Report for Master's Project

A Goodness-of-fit Association Test for Whole Genome Sequencing Data

Li Yang

Advisor: Zheyang Wu

Department of Mathematical Sciences

Worcester Polytechnic Institute

Abstract

Although many genetic factors have been successfully identified for human diseases in genome-wide association studies (GWAS), genes discovered to date only account for a small proportion of overall genetic contributions to many complex traits. Association studies have difficulty in detecting the remaining true genetic variants that are either common variants with weak allelic effects, or rare variants that have strong allelic effects but are weakly associated at the population level. In this work we applied a goodness-of-fit test for detecting sets of common and rare variants associated with quantitative or binary traits by using whole genome sequencing (WGS) data. This test has been proved optimal for detecting weak and sparse signals in the literature, which fits the requirements for targeting the genetic components of missing heritability. Furthermore, this p-value-combining method allows one to incorporate different data and/or research results for meta-analysis. The method was used to simultaneously analyse the WGS and GWAS data of Genetic Analysis Workshop (GAW) 18 for detecting true genetic variants. The results show that goodness-of-fit test is comparable or better than the influential sequence kernel association test in many cases.

Background

According to the Catalog of Genome-Wide Association Studies updated by the National Human Genome Research Institute, about 7,260 SNPs have been identified for 770 traits in 1,360 publications till November 2012. However, researchers believe that a significant proportion of heritability of many complex traits is still missing [1, 2]. The remaining genetic variants to be detected are either common variants with small allelic effects, or rare variants with relatively strong allelic effects. In both cases, the genetic effects are weak at the population level. Furthermore, only a small proportion of the avalanche of candidate variants are likely associated with a trait, which is a problem closely related to sparse signal discovery in statistics. It is very challenging to detect weak and sparse genetic effects via association.

To address this issue, we adopt a goodness-of-fit test (GOFT) [3] that has been proved to be optimal under a Gaussian means model [4]. That is, the boundary of the reliable detection of this method reaches the lowest possibility among all statistical tests when the signals are weak and sparse. Since the Gaussian means model is asymptotically equivalent to regression models [5], the goodness-of-fit test is promising in detecting weak and sparse genetic effects through regression model fitting. In this work, we illustrate how to apply the test to WGS data by using the GAW18 data. The method is assessed under various rare variant collapsing strategies, and compared with the sequence kernel association test (SKAT) [6]. Moreover, because GOFT is a method combining p-values, it has the potential to be used as a meta-analysis for incorporating data from different sources. Since in this GOFT statistic we only predetermined one integer and it may difficult to make a sensible choice, we also develop adaptive goodness-of-fit test (AGOFT) which allows us to take multiple choices of truncated point. We apply both GOFT and AGOFT to simultaneously analyse WGS data and GWAS data for detecting genetic loci associated with systolic blood pressure (SBP). The results show that even without the sophisticated weighting scheme, GOFT is comparable to, and sometimes better than, SKAT under its best weighting scheme. In addition, at small p-value cut-offs, the GOFT meta-analysis provides higher power than that when only WGS data was used.

Methods

Method 1: Sequence kernel association test (SKAT)

Sequence kernel association test (SKAT) [6] is a supervised and flexible test for the associations between sets of genetic variants and a continuous or dichotomous trait. Through adjusting the variance of the random effect coefficients of the genetic variants, SKAT can consider different weights for different variants in contributing to the response trait. Typically, the rare variants are assigned with larger weights than the common variants based on the rare-variant-common-disease model [7]. We use the R package SKAT [8] for the WGS data analysis.

Method 2: Goodness-of-fit test (GOFT)

The problem of determining the associations between a set of genetic variants and the trait can be viewed as a multiple hypotheses testing problem. Under the null hypothesis that there is no genetic association, the p-value from each genetic variant follows a uniform(0,1) distribution. So testing a group of variants can be considered as a goodness-of-fit test (GOFT) that measures the consistency between the empirical

distribution of the observed p-values and the uniform distribution. Here we adopt a GOFT from Berk and Jones [3], which was proposed from large deviation theory, and then was proved optimal in detecting weak and sparse signals [4]. Let $p_{(1)} \leq \cdots \leq p_{(L)}$ be the sorted p-values from L individual variants and the trait. The GOFT statistic is

