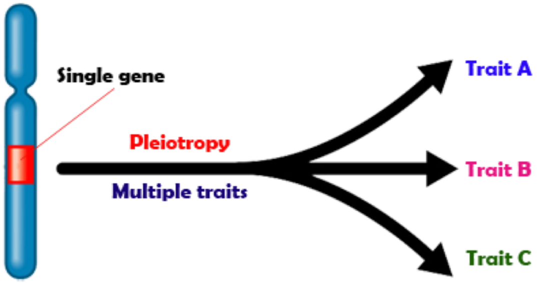
Associating Multivariate Quantitative Phenotypes with Genetic Variants in Family Samples with a Novel Kernel Machine Regression Method

Qi Yan

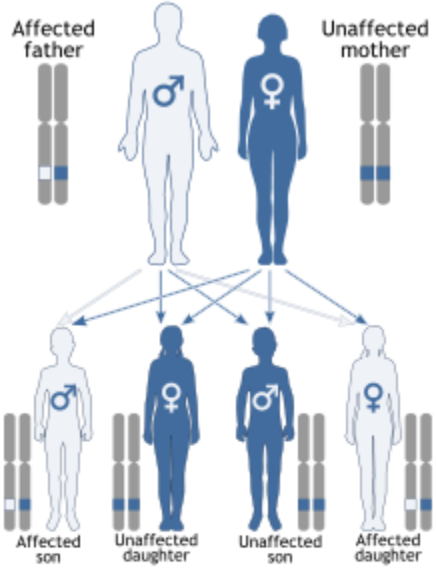
Department of Pediatrics, University of Pittsburgh
Children's Hospital of Pittsburgh of UPMC

June 13th, 2016

Motivation



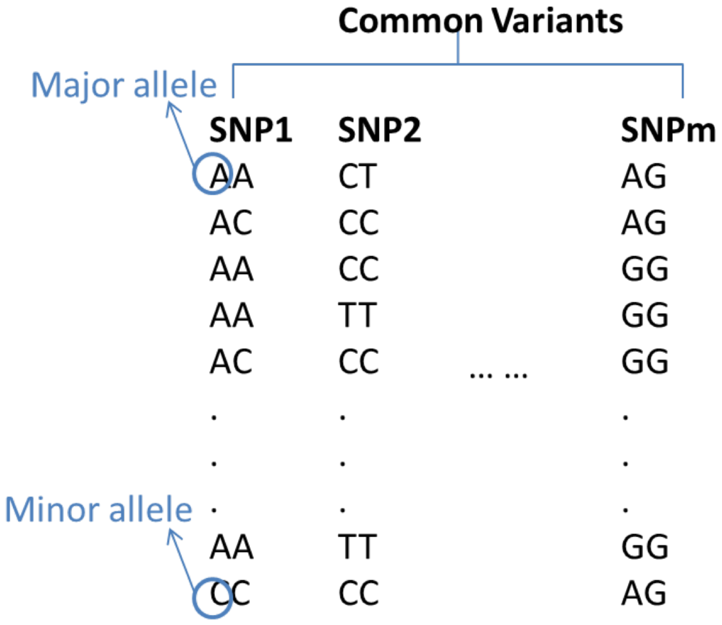
- Phenotypes:
  - Genetic studies have been conducted to collect multiple correlated phenotypes for one complex disease. Jointly modeling multiple phenotypes can improve the statistical power [Sivakumaran S, et al. AJHG. 2011];
  - Family based designs have been widely used [Spielman RS, et al. AJHG. 1993]. Appropriately handling familial correlation can retain Type I error rate;



http://dragonflyissuesinevolution13.wikia.com/wiki/Pleiotropy
http://en.wikipedia.org/wiki/Hereditary_hemorrhagic_telangiectasia

Motivation

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

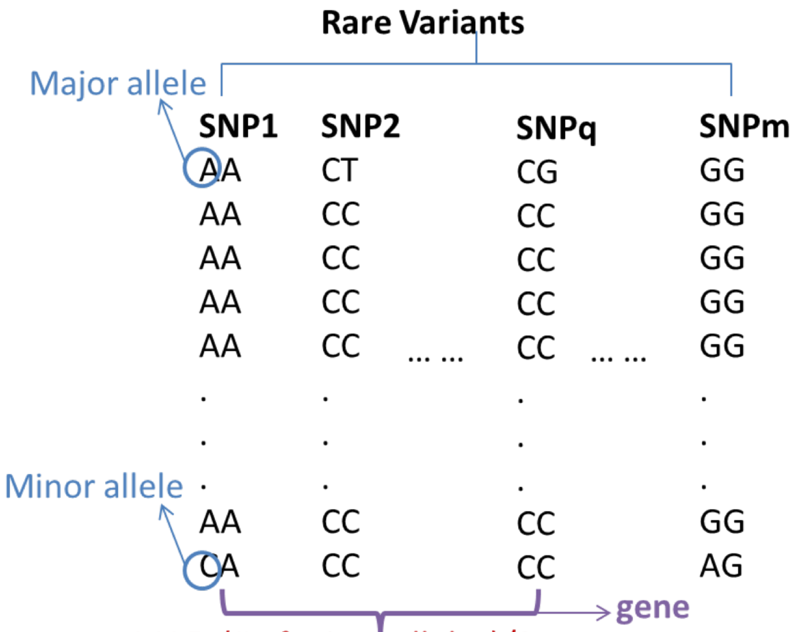


Common Variants

SNP1	SNP2		SNPm
AA	CT		AG
AC	CC		AG
AA	CC		GG
AA	TT		GG
AC	CC	… …	GG
.	.		.
.	.		.
.	.		.
AA	TT		GG
CC	CC		AG

MAF=(# of minor alleles)/2n

MAF>0.05 (common variant)



