Rare-Variant Kernel Machine Test for Longitudinal Data for Population and Family Samples

Qi Yan

Department of Pediatrics, Children's Hospital of Pittsburgh of UPMC

Motivation

- Phenotypes:
  - In many genetic studies, phenotypes are measured at multiple time points for each subject. It is expected that a method that is able to take into account all time points jointly in an association test could improve the power;
  - Family based designs have been widely used. Appropriately handling familial correlation can retain Type I error rate;

- Genotypes:
  - Common variants (MAF≥0.05): single marker test;
  - Rare variants (MAF<0.05): test at gene level (e.g. SKAT).

Aims

- Association test between quantitative phenotypes and genes;
- Rare variants are assigned into genes;
- multiple time points for each subject are tested simultaneously.
- Family structure is either (1). not considered or (2). considered;

Methods

- **Kernel Machine (KM) Regression for Linear Mixed Model:**

$$y = X\beta + G\gamma + u + \varepsilon$$

1. ***y***: quantitative phenotypes (multiple correlated phenotypes);
2. ***Xβ***: fixed effects of covariates;
3. ***Gγ***: genetic effects from one gene consisted of SNPs;
4. ***u***: random effects of covariates;
5. ***ε***: random error.

- Assume $\gamma \sim N(0, \tau W)$, H_0: $\gamma=0 \rightarrow$ **H_0: $\tau=0$**;
- $u \sim N(0, \ K)$ and $\varepsilon \sim N(0, \ \sigma_E^2 I)$

$$Q = (y - X\hat{\beta})' \hat{\Sigma}^{-1} G W G' \hat{\Sigma}^{-1} (y - X\hat{\beta})$$

➢ Longitudinal Kernel Machine (L-KM) regression for Quantitative Traits for Population Data:

Under the null hypothesis, the random intercept and time model for the i-th subject at time point j is

$$y_{ij} = \beta_0 + t_{ij}\beta_1 + b_{0i} + t_{ij}b_{1i} + \varepsilon_{ij}$$

For one subject,

$$y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

$$Var(b_i) = \begin{pmatrix} \sigma^2_{int} & \sigma_{cov} \\ \sigma_{cov} & \sigma^2_{time} \end{pmatrix}$$

$$\boldsymbol{Var(y_i) = Z_i Var(b_i) Z_i{}' + \sigma^2_E I_{m\times m}}$$

For the whole data set, the variance term is

$$\boldsymbol{Var(y) = I \otimes Z_i Var(b_i) Z_i{}' + \sigma^2_E I = \Sigma}$$

➢ Longitudinal Family Kernel Machine (LF-KM) regression for Quantitative Traits for Family Data:

Under the null hypothesis, the random intercept and time model for the i-th subject in the k-th family at time point j is

$$y_{ijk} = \beta_0 + t_{ijk}\beta_1 + b_{0ik} + t_{ijk}b_{1ik} + \delta_{ik} + \varepsilon_{ijk}$$

For one subject,

$$y_{ik} = X_{ik}\beta + Z_{ik}b_{ik} + \delta_{ik} + \varepsilon_{ik}$$

For one trio family,

$$y_k = X_k\beta + Z_k b_k + \delta_k + \varepsilon_k$$

$$Var(Z_k b_k) = I_{3\times 3} \otimes Z_{ik} Var(b_{ik}) Z_{ik}{}' = I_{3\times 3} \otimes Z_{ik} \begin{pmatrix} \sigma_{int}^2 & \sigma_{cov} \\ \sigma_{cov} & \sigma_{time}^2 \end{pmatrix} Z_{ik}{}'$$

$$Var(\delta_k) = \sigma_G^2 \cdot J_k \Phi_k J_k{}' = \sigma_G^2 \cdot \begin{bmatrix} 1_{m\times 1} & 0_{m\times 1} & 0_{m\times 1} \\ 0_{m\times 1} & 1_{m\times 1} & 0_{m\times 1} \\ 0_{m\times 1} & 0_{m\times 1} & 1_{m\times 1} \end{bmatrix} \Phi_k \begin{bmatrix} 1_{m\times 1} & 0_{m\times 1} & 0_{m\times 1} \\ 0_{m\times 1} & 1_{m\times 1} & 0_{m\times 1} \\ 0_{m\times 1} & 0_{m\times 1} & 1_{m\times 1} \end{bmatrix}'$$

$$\Phi_k = \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix}$$

$$\boldsymbol{Var(y_k) = Var(Z_k b_k) + Var(\delta_k) + \sigma_E^2 I_{3m\times 3m}}$$