(1) $$G = L \cdot max_{1 \leq j \leq \frac{L}{2}} K(\frac{j}{L}, p_{(j)}), \quad \text{where,}$$

$$K(t,x) = \begin{cases} tlog\left(\frac{t}{x}\right) + (1-t)\, log\left(\frac{1-t}{1-x}\right), & \text{if } 0 < x < t < 1, \\ 0, & \text{if } 0 \leq t \leq x \leq 1, \\ +\infty, & \text{Otherwise.} \end{cases}$$

Comparing with many p-value based SNP-set testing methods that sum up all p-values together in certain formulas [9], GOFT looks for the most representative p-value to the SNP set. At the same time, unlike the minimal p-value method that fixes the smallest p-value to represent the set, GOFT adapts to the signal pattern through the maximization procedure. Such adaptation is critical because the p-value of a true association is not necessarily the minimal one, especially when true associations are sparse and weak [4]. Another advantage is that the GOFT statistic only requires information from a set of p-values to work, so it can be flexibly applied to different genetic studies based on the corresponding appropriate p-values, or to meta-analyses that incorporate various data sources.

A permutation test can be applied to accommodate the various sizes of variant sets and the LD structures among the variants. Specifically, let G_s and G_{sm}, $s = 1, \ldots, S, m = 1, \ldots, M$, denote the GOFT statistics of the s^{th} genome segment window from the original data and from the m^{th} permutation of the genotype data, respectively. The empirical p-value for the s^{th} window is $p_s = \#\{ G_{sm} \geq G_s, m = 1, \ldots, M\}/M$. The number of permutations M =1000 was used in the following data analysis.

Method 3: Adaptive Goodness-of-fit test (AGOFT)

Moreover, to use the GOFT statistic, we must get the truncated point L a priori. When the number of tests is large, it is hard to take a reasonable choice of L. Then we develop an adaptive goodness-of-fit test (AGOFT) which allows us to have multiple choices of points. Instead of one typical point L, we can obtain the association evidence on each of L candidate points, $K_1 \leq \cdots \leq K_L$. Specifically we define $\hat{s}(K_l)$ be the estimated p-value of $G(K_l)$, $1 < l <$ L, the statistic of minimum p-value can be defined

$$MinP = min_{1 \leq l \leq L} \hat{s}(K_l)$$

The *MinP* algorithm can be shown as below:

1. Calculate the GOFT statistics for each truncated point in each permutation as

$$G_{lm} = l \cdot max_{1 \leq j \leq L} K\left(\frac{j}{l}, p_{(jm)}\right)$$

   Where G_{lm} is calculated from the l truncated point and the m^{th} permutation.

2. Based on G_l from both the observed SNP data and the permuted data, we use all the G_l to form a common distribution for the significance level of each time of permutation. Specifically we estimate the p-value for G_l as

$$s_{lm} = \frac{\sum_{m^*=0}^{M} I(MinP_{(m^*)} \geq MinP_{(m)})}{M+1}$$

3. Use a calculation similar to step 2, the AGOFT statistic from the m^{th} permutation data by

$$\frac{\sum_{m^*=0}^{M} I(MinP_{(m^*)} \leq MinP_{(m)})}{M+1}$$

Collapsing of rare variants

For the association study of complex diseases based on WGS data, a major challenge is to address rare variants that have weak statistical association due to small allele frequency. The GOFT is asymptotically optimal for weak and sparse signals, and is a right fit in this scenario. At the same time, because the effects of missense rare alleles are mostly in the same deleterious direction [10], collapsing the rare variant before GOFT is likely more efficient [11]. Furthermore, because common and rare variants contribute to complex diseases, it is good to combine information from both to facilitate the detection of associated genome

segments. Following a literature work [12], we collapse rare single nucleotide variants (SNVs) that allocate between adjacent common SNVs by summation of their genotype. Then the p-values for associations of both collapsed rare variants and common variants are obtained and fed into GOFT test statistic in (1) to study the overall significance of variant-groups.

Results

For evaluating the above association tests, we used the WGS "dose" file of 1,215,399 SNVs and the GWAS file of 65,519 SNVs on chromosome 3 as the genotype data. The quantitative trait was the systolic blood pressure (SBP) for the 142 independent individuals who have no missing genotype. To assess how the SNV-group size may affect the performance of these tests, we split chr3 into segments of fixed windows with one of three widths: 10kbp, 100kbp and 500kbp. Respectively at those three levels, the grouping strategy resulted in 19,472, 1,950, and 391 windows, among which 87, 37, and 20 windows contain true SNVs that are either non-synonymous or regulatory to SBP according to the GAW18 simulation [13]. The true windows and the 200 simulation replicates were used for evaluating power and type I error rate. We defined a SNV as a rare variant if it has minor allele frequency (MAF) less than 5%. The knowledge of the simulated true SNVs was only used for evaluating the power of the association tests, not for designing the tests and the data analysis strategies.

