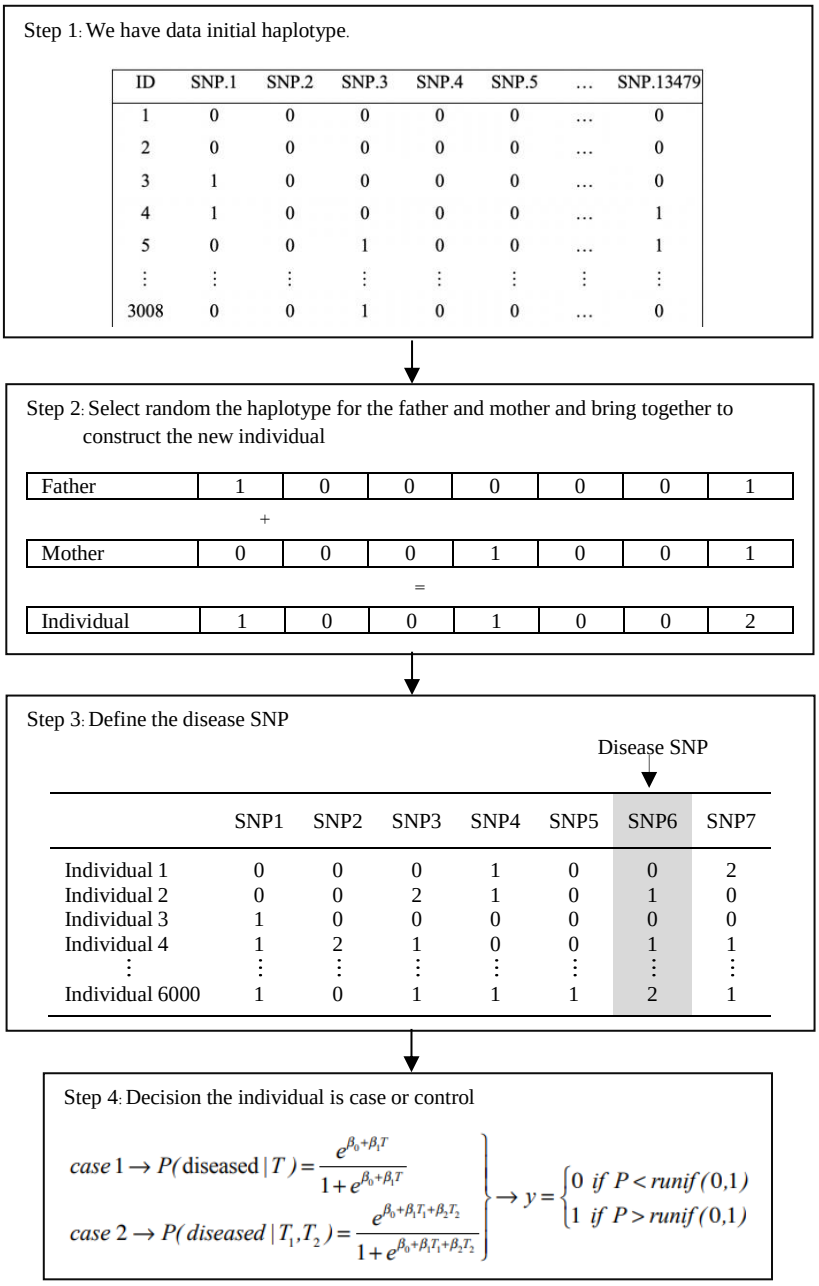


Figure 3. The procedure in the generating simulated data

3.1 The simulation results of one disease SNP

The TP and FP rates of single SNP, SKAT normal, and GHC test of rs3789038 as disease SNP with gene effect size b_1 = 0.2 are provided in Table 2. The FP rate of the single SNP analysis is seen at 0.00105 to be roughly at least 14 and 11 times lower than the SKAT normal and GHC test, respectively. But this comes at the lower TP rate of 0.71. When considering the FP rate, the result shows that the GHC test gives the lowest FP.

