UAB SCHOOL OF PUBLIC HEALTH

A COMBINED SEQUENCE KERNEL ASSOCIATION TEST (SKAT) AND ADAPTIVE RANK TRUNCATED PRODUCT (ARTP) POWERFUL METHOD FOR GENETIC PATHWAY ANALYSIS

Qi Yan

Department of Biostatistics, University of Alabama at Birmingham

Introduction

- Drawbacks of traditional GWAS:

- 1. Single genetic variants that contribute weak but real effects on disease risks are likely to be missed after taking into account multiple comparison adjustment.
- 2. Single-marker association test cannot handle genetic hierarchy appropriately due to ignoring pathways---genes---SNPs structure.

Introduction

- Advantages of grouping test of SNPs

  (e.g. genes---SNPs):

  - 1. It is able to overcome the barrier of stringent significance level by reducing the number of testing came out.
  - 2. It has the potential to improve the power when the joint effect of multiple SNPs is stronger than individual SNPs.

- Sequence Kernel Association Test (SKAT) is one of the popular grouping test methods.

Introduction

- Pathway analysis

  (pathways---genes---SNPs):

- 1. It is more biologically meaningful than single-marker association test by incorporating prior biological knowledge.
- 2. It reduces the number of hypotheses being tested and thus relaxing the stringent significance level.
- 3. It could be powerful to detect the joint effect of multiple SNPs.

Introduction

- A typical pathway analysis:
  - Firstly need to predefine sets of SNPs or genes as pathways
  - Then use statistical approaches to evaluate the significance of test statistics at pathway level.
  - Because of the difficulties of deriving exact distributions of test statistics, most of methods of pathway analysis involve permutation procedure.

A Combined SKAT-ARTP Powerful Method for Genetic Pathway Analysis

- Specific aim:

- We propose a powerful pathway analysis approach that combines SKAT and ARTP method. In other words, SKAT is applied to summarize gene-level statistic and ARTP is used to summarize pathway-level statistic.

- We propose an optimized set of weights allowing good power for both common and rare variants in SKAT.

Methods (SKAT-ARTP Pathway Analysis Method):

SKAT (Wu et al., 2011):

We assume an $n \times 1$ vector of the trait **y**. The link function $h(\cdot)$ is used to map linear combination of predictors for observation i, η_i, to the conditional mean of **y** for observation i, μ_i.

$$h(\mu) = \eta = X\beta + G\gamma$$

1. **X** is an $n \times p$ covariate matrix, $\boldsymbol{\beta}$ is a $p \times 1$ vector representing fixed effects parameters;
2. **G** is an $n \times q$ genotype matrix for q genetic variants of interest, $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of variants;

- The random effects γ_j is assumed to be normally distributed with variance $\tau \mathbf{W}_j$ for variant j, so the null hypothesis being tested is H_0: $\boldsymbol{\gamma}=0$, which is equivalent to test H_0: $\tau=0$

Specifically, the variance-component score statistics is

$$Q = (y - \hat{\mu})' G W G' (y - \hat{\mu})$$

where $\hat{\mu}$ is the predicted mean of y under null hypothesis. In a dichotomous trait case, $\hat{\mu} = \text{logit}^{-1}(X\hat{\beta})$. Here $W = diag(w_1, \ldots, w_q)$ contains the weights of the q variants. In the matrix notation,

$$Q = [y_1 - \hat{\mu} \quad \cdots \quad y_n - \hat{\mu}]_{1\times n} \begin{bmatrix} G_{11} & \cdots & G_{1p} \\ \vdots & \ddots & \vdots \\ G_{n1} & \cdots & G_{np} \end{bmatrix}_{n\times p} \begin{bmatrix} w_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & w_p \end{bmatrix}_{p\times p} \begin{bmatrix} G_{11} & \cdots & G_{n1} \\ \vdots & \ddots & \vdots \\ G_{1p} & \cdots & G_{np} \end{bmatrix}_{p\times n} \begin{bmatrix} y_1 - \hat{\mu} \\ \vdots \\ y_n - \hat{\mu} \end{bmatrix}_{n\times 1}$$

Under the null hypothesis, Q follows a mixture of chi-square distributions, which can be closely approximated with the computationally efficient Davies method.