For GOFT, we assessed its type I error rates estimated by the false positive rate of the 19,385 false 10kpbs-windows on chr3 over a sequence of cut-offs. The type I error rate was well controlled (results are available upon request due to limited publication space). Figure 1 shows the assessment for the power of GOFT in detecting overall genetic associations, which was estimated by the true positive rate of true association windows based on GAW18 simulation replicate 1. We considered various window sizes with and without rare variant collapsing. Larger windows provided higher power at large cut-offs, but not at the small p-values that are often used in practice. This is because large windows likely had more noise variants, which diluted the signals from true variants, and thus were harder to get very small p-values. In the meanwhile, rare variant collapsing did help to increase the power in general.

Under the window size of 10kbps, we assessed SKAT with different strategies of weighting variants: flat weight, Beta(1, 1), Beta(0.5, 0.5), Beta(1, 25), and logistic(0.07, 150). Figure 2 shows the power of detecting the 87 true 10kbps-windows on chr3 over a variety of p-value cut-offs. The Beta(1, 25) and logistic weights performed better for small p-value cut-offs. Figure 2 also shows that GOFT was similar to the best SKAT setups for small p-values. In fact, GOFT had a larger area under curve (AUC) than SKATs when comparing their whole ROC curves (results are available upon request).

In order to study the performance in detection various patterns of genetic effects, we compared the power of GOFT and logistic-weight-SKAT in detecting each of all 87 true 10kbps windows on chr3. The power was estimated by the true positive rate of a true window among 200 replicates. There are three patterns of comparisons: GOFT was better in 42 windows (Figure 3 left panel), SKAT was better in 27 windows (Figure 3 middle panel), and both were similar. Figure 3 illustrates examples of these comparisons based on ROC curve (complete results are available upon request). GOFT seems better overall, but the comparison is not very significant (42 vs. 27, with p-value 0.10). The type I error rate was well controlled here (results available upon request).

Because GOFT only requires p-values as the input, it has a potential to be used in a meta-analysis for incorporating data from different studies. Here we evaluated how much the GWAS data could contribute useful information to the WGS study. By mapping the "rs" IDs to the Chromosome Report from dbSNP, we calculated the p-values of 65,519 GWAS SNVs on chr3. On average 3.4 GWAS SNVs were added into each window (about 5% increase). Figure 4 shows that the type I error rate control after adding the GWAS SNVs was still good (left panel) and that adding GWAS data helped to improve the power of GOFT in detecting true 10kbps windows on chr3 in general.

As we have said before, when the number of individual tests is large, it is difficult to make a sensible choice of the truncated point. Figure 5 states that in those three windows which contain the most significant variance among all the true SNPs, GOFT has a better power than AGOFT.

Discussion

We attempt to address the low power issue of association tests for WGS data from two aspects. First, we prefer to utilize tests specially designed for detecting weak and spares genetic effects. For this purpose, the GOFT and AGOFT are asymptotically optimal in the sense that their asymptotic detection boundary is one of the lowest boundaries among all statistical methods. If signals are weaker or sparser than this boundary, no statistical methods would work well anyway. Second, we try to borrow information from other data sets through meta-analysis. Although there has been some debate on how much of total heritability could be explained by GWAS data [7], the common agreement is that both common and rare variants contribute to complex diseases. It is potentially helpful to add GWAS data into WGS in order to increase the power. Our results show that both attempts are promising.

At the same time, several future works could be considered based on the limitations of the current study. First, the sample size is likely still small for either verifying asymptotical results or larger power of detecting weak genetic effects simulated in the data. It would be nice to further confirm the patterns of comparisons among these association methods by simulated and real data with much larger sample size. Second, gene-based collapsing can be applied as an alternative to the window scheme used here. Third, we have applied a simple rare variant collapsing process by direct summing the genotypes. This collapsing strategy is less sophisticated than the weighting strategy of SKAT. In fact, GOFT can further incorporate more successful collapsing strategies to improve its power, for example, by weighting the SNPs, like what SKAT and other methods have adopted [14-17]. Fourth, GOFT represents a first stage analysis, which only seeks to answer where the associations are located at; additional analyses could be required to determine the number and exact location of causal signals. Lastly, the adaptive idea may cause puzzle when we use multiple truncated point. This method may miss some key point for truncation.

