A Genetic Association Study Comparing Kernel-based Methods, with Application to Crohn's Disease

BSc Peter Tea, Msc Zhe Gao and Dr. Kelly Burkett

Department of Mathematics and Statistics, University of Ottawa

uOttawa

Introduction

- Association studies test for correlation between genetic variation and phenotype variation (Ex: Kernel-based methods). These studies can locate candidate genes that contribute to the onset of a disease
- Kernel methods require: 1) A kernel function and; 2) A kernel association statistic. BUT, there exists **MANY** different kernel functions and association statistics to choose from!
- Research goal: Compare performances of different combinations of kernel functions and kernel statistics, under 2 distinct phenotype models

Background Information

Kernel functions

- Maps the degree of genetic similarity between pairs of individuals
- Two domains of kernel functions:
  - I) Scoring genotype kernels: Count the number of shared alleles across all SNP sites
  - II) Tree-based kernels: Use branch lengths of Gene tree

- Gene tree tracks descent of haplotypes from a common ancestral haplotype
- Haplotypes that are genetically similar tend to cluster next to each other on the tree

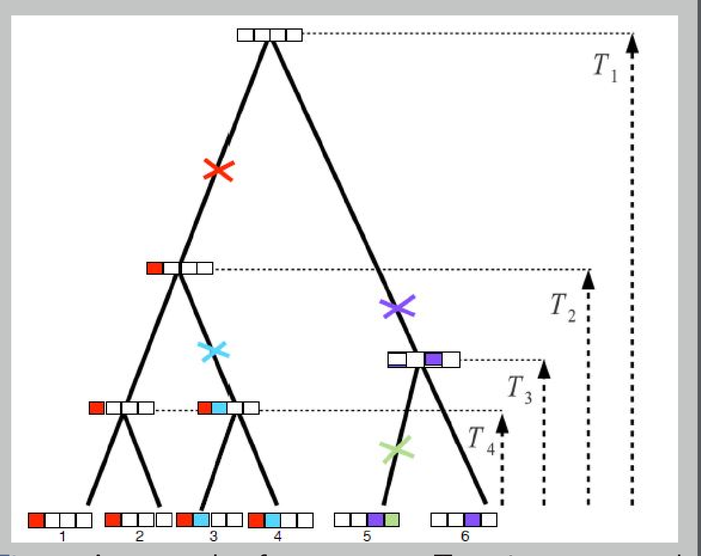

Figure: An example of a gene tree. Tree tips correspond to extant haplotypes; Tree root represents ancestral haplotype.

Kernel Statistics

- Many statistic approaches assume a regression model. For example, SimReg's approach is:

$$Z_{ij} = \beta \cdot S_{ij} + \epsilon_{ij}$$

where:

Z_{ij} is the cross product of the phenotype residuals between subjects i and j $(i \neq j)$; S_{ij} is a kernel function measuring genetic similarity; ϵ_{ij} are error terms

- Hypothesis tests are all similar:

$$H_0 : \beta = 0 \quad \text{vs.} \quad H_1 : \beta \neq 0$$

Simulation

Continuous phenotypes simulated from a normal distribution, where the mean depends on the number of causal variants

- Phenotype 1: Single common causal variant. $(\mathbf{0.2} < MAF < \mathbf{0.35})$
- Phenotype 2: Multiple rare causal variants ($MAF < \mathbf{0.05}$)

Results: Phenotype 1

Statistic \ Kernel	IBS	AM	H1	Skat
SimReg	0.761	0.761	0.743	0.126
MDMR	0.839	0.836	0.873	0.065
SKAT	0.830	0.830	0.871	0.114

Statistic \ Kernel	Tree1	Tree2	Tree3	Tree4	Tree5
SimReg	0.684	0.684	0.531	0.538	0.538
MDMR	0.716	0.721	0.361	0.364	0.365
SKAT	0.733	0.733	0.335	0.349	0.349

Figure: Power of kernel methods applied on simulated phenotype 1 model.