For the whole data set, we assume n individuals from families. The variance term is

$$\boldsymbol{Var(y) = I \otimes Z_{ik} Var(b_{ik}) Z_{ik}{}' + \sigma_G^2 \cdot J\Phi J' + \sigma_E^2 I = \Sigma}$$

$$J = \begin{bmatrix} 1_{m\times 1} & \cdots & 0_{m\times 1} \\ \vdots & \ddots & \vdots \\ 0_{m\times 1} & \cdots & 1_{m\times 1} \end{bmatrix}_{nm\times n}$$

➢ Simulation Studies

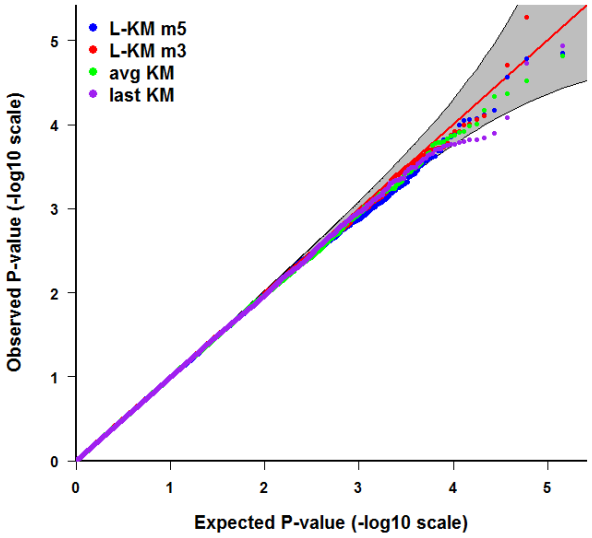
- Genotypes:
  - Population dataset = 1,000 ×30 rare variants;
  - Trio family dataset = 300 trios ×30 rare variants;
  - Three generation family dataset = 100 families ×30 rare variants;
  - Total = 100 genotype datasets.

- Phenotypes:
  - Type I error rate: 1000 sets of phenotypes for each genotype dataset (independent);
  - Power: 1000 sets of phenotypes for each genotype dataset (Causal variants(+/-) = 30%/0%; 20%/10%).