A good choice of weights could improve the power. SKAT adapts Common Disease-Rare Variants hypothesis that assumes rare variants are more likely to be causal variants than common variants. A Beta density with parameters $a_1 = 1$ and $a_2 = 25$ is recommended as weight function, $\sqrt{w_j} = Beta(MAF_j;\ a_1,\ a_2)$

We proposed the sum of beta density (e.g. $0.5 * Beta(MAF_j;\ 1,\ 25)$) and inverse of single marker test p-value (e.g. $1.25/(0.1 + pvalue)$) as a better square root of weight for testing both common and rare variants.

The Adaptive Rank Truncated Product (ARTP) Algorithm (Yu et al., 2009):

- ARTP method is a truncated product method that uses the product of some auto-selected most significant p-values.
- For combining gene level p-values:
- First, we obtain p-values for each gene on the null hypothesis based on the observed data, denoted as $p_1^{(0)}, \ldots, p_L^{(0)}$, where L is the number of genes in the pathway.
- Second, we permute the phenotypes to generate B datasets under the null hypothesis. Based on the b^{th} permuted dataset, $0<b\leq B$, we can also obtain the p-values, $p_1^{(b)}, \ldots, p_L^{(b)}$, for genes.

In sum, the ARTP algorithm is shown as below:

1. Based on $p_1^{(b)}, \ldots, p_L^{(b)}$ from gene level test approach, 0≤b≤B (b=0 is for the observed data set), for any given b, calculate the rank truncated product statistics for each candidate truncation point, denoted as $W_j^{(b)} = \prod_{i=1}^{j} p_{(i)}^{(b)}, 1 \le j \le L$, where $p_{(i)}^{(b)}$ is the ranked p-value in the bth permuted dataset ($p_{(1)}^{(b)}$ is the smallest p-value). Therefore, for the bth permuted dataset, we have $W_1^{(b)}, W_2^{(b)}, \ldots, W_L^{(b)}$.
2. Based on $W_j^{(b)}$, $1 \le j \le L$, $0 \le b \le B$, for any given b, apply $\hat{S}_j^{(b)} = \frac{\sum_{b^*=0}^{B} I\left(W_j^{(b^*)} \le W_j^{(b)}\right)}{B+1}$ to obtain the corresponding p-values for $W_j^{(b)}$.
3. For any b, let $MinP^{(b)} = min_{1 \le j \le L} \hat{S}_j^{(b)}, 0 \le b \le B$. The adjusted p-value for the adaptive rank truncated product statistic $MinP^{(0)}$ is estimated as $\frac{\sum_{b=0}^{B} I\left(MinP^{(b)} \le MinP^{(0)}\right)}{B+1}$ and it is the pathway p-value.

- SKAT-ARTP Method:

  Powerful and efficient SKAT algorithm is used first to obtain gene-level p-values for all genes within the pathway. ARTP algorithm is then applied to evaluate the association between the pathway and disease while excluding genes that do not affect phenotypes.

- Other Pathway Analysis Approaches:

  - *ARTP-ARTP* (*Yu et al., 2009*): both gene-level and pathway-level p-values are evaluated by ARTP;
  - *Individual SKAT*: all SNPs in one pathway as a group, ignore gene-level information;

Simulation study:

- Null gene sets

- Type I Diabetes genotype dataset from WTCCC.
- 2000 samples.
- SNPs assigned to a gene if they are located within the flank of 5kb.
- Overlapped genes were deleted.
- Genes assigned to pathways in KEGG based on HUGO Gene symbols.
- At the end, 6 pathways including 99 genes and 797 SNPs were selected as genotypes
- Simulated 1000 sets of phenotypes and hence 1000 simulated datasets were generated.

The dichotomous phenotypes were generated via the model:

$$logitP(y = 1) = \alpha_0$$

where α_0 was determined to set the prevalence to 5%.

