Efficiency of Single SNP analysis and Sequence Kernel Association Test in Genome-wide Association Analysis

Sirikanlaya Sookkhee, and M. Fazil Baksh and Pianpool Kirdwichai *

Abstract—**This research compares the efficiency of the commonly used single SNP analysis and SNP-set analysis based on the recently proposed Sequence Kernel Association Test (SKAT) in the analysis of genome-wide studies of disease-gene association. False positive (FP) and true positive (TP) rates are evaluated for different genetic models of disease with significance thresholds adjusted for multiple testing based on the permutation method. Simulation results shows that although SKAT tends to be slightly more efficient in identifying true associations, this comes at the cost of a substantial increase in the false positive findings over the single SNP method and that this efficiency gain is also highly dependent on choosing an appropriate weight and correlation for the SNP sets in SKAT.**

Keywords: Single SNP Analysis, SNP-set Analysis, SKAT, GWAS, Bonferroni Correction, Permutation Test

1 Introduction

A genome-wide association study (GWAS) identifies genetic variants associated with disease and other medical conditions by analysing many (10^6) variants, such as single nucleotide polymorphisms (SNPs), distributed across the genome. Historically the most common hypothesis testing method employed is the single SNP analysis whereby each variant is singly tested for association with an adjustment, such as Bonferroni correction, for controlling the Type I error rate. This approach tend to be highly conservative and in addition, a large number of SNPs takes a lot of time to analyze. In attempts to solve these problems, researchers have investigated ways to maintain the efficiency of the test and to reduce the time to analyze by grouping the SNPs into SNP-sets before testing. This grouping may be by genomic features such as a gene or haplotype block and is intuitively appealing when the belief is that a group of closely linked SNPs underlies development of the disease.

Gauderman et al. [1] use principal components (PCs) analysis to compute combinations of SNPs that capture the underlying correlation structure within the locus then uses PCs in a test of disease association. Methods based on PCs include the standard principal components analysis (PCA) technique, supervised PCA, sparse PCA, and functional PCA, kernel PCA and sliced inverse regression (SIR), see [2]-[4]. Wu et al. [5] proposed grouping SNPs in a gene or haplotype block using a kernel machine and developed a test for the association between SNP-set and disease outcome and claim that grouping into SNP-sets correctly can lead to improving the power of the test. Currently, there is a tool implemented in R software [6] for association analysis of a SNP-set and disease status called Sequential Kernel Association Test (SKAT). SKAT tests for association between the set and continuous or dichotomous phenotypes using a kernel regression framework [7] and is potentially powerful under many different scenarios. It does not require any assumptions on the directionality of effects [8] and there are possible kernel choices for grouping such as linear kernel, weighted linear kernel, identical-by-state kernel (IBS), weighted IBS kernel, 2wayIX kernel, quadratic kernel and weighed quadratic kernel [7].

This research uses simulation data to compare efficiency and false positive rates of the single SNP analysis via logistic regression with SNP-set analysis via SKAT. SNP sets were grouped by gene. The haplotypes used in the simulations were constructed from the control dataset in the Welcome Trust Case Control Consortium (WTCCC) study of Crohn's disease [9]. It is well known that the probability of Type I error increases when testing multiple hypotheses and that Bonferroni correction, which replaces the original significance level α by α/m where m is the number of hypothesis tests, is an easy way to control the family-wise error rate (FWER). However, this approach leads to a conservative test and a constringent

*Manuscript received December 08, 2017; revised January 28, 2018. This work was supported in part by Graduate College of King Mongkut's University of Technology North Bangkok, Thailand. S. Sookkhee is with Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand., (e-mail: sirikanlaya.s@sskru.ac.th)., M. F. Baksh is with the Department of Mathematics and Statistics, School of Mathematical, Physical & Computational Sciences, University of Reading, U.K.,(e-mail: m.f.baksh@reading.ac.uk)., P. Kirdwichai is with Department of Applied Statistics, Faculty of Applied Science, King Mongkut's University of Technology North Bangkok, Bangkok, 10800, Thailand., (e-mail: pianpool.k@sci.kmutnb.ac.th).