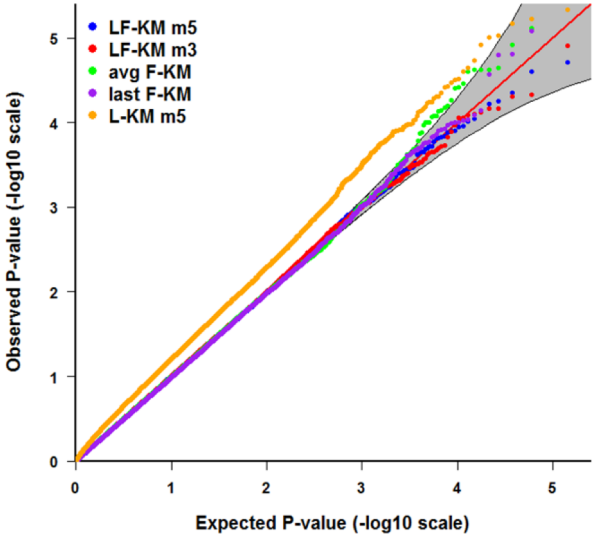
Results

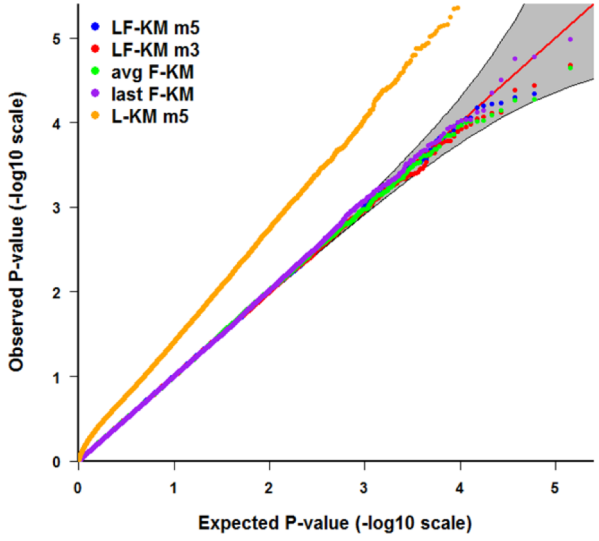
➢ Simulation of the Type I Error Rate:

Population

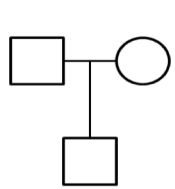


Family



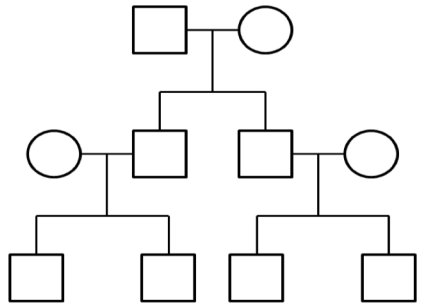


(A) (B)



Trio

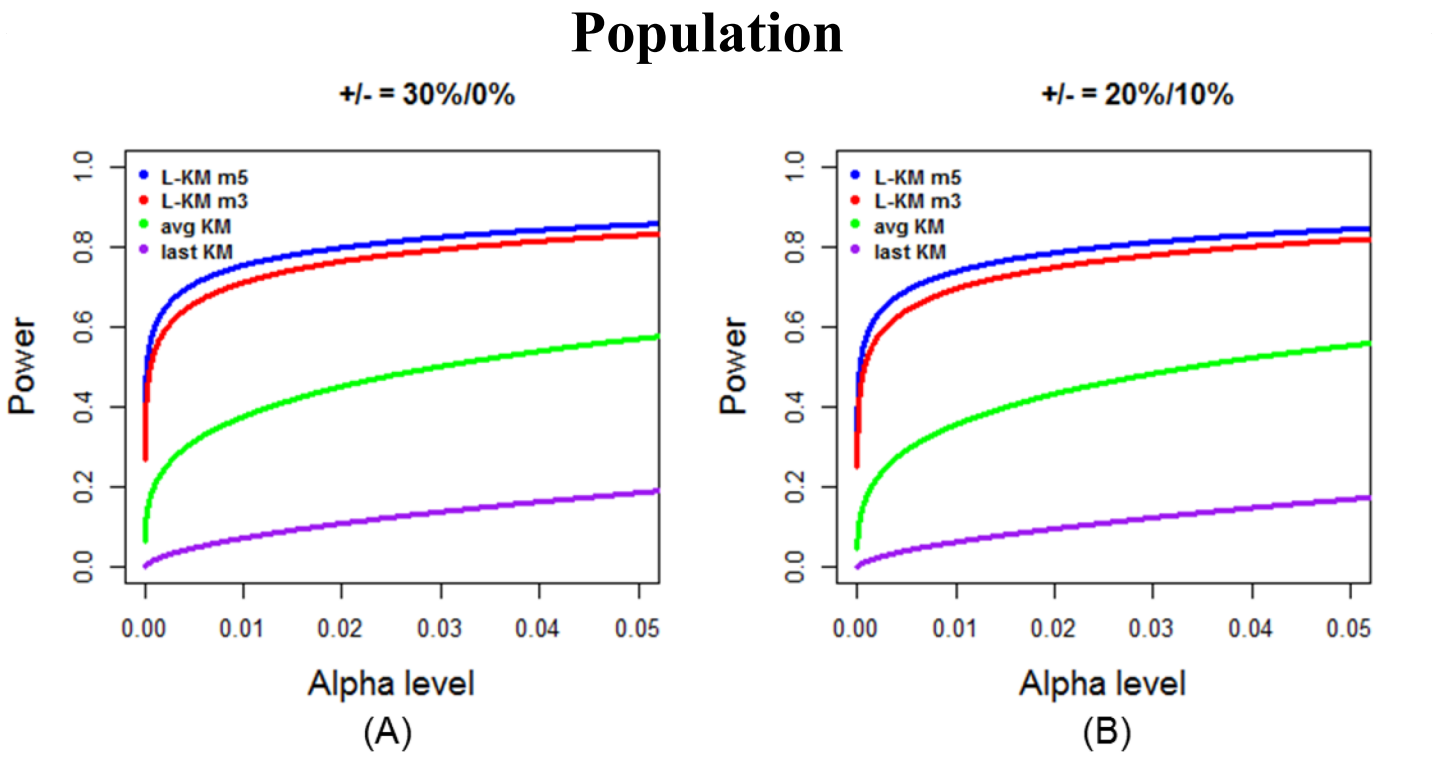
(A)