- Causal gene sets

The dichotomous phenotypes were generated via the model:

$$logitP(y = 1) = \alpha_0 + 0.5X_1 + 0.5X_2 + \beta_1 G_1 + \beta_2 G_2 + \cdots + \beta_p G_p$$

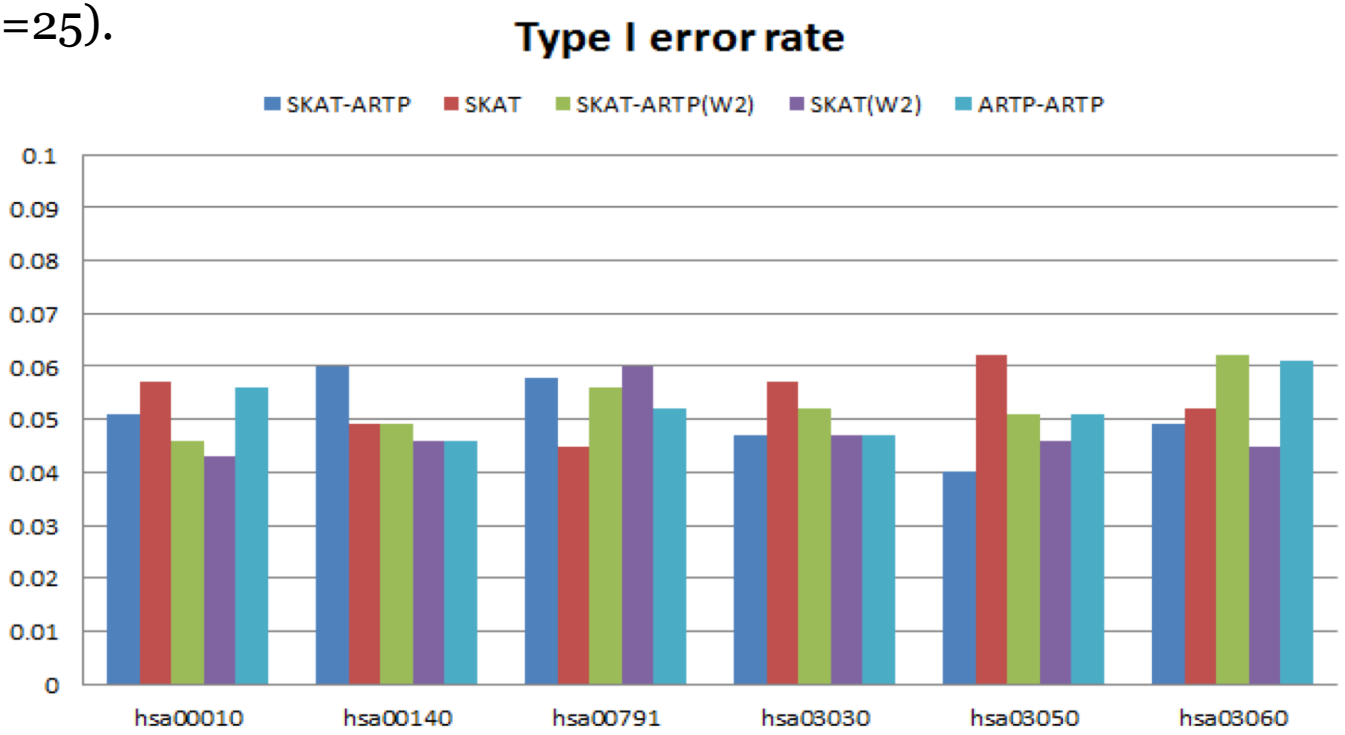
where X_1 is a continuous covariate generated from a standard normal distribution, X_2 is a dichotomous covariate from a Bernoulli distribution with probability of 0.5, $G_1, G_2, \ldots, G_p$ are the genotypes for causal SNPs and $\beta_1, \beta_2, \ldots, \beta_p$ are log ORs for the causal SNPs. α_0 was determined as described in null gene set. $\beta_1, \beta_2, \ldots, \beta_p$ were set to $c|log_{10}MAF_j|$ in order to assign large effects to rare variants, and c=0.8.

First scenario is there are totally 6 causal variants and each one is from one pathway;

Second scenario is the number of causal SNPs is proportional to the length of the pathway.