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

threshold when m is large. The gold standard for finding an appropriate significance threshold that controls the type I error is the permutation method and therefore this method is instead used in this research for the single SNP and SNP-set analyses.

2 Methodology

2.1 Single SNP Analysis

Logistic regression models are used to evaluate the relationship between the disease and each SNP, separately. Logistic regression is a powerful and flexible technique for the statistical modeling of a binomial outcome. Logistic regression [10] 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. Logistic regression is often the preferred approach because it allows for adjustment for clinical covariates (and other factors), and can provide adjusted odds ratios as a measure of effect size. Logistic regression has been extensively developed, and numerous diagnostic procedures are available to aid interpretation of the model. The genotypes for an SNP can also be grouped into genotype classes or models, such as dominant, recessive, multiplicative, or additive models [11]. The additive model is commonly used as it has reasonable power to detect both additive and dominant effects, but it is important to note that an additive model may be underpowered to detect some recessive effects [12]. 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 T_{ij} be a genotype for the j^{th} SNP ($j = 1, 2, \ldots, m$) on the i^{th} individual ($i = 1, 2, \ldots, n$) and let P_{ij} be the probability of disease for this individual given covariates $\mathbf{C}_i = (C_{i1}, C_{i2}, \ldots, C_{iq})$ and the genotype at the j^{th} SNP. The logistic regression model is

$$logit(P_{ij}) = \alpha_0 + \alpha' \mathbf{C}_i + \beta_j T_{ij}, \tag{1}$$

where α_0 is an intercept term and $\alpha = [\alpha_1, ..., \alpha_q]'$ is a vector of regression coefficients for the q covariates. Here, $\beta_j = 0$ corresponds to the null hypothesis of lack of association between the j^{th} SNP and disease. The likelihood ratio test is used in the tests of association in this paper although either of the three asymptotically equivalent likelihood ratio, score or Wald tests, could have been used. Under the null hypothesis, any of the three test statistics has a chi-squared distribution with 1 degree of freedom.

2.2 Sequence Kernel Association Test

Sequence Kernel Association Test (SKAT) is a supervised test for the joint effects of multiple variants in a region of the genome on the disease or condition of interest. Regions can be defined by genes, haplotype block, principal component analysis, or sliding window. For each region, SKAT analytically calculates a p-value for association see [5], [11] and [12]. The logistic regression model now becomes

$$logit(P_{ik}) = \alpha_0 + \alpha' \mathbf{C}_i + \beta' \mathbf{T}_{ik} \tag{2}$$

where P_{ik} is the disease probability given the covariates $\mathbf{C}_i$ and the p variants $\mathbf{T}_{ik} = (T_{i1}, T_{i2}, ..., T_{ip})_k$ in the k^{th} SNP-set. Here $\beta = [\beta_1, \ldots, \beta_p]'$ is a vector of regression coefficients for the p observed variants and evaluating whether the variants are jointly associated with the disease corresponds to testing the null hypothesis $H_0 : \beta = 0$ that is $\beta_1, \beta_2, \ldots, \beta_p = 0$. SKAT tests H_0 by assuming each β_r follows an arbitrary distribution with mean of zero and a variance of $w_r \tau$, where τ is a variance component and w_r is a pre-specified weight for variant r. The null hypothesis $H_0 : \beta = 0$ is equivalent to the hypothesis $H_0 : \tau = 0$.

The SKAT test statistic can be written as

$$S = (\mathbf{y} - \hat{\mu})' \mathbf{K} (\mathbf{y} - \hat{\mu}) \tag{3}$$

where $\mathbf{K} = \mathbf{T}_k \mathbf{W} \mathbf{R}_\rho \mathbf{W} \mathbf{T}_k'$ is an $n \times n$ kernel matrix, $\mathbf{R}_\rho = (1 - \rho)\mathbf{I} + \rho \underline{1}\underline{1}'$ is an $p \times p$ compound symmetric matrix, $\underline{1} = (1, ..., 1)'$ and $\hat{\mu}$ is the predicted mean of $\mathbf{y}$ under H_0, that is $\hat{\mu} = logit^{-1}(\hat{\alpha}_0 + \mathbf{C}\hat{\alpha})$ where α_0 and $\hat{\alpha}$ are estimated under the null model by fitting the logistic regression model on only the covariates $\mathbf{C}$. Here T_k is a $n \times p$ matrix with the $(i, r) - th$ element being the genotype of variant r of the i^{th} individual in the k^{th} SNP-set and $\mathbf{W} = diag(w_1, ..., w_p)$ contains the weights of the p variants. In the case when the correlation $\rho = 0.0$ the SKAT test statistic in equation (3) can be simplified as the weighted sum