Rare Variants

SNP1	SNP2		SNPq		SNPm
AA	CT		CG		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC		CC		GG
AA	CC	… …	CC	… …	GG
.	.		.		.
.	.		.		.
.	.		.		.
AA	CC		CC		GG
CA	CC		CC		AG

MAF=(# of minor alleles)/2n

MAF<0.05 (rare variant)

Aims

- Association test between multiple quantitative phenotypes and genes in family samples
  - Rare variants are assigned into genes;
  - Family structure is considered;
  - Correlated quantitative phenotypes are tested simultaneously.

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Let there be n subjects with q genetic variants. The $n \times 1$ vector of the quantitative trait $\boldsymbol{y}$ follows a linear mixed model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\varepsilon}$$

- $\mathbf{X}$ is an $n \times p$ covariate matrix,
- $\boldsymbol{\beta}$ is a $p \times 1$ vector containing parameters for the fixed effects (an intercept and $p - 1$ covariates),
- $\mathbf{G}$ is an $n \times q$ genotype matrix for the q genetic variants of interest,
- $\boldsymbol{\gamma}$ is a $q \times 1$ vector for the random effects of the q genetic variants,
- $\boldsymbol{\varepsilon}$ is an $n \times 1$ vector for the random error,
- $\boldsymbol{u}$ is an $n \times 1$ vector for the random effects due to covariates (e.g., correlation between phenotypes or relatedness in families)

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W}) \qquad H_0\!: \tau = 0$$

$$\mathbf{u} \sim N(0, \mathbf{K})$$

$$\boldsymbol{\varepsilon} \sim N\left(0, \sigma_{\mathrm{E}}^2\mathbf{I}\right)$$

where $\mathbf{W}$ is a predefined $q \times q$ diagonal weight matrix for each variant, and $\mathbf{K}$ is an $n \times n$ covariance matrix

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma} + \mathbf{u} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{\gamma} \sim N(0, \tau\mathbf{W})$$

For the linear mixed model, the log likelihood is

$$l = C - \frac{1}{2}\log|\boldsymbol{\Sigma}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

$$\boldsymbol{\Sigma} = \tau\mathbf{G}\mathbf{W}\mathbf{G}' + \mathbf{K} + \sigma_{\mathrm{E}}^{2}\mathbf{I}$$

To derive the score test for H_0: $\tau = 0$, we take the first derivative with respect to τ

Score function: $$\frac{dl}{d\tau} = -\frac{1}{2}tr(\underset{\substack{\uparrow \\ \text{Fixed}}}{\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'}) + \frac{1}{2}\underset{\substack{\uparrow \\ \text{Test statistic}}}{(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})'\boldsymbol{\Sigma}^{-1}\mathbf{G}\mathbf{W}\mathbf{G}'\boldsymbol{\Sigma}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})}$$

Methods

➢ Kernel Machine (KM) Regression for Linear Mixed Model:

Under the null hypothesis, the linear mixed model is $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u} + \boldsymbol{\varepsilon}$, and the estimates are

$$\widehat{\boldsymbol{\Sigma}} = \widehat{\mathbf{K}} + \hat{\sigma}_E^2\mathbf{I}$$

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{y}$$

Replacing the variance components with their maximum likelihood estimators (MLEs), we have

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{GWG}'\widehat{\boldsymbol{\Sigma}}^{-1}\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

as the test statistic. Under the null hypothesis, the variance of the residual is: $Var\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right) = \mathbf{P_0}$

The statistic Q is a quadratic form of $\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}$ and follows a mixture of chi-square distributions under H_0. Thus,

$$\mathrm{Q} \sim \sum_{i=1}^{q} \lambda_i \chi_{1,i}^2$$

where λ_i are the eigenvalues of the matrix $\mathbf{W}^{\frac{1}{2}}\mathbf{G}'\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{P_0}\widehat{\boldsymbol{\Sigma}}^{-1}\mathbf{GW}^{\frac{1}{2}}$

Methods

➢ Kernel Machine Regression for Quantitative phenotypes in Family Data (MF-KM):

Under the null hypothesis,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{h} + \boldsymbol{\varepsilon}$$

- **y** is a vector of quantitative trait (i.e., **y** = (*y11, y12, y21, y22, …, ym1, ym2*) where m is the number of individuals),
- **Xβ** is the fixed effects of covariates,
- **h** is the random effect of correlated phenotypes corresponding to the polygenic contribution,
- **ε** is the random effect of correlated phenotypes corresponding to the random environmental contribution.