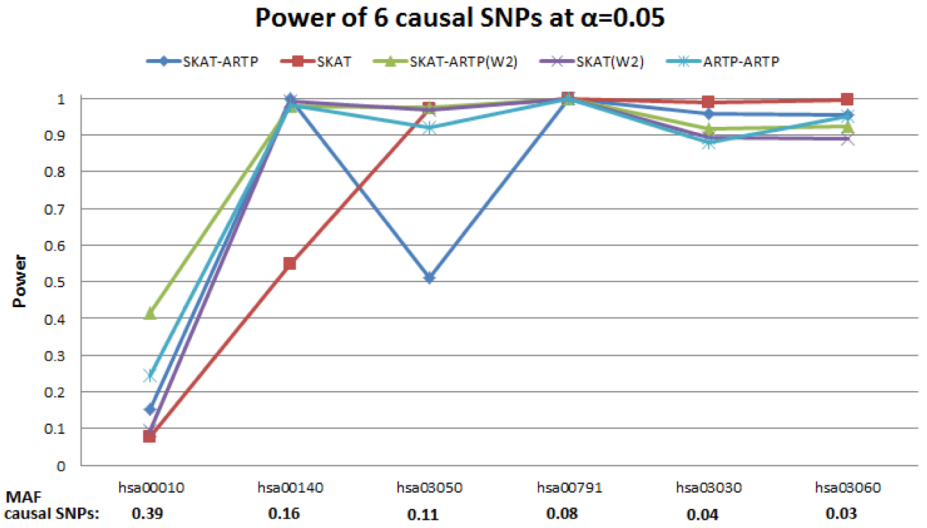
Simulation Study Results:

(W2) indicates that the weights for each SNP used in SKAT are [1.25/(0.1+p-value) + Beta(MAF; 1, 25)]^2, where p-value is from single marker test;
No specification indicates that SKAT uses default weights (Beta function with a_1=1 and a_2=25).

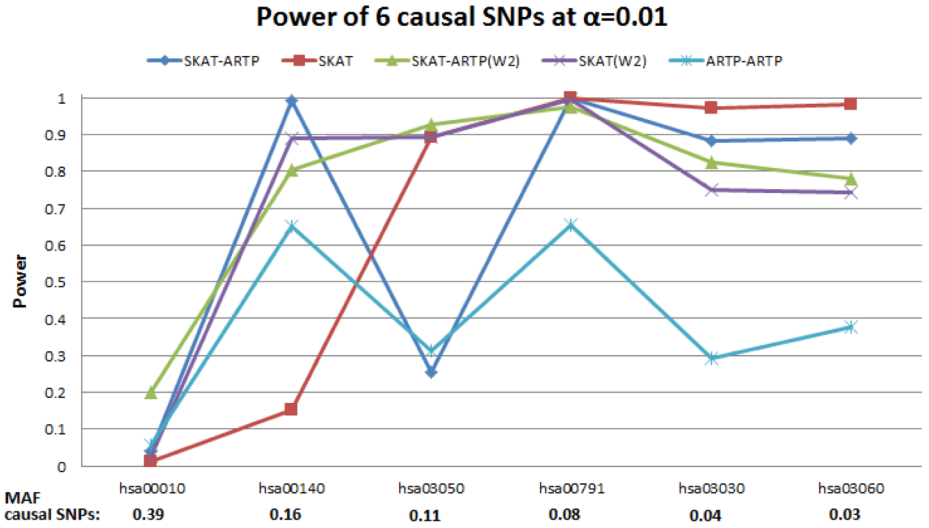


Type I error rate under the significant level of 0.05.