The TP and FP rates of the single SNP and SKAT analyses with rs3785142 as disease SNP with gene effect size b_1 = 0.2. The FP rate of the single SNP analysis is 0.00132 and TP rate is 0.89 showing high the TP and FP. When we considered SNP-set analysis, the efficient model of SKAT is SKAT normal showing the TP as 0.92 more than GHC at least

Table 2. The TP and FP rates of single SNP, SKAT, and GHC test with rs3789038 and rs3785142 as disease SNP and effect size $b_1 = 0.2$

Method	Disease SNP rs3789038		Disease SNP rs3785142	
	FP rate	TP rate	FP rate	TP rate
Single SNP analysis	0.00105	0.71	0.00132	0.89
SKAT default	0.00154	0.80	0.00209	0.00
SKAT Madsen & Browning	0.01073	0.86	0.03705	0.65
SKAT inverse mean	0.00981	0.86	0.02734	0.57
SKAT normal	0.00939	0.85	0.04867	0.92
GHC	0.00096	0.82	0.00075	0.86

6 times while the FP rate of GHC is lower than all method. An appropriate weight in this case is the normal and this is confirmed by the simulation results.

3.2 The simulation results of two disease SNPs

The comparisons of TP and FP rates for single SNP, SKAT and GHC test based on the permutation threshold were provided in Table 2. It was computed from 1,500 replicates with a gene effect size for disease SNP rs3789038 of $b_1 = 0.2$ and effect size for rs3785142 of $b_2 = 0.1$. Here the single SNP method is found to have the lowest TP rate (0.85) and FP rate (0.00220). The TP is consistent with the findings in Table 2. We found that when we defined two disease SNP the FP from SKAT and GHC test gave a high TP and FP. SKAT gave higher FP than GHC test while the rates were close to TP.

When considering the gene effect size for disease SNP rs3789038 of $b_2 = 0.1$ and effect size for rs3785142 of $b_1 = 0.2$, the FP rate of single SNP analysis is the lowest while the TP is quite high. SKAT and GHC test give a high TP and high FP. Totally, the efficient model of SKAT is SKAT normal because this weight can pick up signals from both common and rare variants. The efficient SKAT normal was chosen as a representative of the SKAT method in applying to real data.

4. Real Data Application

The three methods were evaluated in the previous section can be applied to real data to define the locate of SNP or region that cause the disease. The data set used has 13,479 SNPs on Chromosome 16 comprise 2,005 cases and 1,500 controls of Crohn's disease studies. The result showed the SNPs which were declared as significant by Single SNP analysis, SKAT normal, and GHC in Table 3. The example of horizontal line of permutation threshold for declaring significant of Single SNP analysis and SNP-set method are presented in Figure 4.

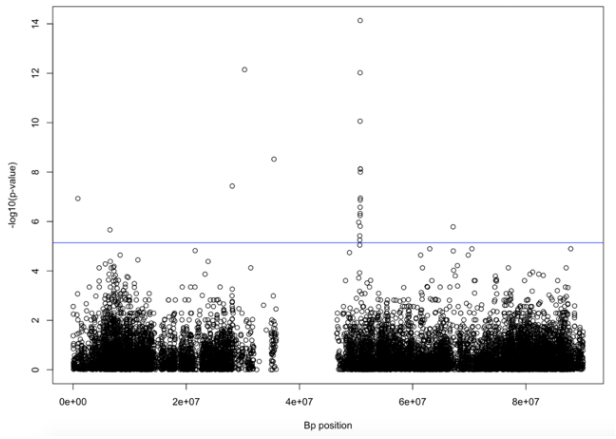
The genes that were declared as significant by three methods which single SNP analysis, SKAT and GHC are shown in Table 4. Single SNP analysis found eight significant regions. SKAT normal found four regions and the GHC method found four region.