$$S_{SKAT} = \sum_{r=1}^{p} w_r^2 D_r^2 = \sum_{r=1}^{p} w_r^2 \left\{ \sum_{i=1}^{n} T_{ir}(y_i - \hat{\mu}_i) \right\}^2, \tag{4}$$

whereas when correlation $\rho = 1.0$, equation (3) becomes

$$S_{Burden} = \left\{ \sum_{r=1}^{p} w_r D_r \right\}^2 = \left\{ \sum_{r=1}^{p} w_r \sum_{i=1}^{n} T_{ir}(y_i - \hat{\mu}_i) \right\}^2, \tag{5}$$

where $D_r = \sum_{i=1}^{n} T_{ir}(y_i - \hat{\mu}_i)$ is the score statistic for testing $H_0 : \beta_r = 0$. Finally, it should be noted that $\mathbf{K}$ is an $n \times n$ symmetric matrix with elements $\mathbf{K}(\mathbf{T}_i, \mathbf{T}_{i'})$ that measures genetic similarity between the $i - th$ and $i' - th$ individuals in the study. Many choices for $\mathbf{K}$ are possible such as linear , weighted linear, identical-by-state (IBS), weighted IBS, 2wayIX, quadratic and weighed quadratic. The weighted linear kernel $\mathbf{K}(\mathbf{T}_i, \mathbf{T}_{i'}) = \sum_{j=1}^{p} w_{ij} T_i T_{i'}$ is used in this study. This kernel assumes that the disease depends on the variants in a linear fashion and is equivalent to the classical logistic model.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Choosing an appropriate weight is very important in SKAT because a good choice of weights can improve power of the test. Weight functions can be specified in the SKAT package in R using the Beta density function $Beta(p_r : a_1, a_2)$, where p_r is the estimated minor allele frequency (MAF) for SNP r in the SNP-set and a_1 and a_2 are pre-specified scale parameters of the Beta distribution. The default in the SKAT package $Beta(p_r : 1, 25)$ substantially up-regulates rare variants and down-regulates common variants. The Madsen and Browning weight $w_r = 1/\sqrt{p_r(1-p_r)}$ is equivalent to $Beta(p_r : 0.5, 0.5)$ and can pick up signals from both common and rare variants but is thought to suffer from low power. A third option $w_r = 1/\sqrt{p_r}$, which we call the inverse mean, is equivalent to $Beta(p_r : 0.5, 1)$. The final option considered in this paper is $Beta(p_r : 10, 10)$ which gives the appearance of a symmetrical distribution similar to the normal distribution. These weight functions are illustrated in Figure 1.

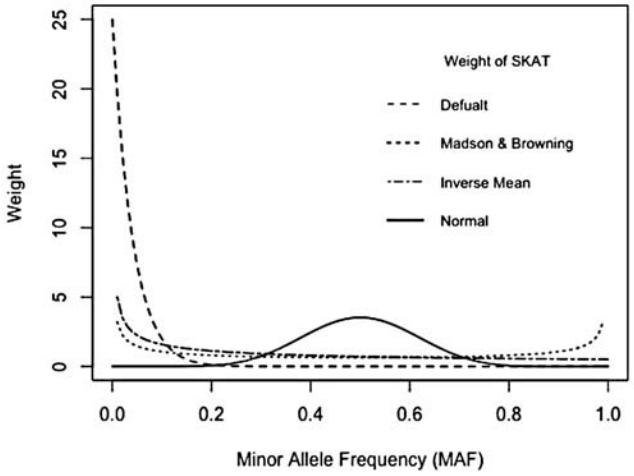

Figure 1: Examples of weight functions used in SKAT analysis.