Conclusions

We adopt goodness-of-fit test (GOFT) and adaptive goodness-of-fit test (AGOFT) to WGS data analysis for detecting disease-associated genomic segments. It is compared with the sequence kernel association test (SKAT) by using the GAW18 simulation data with SBP_1 as response. Even without a sophisticated weighting scheme, GOFT is comparable to or better than SKAT with the best weighting scheme in many cases. GOFT can be applied to a combination of GWAS and WGS data. Our results show that such meta-analysis has potential to provide higher power over WGS data analysis only. In all cases, the power is still low for detecting overall heritability under the sample size of 142 independent individuals for genetic association study.

Acknowledgements

We are grateful to the NIH funding support for GAW18 and for the student travel awards to Yang and Xuan. We are grateful to WPI Computing and Communications Center for computational support. Also thanks for the academic advice from Professor Zheyang Wu, and the help from Jing Xuan, Taoye Chen and Yiwei Liu.

References

1. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, Ioannidis JPA, Hirschhorn JN: **Genome-wide association studies for complex traits: consensus, uncertainty and challenges.** Nature Reviews Genetics 2008, **9**(5):356-369.

2. Kraft P, Hunter DJ: **Genetic Risk Prediction--Are We There Yet?** N Engl J Med 2009, **360**(17):1701.

3. Berk RH, Jones DH: **Goodness-of-fit test statistics that dominate the Kolmogorov statistics.** Probability Theory and Related Fields 1979, **47**(1):47-59.

4. Donoho D, Jin J: **Higher criticism for detecting sparse heterogeneous mixtures.** The Annals of Statistics 2004, **32**(3):962-994.

5. Cai TT, Zhou HH: **Asymptotic equivalence and adaptive estimation for robust nonparametric regression.** The Annals of Statistics 2009, **37**(6A):3204-3235.

6. 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.** American Journal of Human Genetics 2011, **89**:82-93.

7. Iyengar SK, Elston RC: **The genetic basis of complex traits: rare variants or "common gene, common disease"?** Methods Mol Biol 2007, **376**:71-84.

8. Lee S, Miropolsky L, Wu M, Lee MSS: **Package 'SKAT'.** 2011, .

9. Luo L, Peng G, Zhu Y, Dong H, Amos CI, Xiong M: **Genome-wide gene and pathway analysis.** European Journal of Human Genetics 2010, **18**(9):1045-1053.

10. Kryukov GV, Pennacchio LA, Sunyaev SR: **Most rare missense alleles are deleterious in humans: implications for complex disease and association studies.** The American Journal of Human Genetics 2007, **80**(4):727-739.

11. Li B, Leal SM: **Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data.** The American Journal of Human Genetics 2008, **83**(3):311-321.

12. Xiong M, Zhao J, Boerwinkle E: **Generalized T^2 Test for Genome Association Studies.** The American Journal of Human Genetics 2002, **70**(5):1257-1268.

13. Almasy L, Dyer TD, Peralta JM, Jun G, Fuchsberger C, Almeida MA, Kent JW Jr, Fowler S, Duggirala R, Blangero J: **Data for Genetic Analysis Workshop 18: Human Whole Genome Sequence, Blood Pressure, and Simulated Phenotypes in Extended Pedigrees.** .

14. Madsen BE, Browning SR: **A groupwise association test for rare mutations using a weighted sum statistic.** PLoS genetics 2009, **5**(2):e1000384.

15. Ionita-Laza I, Buxbaum JD, Laird NM, Lange C: **A new testing strategy to identify rare variants with either risk or protective effect on disease.** PLoS genetics 2011, **7**(2):e1001289.

16. Han F, Pan W: **A data-adaptive sum test for disease association with multiple common or rare variants.** Hum Hered 2010, **70**(1):42-54.

17. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR: **A method and server for predicting damaging missense mutations.** Nature methods 2010, **7**(4):248-249.