5. Discussion and Conclusion

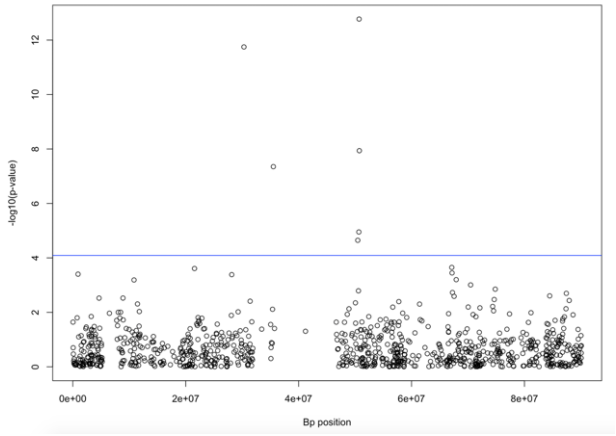
The findings confirm that the single SNP analysis is robust in producing an equal FP value when we change the disease SNP rs3789038 to rs3785142. On the other hand, the efficiency of the SKAT analysis for genome-wide association analysis is highly dependent on the disease-causing SNPs. Determining the disease SNP are affects TP and FP rate. We may have to consider the value of MAF, the correlation between SNP in the gene. The important thing in SKAT is the choice of weight and this must be carefully selected. It is necessary to find the appropriate weight for the data being studied by tuning the appropriate parameters to get a powerful test. Incorrect weight and data misalignment may lead to low

Table 3. The TP and FT rates of single SNP and SKAT analysis with rs3789038 and rs3785142 as disease SNPs and respective effect sizes of $b_1 = 0.2$ and $b_2 = 0.1$. and effect sizes of $b_1 = 0.1$ and $b_2 = 0.2$

Method	$b_1 = 0.2$ and $b_2 = 0.1$		$b_1 = 0.1$ and $b_2 = 0.2$	
	FP rate	TP rate	FP rate	TP rate
Single SNP analysis	0.00220	0.85	0.00270	0.94
SKAT default	0.00344	0.92	0.00482	0.38
SKAT Madsen & Browning	0.04139	0.93	0.06329	0.80
SKAT inverse mean	0.03321	0.94	0.05624	0.74
SKAT normal	0.03329	0.94	0.04820	0.93
GHC	0.01716	0.93	0.02308	0.96



(a)



(b)

Figure 4. The horizontal line of the permutation threshold for declaring significance of Single SNP analysis (a) and SNP-set method (b)

Table 4. Genes on Chromosome 16 declared as significant by Single SNP analysis, SKAT and GHC method using the permutation method threshold

Region	Single SNP analysis	SKAT	GHC
1	LMF1	-	-
2	RBFOX1	-	-
3	-	LOC646828	-
4	XPO6	-	-
5	SMG1P5	SMG1P5	SMG1P5
6	Intron Gene 151	Intron Gene 151	Intron Gene 151
7	Intron Gene 173	-	Intron Gene 173
8	Intron Gene 174, NOD2, CYLD	Intron Gene 174, NOD2, CYLD	Intron Gene 174, NOD2, CYLD
9	C16ORF70	-	-

test power. GHC results were more robust when the disease SNP were changed. In the case one disease SNP, GHC gave a small FP while TP was close to that of the other two methods. When considering the test power, GHC is better than the other two methods. But it is difficult to specify which method is better because the meaning of an effective model covers many issues including the power of the test and computational efficiency. We find that the GHC test is a statistical method for testing the SNP-set suitable for the SNP-set containing a finite set of the correlated marker. These SNP data have no variation so the covariance matrix cannot be calculated. It is necessary to remove the no variation SNP from the data set. In the process of checking and deleting no variation data, the analysis takes a lot of time when compared with the single SNP and SKAT test, so that while GHC is shown to be preferable in terms of error rates, it is disadvantageous in terms of computational efficiency. It is obvious that the testing power of the three methods are equal, but the time it takes to test the SKAT normal is minimal as shown in Table 5.

In the section of real data analysis, all of the methods found that many regions are significant Especially, Region 4 contributed the intron gene 174, NOD2 and CYLD gene, corresponding to many pieces of research that found NOD2 is identified as a highly significant variant in Crohn's disease (Michail, Bultron, & DePaolo, 2013). Sidiq and Yoshihama (2016) said that NOD2 plays a key role in the regulation of microbiota in the small intestine and is the strongest risk factor in ileal Crohn's disease. It is confirmed that NOD2 is one of the most critical genetic factors causing the disease. In this research, NOD2 was found in all of the methods. In this situation, SKAT has the advantage when compared to the other methods in terms of providing the same results but taking less time. We expect this method will help to get accurate and fast answers in tests to find genes region that affect other diseases as well.