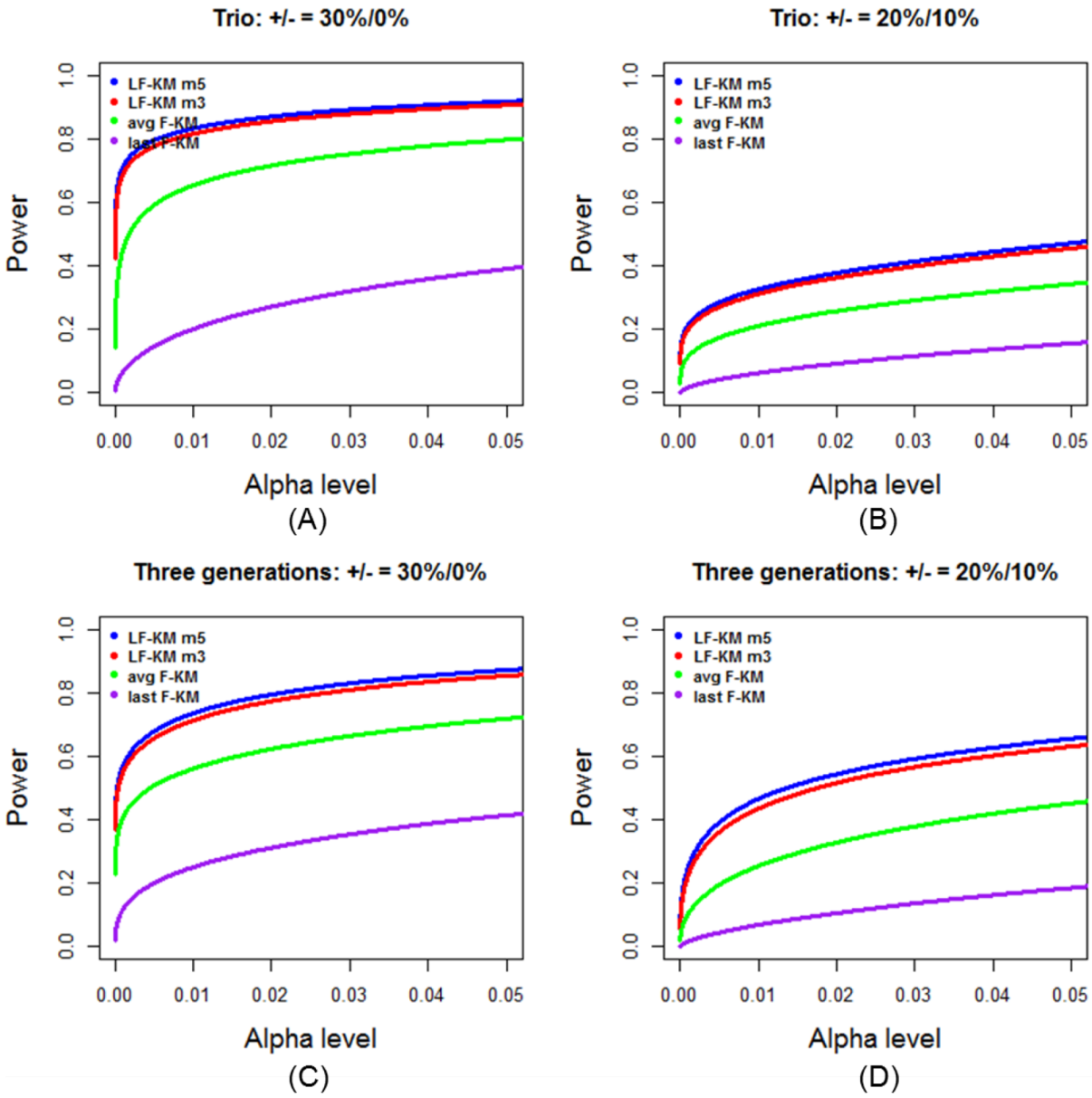
Three generations

(B)

➢ Statistical Power Comparison:

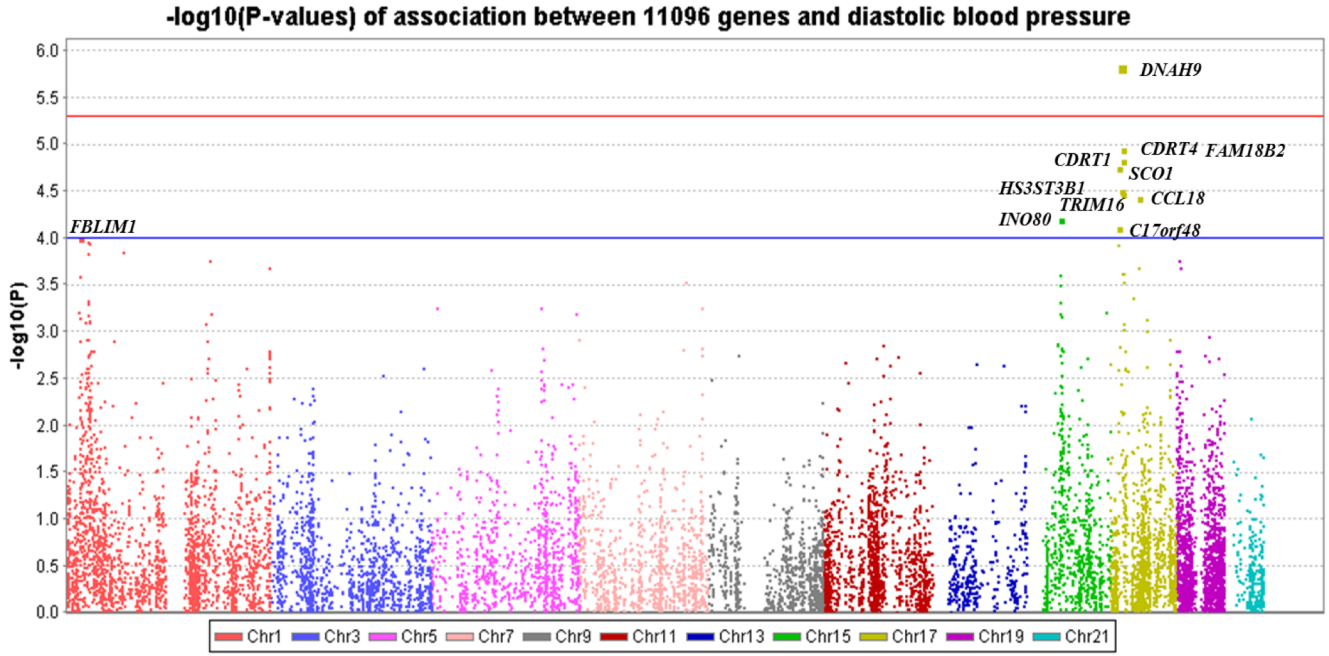


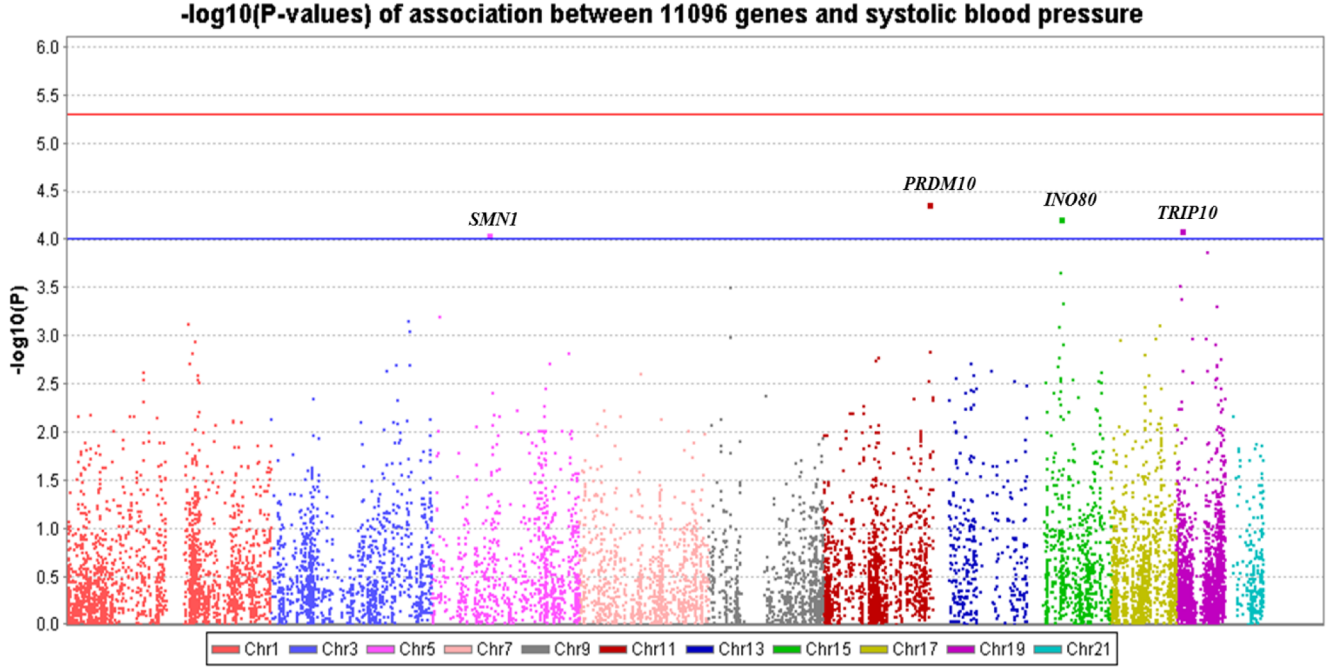
Family


Trio: +/- = 30%/0%
LF-KM m5
LF-KM m3
avg F-KM
last F-KM
Power
Alpha level
(A)
Trio: +/- = 20%/10%
(B)
Three generations: +/- = 30%/0%
(C)
Three generations: +/- = 20%/10%
(D)

➢ GAW18 Data Analysis Results:

- 855 subjects from 20 families were used in the analysis and each subject has up to 4 exam points;
- Assigned rare variants to a gene if they are located within a 5kb flank;
- 11,096 genes were used in the analysis;
- Used the LF-KM statistic to analyze the association of genetic variants with diastolic and systolic blood pressure that are considered heritable traits.


-log10(P-values) of association between 11096 genes and diastolic blood pressure
-log10(P)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
DNAH9
CDRT1
CDRT4
FAM18B2
SCO1
HS3ST3B1
CCL18
TRIM16
INO80
C17orf48
FBLIM1
Chr1
Chr3
Chr5
Chr7
Chr9
Chr11
Chr13
Chr15
Chr17
Chr19
Chr21


-log10(P-values) of association between 11096 genes and systolic blood pressure
-log10(P)
6.0
5.5
5.0
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0.0
PRDM10
INO80
TRIP10
SMN1
Chr1
Chr3
Chr5
Chr7
Chr9
Chr11
Chr13
Chr15
Chr17
Chr19
Chr21

Summary

- Implement L-KM for testing the association of rare variants in population samples, which simultaneously considers multiple measurements as well as LF-KM for testing the association of rare variants in family samples.

- L-KM retains the correct Type I error rate, and achieves the best power performance in population samples; LF-KM retains the correct Type I error rate, and achieves the best power performance in family samples.

- Observe potential important genes associated with blood pressure.

- The software is available (http://www.pitt.edu/~qiy17/Softwares.html).