Results: Phenotype 2

Statistic \ Kernel	IBS	AM	H1	Skat
SimReg	0.259	0.259	0.172	0.864
MDMR	0.375	0.371	0.306	0.806
SKAT	0.349	0.349	0.286	0.909

Statistic \ Kernel	Tree1	Tree2	Tree3	Tree4	Tree5
SimReg	0.152	0.152	0.440	0.438	0.438
MDMR	0.199	0.196	0.477	0.465	0.470
SKAT	0.188	0.188	0.242	0.246	0.246

Figure: Power of kernel methods applied on simulated phenotype 2 model.

Crohn's Disease Analysis

- Real dataset composed of 258 trios (father, mother and child) affected with Crohn's disease
- 103 SNP markers across 500 kb of the 5q31 region of chromosome 5; region divided into 100 focal points

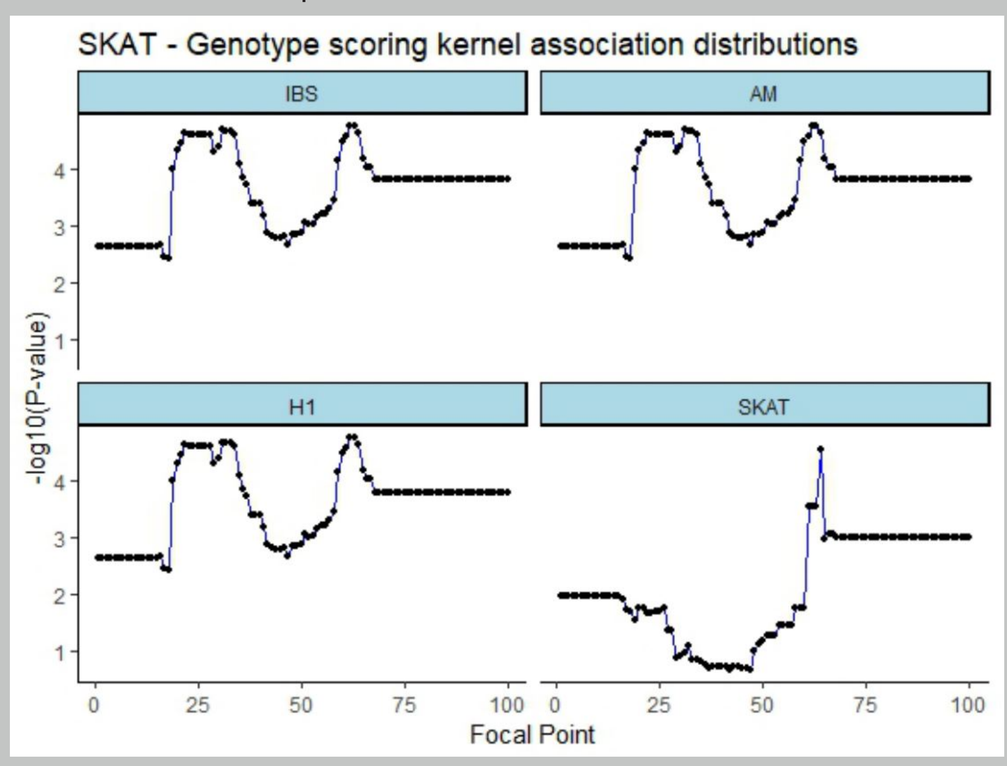

Figure: Upper Plot: P-value distribution plots of the SKAT statistic across the 100 focal points. The four different panels represent the 4 different genotype scoring approaches. Lower Plot: Distribution plots of the SimReg statistic across the 100 focal points. The different panels represent the 5 different tree kernel approaches.

Conclusion

- Under a common causal genetic variant model, power is best when using kernels that score genotype similarity.
- Under a multiple rare causal variant model, power is best when using the tree or SKAT kernels.
- For the Crohn's disease analysis, results depend on the choice of kernel and kernel-based statistic.

ptea035@uottawa.ca