2.3 Multiple Hypothesis Testing

The multiple hypothesis testing problem occur when we test many hypotheses simultaneously. For m independent tests and α the rejection level for each test, the probability of falsely rejecting at least one true null hypothesis, otherwise known at the family-wise error rate (FWER) increases with m in such a way that for even a moderate number of tests we will almost surely incorrectly reject at least one true null hypothesis, see Figure 2. The simplest method for controlling the FWER so that the probability of observing at least one significant result remains below the desired significance level is Bonferroni correction [13].

2.3.1 Bonferroni Threshold

Bonferroni correction adjusts the desired significance level from α to $\alpha' = \alpha/m$ where m is the number of statistical tests conducted. A SNP is then considered to be significant if its p-value is less than the α' adjusted significance level. As they are 13,479 SNPs and 914 SNP-sets in the WTCCC dataset of Crohn's disease, in this paper the α' adjustment for the single SNP analysis and SNP-set analysis when $\alpha = 0.05$ is 3.71×10^{-6} and 5.47×10^{-5}, respectively. Simulation results of single SNP analysis and SKAT analysis with $\rho = 0.0$ under the null model of no gene effect in Table 1 confirms that this adjustment leads to a type I error that is much lower than the desired level and therefore Bonferroni correction is quite conservative and constringent.

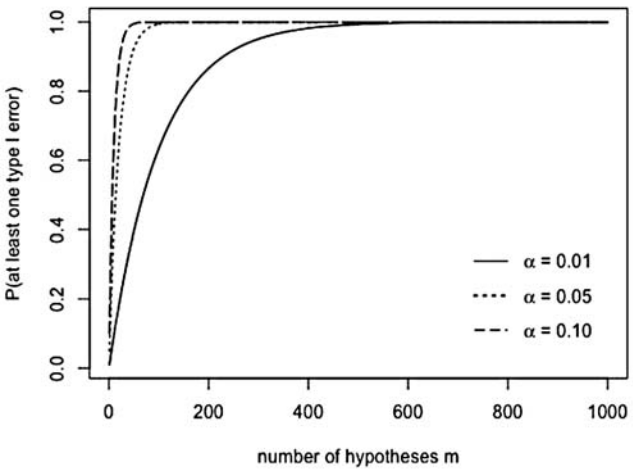

Figure 2: Probability of at least one false positive finding for different number of hypotheses m and significance level α.

Table 1: Achieved type I error of single SNP and SKAT analysis with correlation 0.0 under the null model in tests with $\alpha = 0.05$.

Method	Bonferroni Threshold	New Threshold
Single SNP analysis	0.033	0.059
Default	0.041	0.057
Madson and Browning	0.031	0.055
Inverse mean	0.034	0.059
Normal	0.035	0.055

2.3.2 Threshold based on permutation test

A permutation test is a nonparametric method for estimating the sampling distribution of a test statistic under the null hypothesis that a set of genetic variants has no effect on the outcome. This approach provides a highly reliable distribution of the test statistic but requires many samples generated under the null model. In this research, we use 10,000 replicates for computing the multivariate sampling distribution under the null hypothesis with no gene effect and to establish significance thresholds giving

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 2: The TP and FP rates of single SNP and SKAT analyses with rs3789038 as disease SNP and effect size β_1 = 0.2.

Correlation	Method	New Threshold	
		FP	TP
-	Single SNP	0.00105	0.71
$\rho = 0.0$	Default	0.00154	0.80
$\rho = 0.0$	Madsen & Browning	0.01073	0.86
$\rho = 0.0$	Inverse Mean	0.00981	0.86
$\rho = 0.0$	Normal	0.00939	0.85
$\rho = 1.0$	Default	0.00209	0.83
$\rho = 1.0$	Madsen & Browning	0.01235	0.86
$\rho = 1.0$	Inverse Mean	0.01156	0.86
$\rho = 1.0$	Normal	0.01164	0.86

Table 3: The TP and FP rates of single SNP and SKAT analysis with rs3785142 as disease SNP and effect size β_1 = 0.2.