Table 5. Processing time for testing in each method

Method	Average time per one replicate	Total time (hours)
Single SNP analysis	6 minutes	150 H (6.25 days)
SKAT normal	4 minutes	100 H (4.16 days)
GHC	30 minutes	750 H (31.25 days)

Currently, there are many genomic datasets that need to be analyzed to yield fast, accurate and efficient answers. For large datasets and high dimensional data, the SNP-set method such as SKAT and GHC test is an interesting tool in the repertoire of statistical analysis methods with the advantage that it is potentially powerful and provides the correct assumptions. For SKAT, this paper only considered the weighted linear kernel; planned further work will evaluate the use of different kernels within the genomic setting. Although the two methods for analysis the SNP-set offer more advantage than the single SNP analysis in terms of computation, the high false positive rate in the SKAT method remains a concern.

References

Barnett, I. (2015). GHC: Computes P-values for the Generalized Higher Criticism. R package version 1.0.

Barnett, I. (2018, October 9). Generalized Higher Criticism. Retrieved from https:// scholar.harvard.edu/ibarnett/ software/generalized-higher-criticism

Barnett, I., Mukherjee, R., & Lin, X. (2017). The Generalized Higher Criticism for testing SNP-set effects in genetic association studies. *Journal of the American Statistical Association*, *112*(517), 64-76.

Bush, W. S., & Moore, J. H. (2012). Chapter 11: Genome-wide association studies. *PLoS Computational Biology*, *8*(12), e1002822.

Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P., & Zondervan, K. T. (2011). Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6, 121–133.

Ionita-Laza, I., Lee, S., Makarov, V., Buxbaum, J. D., & Lin, X. (2013). Sequence kernel association tests for the combined effect of rare and common variants. *American Journal of Human Genetics*, *92*(6), 841-853.

Kirdwichai, P., & Baksh, M. F. (2019). The analysis of genomewide SNP data using nonparametric and kernel machine regression. *Journal of Applied Science*, *18*(1), 20-30.

Lee, S. (2017, April 7). SKAT Package. Retrieved from https:// cran.r-project.org/web/packages/SKAT/vignettes/SKAT.pdf

Lee, S., Emond, M. J., Bamshad, M. J., Barnes, K. C., Rieder, M. J., Nickerson, D. A., . . . Lin, X. (2012). Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *American Journal of Human Genetics*, *91*(2), 224-237.

Lee, S., Miropolsky, L., & Wu, M. C. (2017). SKAT: SNP-Set (Sequence) Kernel Association Test. R package version 1.3.2.1.

Lee, S., Wu, M. C., & Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, *3*(4), 762–775.

Lettre, G., Lange, C., & Hirschhorn, J. N. (2007). Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet Epidemiology*, *31*(4), 358-362.

Lewis, C. M. (2002). Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, *3*(2), 146–153.

Michail, S., Bultron, G., & DePaolo, R. W. (2013). Genetic variants associated with Crohn's disease. *The Application of Clinical Genetics*, 6, 25–32.

Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M., & Poland, G. A. (2001). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics*, *70*(2), 425-434.

Sidiq, T., Yoshihama, S., Downs, I., & Kobayashi, K. S. (2016). Nod2: A critical regulator of ileal Microbiota and Crohn's Disease. *Frontiers in Immunology*, *7*(367), 1-11.

Sookkhee, S., Baksh, M. F., & Kirdwichai, P. (2018). Efficiency of single SNP analysis and Sequence Kernel Association Test in genome-wide Association Analysis. *Proceedings of the International MultiConference of Engineers and Computer Scientists 2018*.

The Wellcome Trust Case-Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls.*Nature Publishing Group, 447*, 661-678.

Wu, M. C., Kraft, P., Epstein, M. P., Taylor, D. M., Chanock, S. J., Hunter, D. J., & Lin, X. (2010). Powerful SNP-set analysis for case-control genome-wide association studies. *American Journal of Human Genetics*, *86*(6), 929-942.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., & Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *American journal of Human Genetics*, *89*(1), 82-93.

Zeng, P., Zhao, Y., Qian, C., Zhang, L., Zhang, R., Gou, J., . . . Chen, F. (2014). Statistical analysis for genome-wide association study. *Journal of Biomedical Research*, *29*(4), 285-297.

Zhao, Y., Chen, F., Zhai, R., Lin, X., Diao, N., & Chritiani, D. C. (2012). Association test based on SNP set: Logistic kernel machine based test vs. Principal component analysis. *PLoS ONE*, *7*(9), 1-11.