Figures

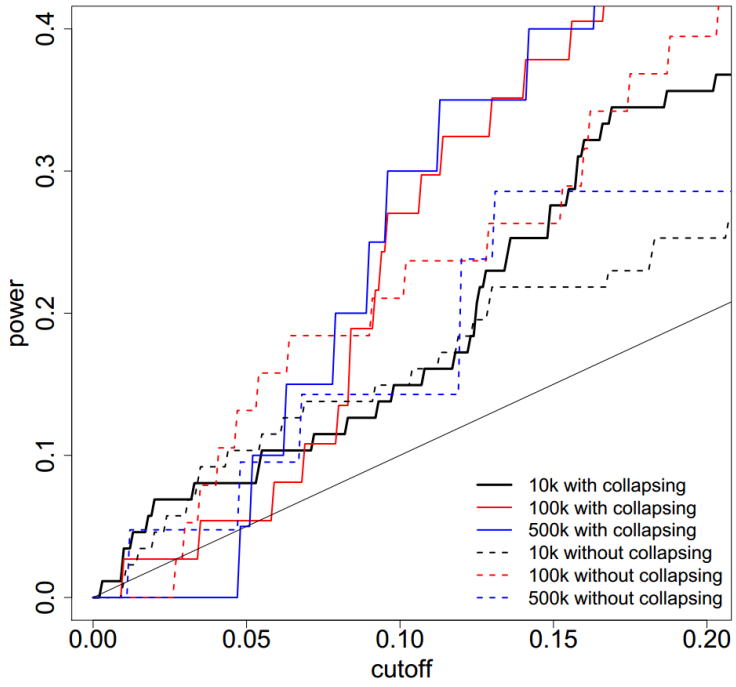

Figure 1 - Power of GOFT for different window sizes with or without collapsing variants

Power is estimated by the true positive rate of true association windows on chr3 based on GAW18 simulation replicate 1.

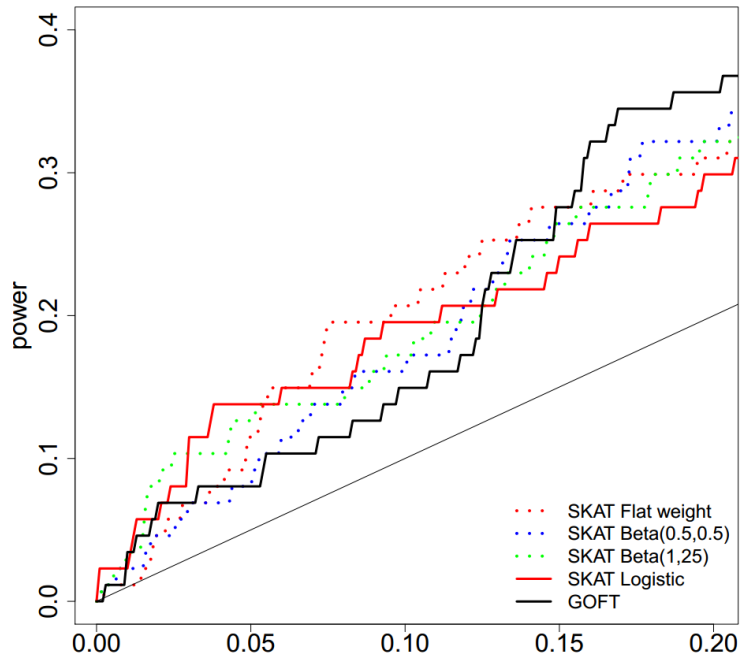

Figure 2 - Power of GOFT and SKAT under different weighting schemes

Power is estimated by the true positive rate of 87 true 10kbps windows on chr3.