Correlation	Method	New Threshold	
		FP	TP
-	Single SNP	0.00132	0.89
$\rho = 0.0$	Default	0.00209	0.00
$\rho = 0.0$	Madsen & Browning	0.03075	0.65
$\rho = 0.0$	Inverse Mean	0.02734	0.57
$\rho = 0.0$	Normal	0.00939	0.85
$\rho = 1.0$	Default	0.00467	0.00
$\rho = 1.0$	Madsen & Browning	0.03555	0.13
$\rho = 1.0$	Inverse Mean	0.03298	0.09
$\rho = 1.0$	Normal	0.03176	0.29

a type I error close to 0.05. We use linear interpolation for finding the thresholds. In this research, we called the threshold based on the permutation test new threshold. The $-\log_{10}$ transformation of the new threshold for single SNP analysis is 5.140. For SKAT analyses, the new thresholds for default weight with correlation 0.0 and 1.0 are 4.115 and 4.115 respectively, while the new thresholds for Madson and Browning weight with correlation 0.0 and 1.0 are 4.040 and 3.998. The new thresholds for inverse mean weight with correlation 0.0 and 1.0 are 4.00 and 3.970 and finally, the new threshold for normal weight with correlation 0.0 and 1.0 are 3.995 and 3.986, respectively. Type I error rates based on the new thresholds, shown in Table 1, suggest that the nominal type I error of 0.05 is achieved in all cases and therefore the new thresholds based on permutation test were selected for comparing the efficiency of the single SNP and SKAT methods.

2.4 The Data and Disease Model Simulation

The genotype data used in this simulations are 13,479 SNPs on Chromosome 16 from 1,504 unaffected individuals in the WTCCC study of Crohn's disease. Using 3008 haplotypes constructed from the 1504 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. There are a total of 7 SNPs in the data on gene HMOX2 with pairwise correlation ranging between 0.93 and 0.99, with median equal 0.99 while there are 8 SNPs in the data on CYLD having pairwise correlations between 0.51 and 0.99, with median equal to 0.93.b

The model for one disease SNP used to generate disease status is

$$P(\text{diseased}|T) = \frac{e^{\alpha_0+\beta_1 T}}{1+e^{\alpha_0+\beta_1 T}}, \tag{6}$$

where T is the number of copies of the rare allele of the disease SNP, α_0 is a pre-specified baseline relative risk of disease and β_1 is the gene effect, which in this study was set equal to 0.2. The disease model for two disease SNP is

$$P(\text{diseased}|T_1, T_2) = \frac{e^{\alpha_0+\beta_1 T_1+\beta_2 T_2}}{1+e^{\alpha_0+\beta_1 T_1+\beta_2 T_2}}. \tag{7}$$

This model assumes the two disease SNPs act linearly on the logit scale and two situation are investigated. The first is for gene effect $\beta_1 = 0.1$ and $\beta_2 = 0.2$ while in the second case the gene effects are fixed at $\beta_1 = 0.2$ and β_2 = 0.1.

3 Result

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) and whether the disease SNP is correctly identified (true positive) by the single SNP and SKAT methods described above. Results presented show the false positive (FP) and true positive (TP) detection rates for the two methods.

Case 1: One disease SNP

Shown in Table 2 are the TP and FP rates of single SNP and SKAT analyses of rs3789038 as disease SNP with gene effect size $\beta_1 = 0.2$. The FP rate of the single SNP

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

Table 4: The TP and FT rates of single SNP and SKAT analysis with rs3789038 and rs3785142 as disease SNPs and respective effect sizes of $\beta_1 = 0.1$ and $\beta_2 = 0.2$.

Correlation	Method	New Threshold	
		FP	TP
-	Single SNP	0.00270	0.94
$\rho = 0.0$	Default)	0.00482	0.38
$\rho = 0.0$	Madsen & Browning	0.06329	0.80
$\rho = 0.0$	Inverse Mean	0.05624	0.74
$\rho = 0.0$	Normal	0.04820	0.93
$\rho = 1.0$	Default	0.00798	0.40
$\rho = 1.0$	Madsen & Browning	0.07008	0.51
$\rho = 1.0$	Inverse Mean	0.06644	0.49
$\rho = 1.0$	Normal	0.06204	0.60

Table 5: The TP and FT rates of single SNP and SKAT analysis with rs3789038 and rs3785142 as disease SNPs and respective effect sizes of $\beta_1 = 0.2$ and $\beta_2 = 0.1$.