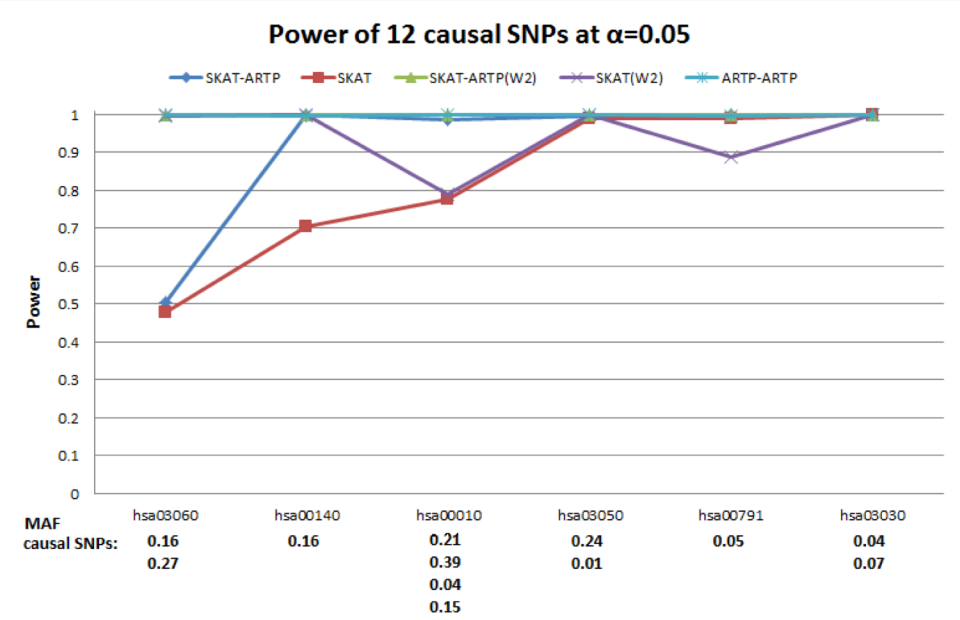
For the average type I error rate over all six pathways, SKAT-ARTP is 0.0508, individual SKAT is 0.0537, SKAT-ARTP(W2) is 0.0527, SKAT(W2) is 0.0478 and ARTP-ARTP is 0.0522.



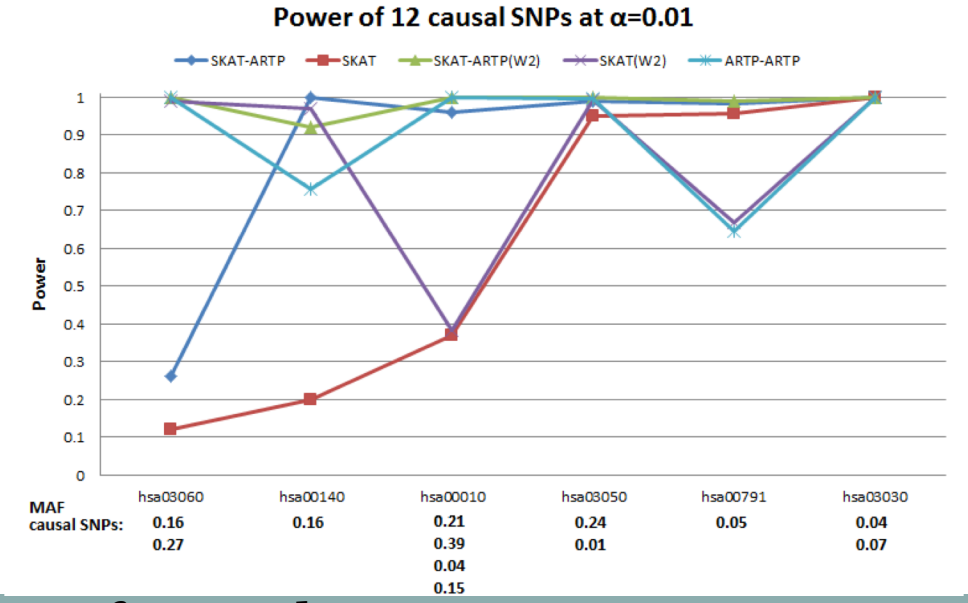
Power of 6 causal SNPs at α=0.05, $\beta_i = 0.8|\log_{10}MAF_j|$



Power of 6 causal SNPs at α=0.01, $\beta_i = 0.8|\log_{10}MAF_j|$



Power of 6 causal SNPs at α=0.05, $\beta_i = 0.8|\log_{10}MAF_j|$



Power of 6 causal SNPs at α=0.01, $\beta_i = 0.8|\log_{10}MAF_j|$

Real Data Results:

Application to Wellcome Trust Case Control Consortium Bipolar Disorder dataset

1998 cases and 1504 controls
10000 permutations

Pathway name	Total genes in the pathway	Raw p-value	Adjusted p-value
Cation channel activity	91	9.99e-5	3.00e-4
Gated channel activity	87	9.99e-5	3.00e-4
Metal ion transmembrane transporter activity	110	9.99e-5	3.00e-4

Questions

Contact information

E-mail: kid1412@uab.edu
Cell phone: 205-396-6942