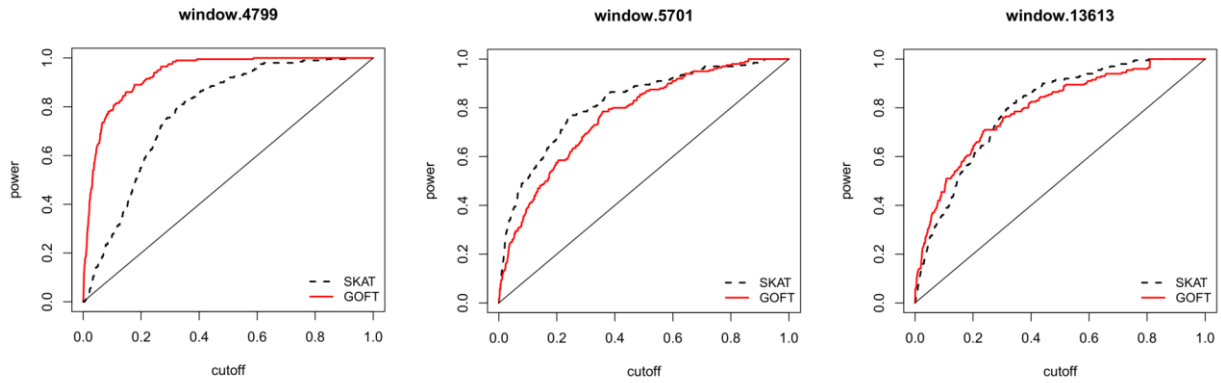

Figure 3 - Comparison patterns between GOFT and SKAT for detecting true windows

Left: window 4799 illustrates a case where GOFT is better; middle: window 5701 is an example where SKAT with logistic-weight is better; right: window 13613 is an example that both methods are similar.

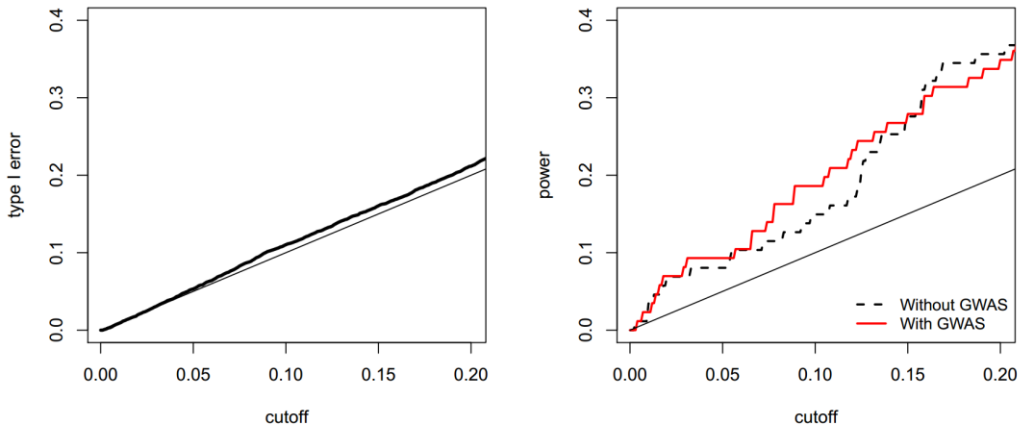

Figure 4 - Type I error rate and power for GWAS-WGS meta-analysis

Left: Empirical type I error rate (i.e., false positive rate) in the meta-analysis; right: power of detecting the 87 true 10kpbs windows on chr3 when GWAS data were added or not.

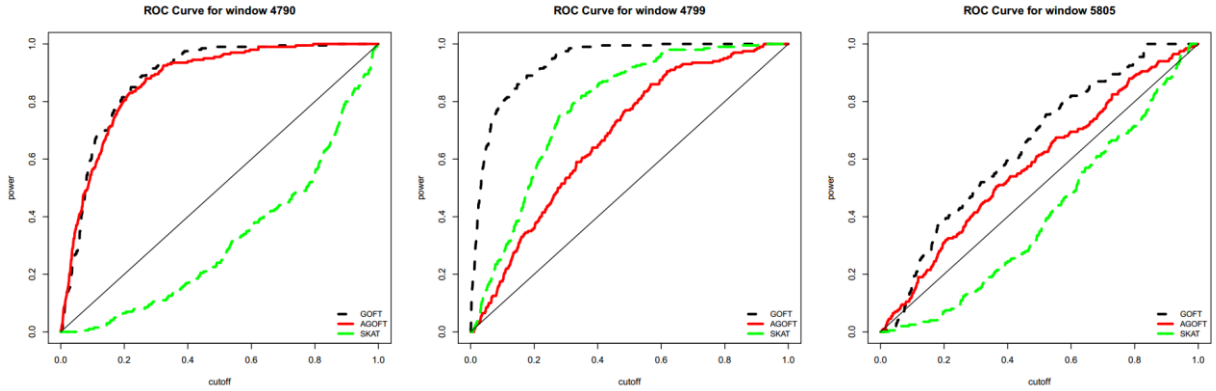

Figure 5 - Comparison patterns between GOFT and SKAT for detecting true windows

Left: window 4790 illustrates a case where GOFT is the best; middle: window 4799 is an example where SKAT with logistic-weight is better than AGOFT; right: window 5805 is an example that all methods are similar.