Correlation	Method	New Threshold	
		FP	TP
-	Single SNP analysis	0.00290	0.88
$\rho = 0.0$	Default	0.00408	0.93
$\rho = 0.0$	Madsen & Browning	0.05263	0.95
$\rho = 0.0$	Inverse Mean	0.04692	0.95
$\rho = 0.0$	Normal	0.04113	0.95
$\rho = 1.0$	Default	0.00708	0.94
$\rho = 1.0$	Madsen & Browning	0.05875	0.95
$\rho = 1.0$	Inverse Mean	0.05573	0.95
$\rho = 1.0$	Normal	0.05222	0.95

analysis is seen at 0.00105 to be roughly at least 9 times lower than the SKAT analyses but this comes at the lower TP rate of 0.71. Considering the result of SKAT analysis, all have very similar TP rates with TP between 0.80 - 0.86 when ρ as 0.0 and TP between 0.83 - 0.86 when ρ as 1.0. With SNP rs3789038 as disease SNP, use of the default weight gives the lowest TP and lowest FP, while there is very little difference in the rates for the other three methods, irrespective of the value for the correlation.

The TP and FP rates of the single SNP and SKAT analyses with rs3785142 as disease SNP and gene effect $\beta_1 = 0.2$ are provided in Table 3. Here it can be seen that the single SNP analysis outperforms SKAT under all conditions. The FP rate of 0.00132 is comparable to that for rs3789038 and is at least 60% lower than for the SKAT analyses while the TP of 0.89 is substantially higher than all the SKAT methods, except the normal method with correlation 0.0 which has TP of 0.85 but also has an FP rate 9 times higher.

Considering only the results of the SKAT analyses in Table 3, the dependence on the correlation is highly evident with ρ equal to 0.0 tending to lead to higher TP rates than when it takes the value 1. The exception to this is the default method which has no power to detect the disease SNP, irrespective of the correlation used and illustrates the influence of the weights used on the findings; the minor allele of rs3785142 is not rare (MAF=0.48) and therefore is not consistent with the default weight which up-weighs rare variants and down-weighs common variants. An appropriate weight in this case is the normal and this is confirmed by the simulation results which shows that analysis under the normal weight is optimal with highest TP and lowest FP.

Case 2: Two disease SNPs

Shown in Table 4 are comparisons of TP and FP rates for single SNP and SKAT analyses using the new threshold and computed from 1,500 replicates with a gene effect size for disease SNP rs3789038 of $\beta_1 = 0.1$ and effect size for rs3785142 of $\beta_2 = 0.2$. Here the single SNP method is found to be optimal with the highest TP rate (0.94) and lowest FP rate (0.00270) of all methods considered. This is consistent with the findings in Table 3 above and confirms that the single SNP analysis is preferable to SKAT when the disease SNPs have high MAF. Considering the results of the SKAT analyses only, the dependence on correlation is again clearly seen with the assumption of uncorrelated SNPs in the SNP-set leading to higher TP and slightly lower FP rates in all but the default method.

Comparisons of single SNP analysis and SKAT analyses using the new threshold for respective gene effects of $\beta_1 = 0.2$ and $\beta_2 = 0.1$ for disease SNPs rs3789038 and rs3785142 are shown in Table 5. The results show that the single SNP analysis FP rate of 0.00290 is once again much lower than for any of the SKAT analyses but her the TP rate of 0.88 is also slightly lower. Considering the SKAT analysis only, a high TP rate independent of the correlation is observed. This is consistent with the findings in Table 2 which is not surprising as SNP rs3789038 is driving the disease-gene relationship in the simulation model.

4 Conclusions and Future Work

The findings confirm that the single SNP analysis is consistent in producing lower numbers of false positives than SKAT while also having a competitive and fairly consistent true positive rate. On the other hand, efficiency of

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the SKAT analysis for genome-wide association analysis is highly dependent on the disease causing SNPs. It is clearly seen that the choice of weight is very important and so must be carefully selected but how this can be achieved is uncertain as very little is known about the disease predisposing loci in a genome-wide study. In addition, as the value of the correlation used impacts on the SKAT results in the sense that the assumption of no correlation of the SNPs in a SNP set can lead to a higher true positive rate when in fact the SNPs on the disease causing gene are highly correlated, the decision as to what value of correlation to use will require careful thought.

Currently, there are many genomic datasets that need to be analyzed for fast, accurate and efficient answers and SKAT is an interesting tool in the repertoire of statistical analysis methods with the advantage that it is potentially powerful, provided the correct assumptions are made. However the high false positive rate remains a concern and methodology for reducing this is currently being developed and evaluated. Additionally this paper only considered the weighted linear kernel; planned further work will evaluate use of different kernels within the genomic setting.

References

[1] W.J Gauderman, C. Murcray, F. Gilliland, D.V. Conti, "Testing Association Between Disease and Multiple SNPs in Candidate Gene," *Genet Epidemiol*, vol.31, no.5, pp. 383-395, Jul 2007.

[2] S. Ma, Y. Dai, "Principal component analysis based method in bioninformatics studies," *Briefings in Bioinformatics*, vol.12, no.6, pp. 714-722, January 2011.

[3] Y. Zhao, F. Chen, R. Zhai, X. Lin, N. Diao, D.C. Christiani, "Association Test Based on SNP Set: Logistic Kernel Machine Based Test vs. Principal Component Analysis," *PLOS ONE*, vol.7, no.9, September 2012.

[4] M. Cai, H. Dai, Y. Qiu, Y. Zhao, R. Zhang, M. Chu, J. Dai, Z. Hu, H. Shen, F. Chen, "SNP Set Association Analysis for Genome-wide Association Studies," *PLOS ONE*, vol. 8, no. 5, May 2013.

[5] M.C. Wu, P. Kraft , M.P. Epstein, D.M. Taylor, S.J. Chanock, D.J. Hunter, X. Lin, "Powerful SNP-Set Analysis for Case-Control Genome-wide Association Studies," *American Journal of Human Genetics*, vol. 86, pp. 929-942, June 2010.

[6] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing,Vienna, Austria, 2016. Available: https://www.R-project.org/

[7] S. Lee, with contributions from L. Miripolsky, M. Wu., *Package SKAT*, Available: https://cran.r-project.org/web/packages/SKAT/SKAT.pdf (Visited on 2016, June, 28).

[8] H. Chen, J.B. Meigs, J. Dupuis, "Sequence Kernel Association Test for Quantitative Traits in Family Samples," *Genet Epidemiol*, vol.37, no.2, pp. 196-204, February 2013.

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

[10] W.S. Bush, J.H. Moore, "Chapter 11: Genome-Wide Association Studies," *PLOS Computational Biology*, vol.8, no.12, December 2012.

[11] C.M. Lewis, "Genetic association studies: design, analysis and interpretation," *Briefing in Bioinformatics*, vol.3, no.2, pp.146–153, June 2002.

[12] G. Lettre, C. Lange, J.N. Hirschhorn, "Genetic model testing and statistical power in population-based association studies of quantitative traits," *Genet Epidemiol*, vol.31, no.4, pp. 358-362, May 2007.

[13] P. Zeng, Y. Zhao, C. Qian, L. Zhang, R. Zhang, J. Gou, J. Liu, L. Liu, F. Chen, "Statistical analysis for genome-wide association study," *The Journal of Biomedical Research*, vol. 29, no. 4, pp. 285-297, 2015.

[14] M.C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, X. Lin, "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test," *The American Journal of Human Genetics*, vol. 89, pp. 82-93, July 2011.

[15] S. Lee, M.J. Emond, M.J. Bamshad, K.C. Barnes, M.J. Rieder, D.A. Nickerson, D.C. Christiani, M.M. Wurfel, "Optimal Unified Approach for Rare-Variant Asssociation Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies," *The American Journal of Human Genetics*, vol. 91, pp. 224-237, August 2012.

ISBN: 978-988-14047-8-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)