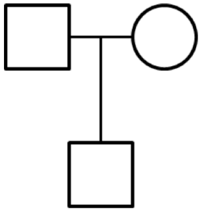
$$\mathrm{Var}(\mathbf{y}) = \underset{\substack{\uparrow \\ \text{kinship}}}{\boldsymbol{\Phi}} \otimes \underset{\substack{\uparrow \\ \text{polygenic variances}}}{\overset{\mathrm{Var}(\mathbf{h})}{\begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix}}} + \mathbf{I} \otimes \underset{\substack{\uparrow \\ \text{environmental variances}}}{\overset{\mathrm{Var}(\boldsymbol{\varepsilon})}{\begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix}}} = \boldsymbol{\Sigma}$$

Methods

➢ Simulation Studies

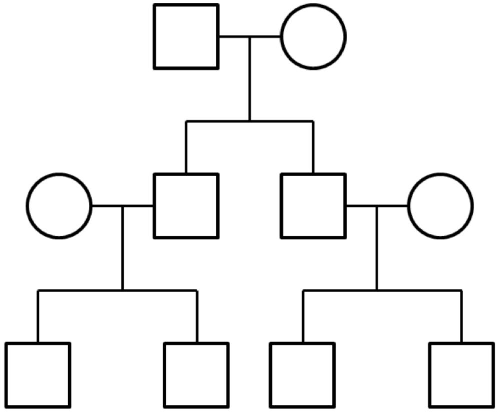
Genotypes:

- Trios:
  - One genotype dataset = 300 trios × 30 rare variants;
  - Total = 100 genotype datasets (1000 sets of phenotypes for each set of genotypes).
- Three-generation families:
  - One genotype dataset = 100 families × 30 rare variants;
  - Total = 100 genotype datasets (1000 sets of phenotypes for each set of genotypes).



Trio

(A)



Three generations

(B)

Methods

➢ Simulation Studies

Phenotypes:

- Type I error rate: 1000 sets of phenotypes for each genotype dataset (independent);

$$\mathbf{y}_i = 0.05 \cdot \mathbf{X}_{1i} + 0.5 \cdot \mathbf{X}_{2i} + \mathbf{e}_i$$

$$\begin{aligned}\mathrm{Var}(\mathbf{y}_i) &= \mathbf{\Phi}_i \otimes \begin{pmatrix} \sigma_{G1}^2 & \sigma_{G12} \\ \sigma_{G12} & \sigma_{G2}^2 \end{pmatrix} + \mathbf{I}_{3\times3} \otimes \begin{pmatrix} \sigma_{E1}^2 & \sigma_{E12} \\ \sigma_{E12} & \sigma_{E2}^2 \end{pmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0.5 \\ 0 & 1 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \otimes \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix} + \mathbf{I}_{3\times3} \otimes \begin{pmatrix} 1 & 0.8 \\ 0.8 & 1 \end{pmatrix}\end{aligned}$$

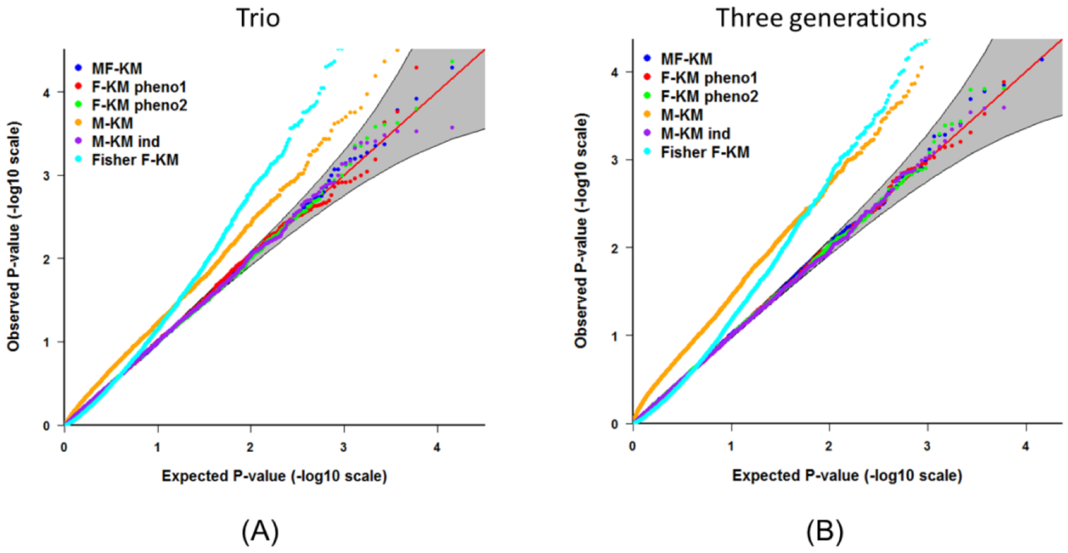
- Power: 1000 sets of phenotypes for each genotype dataset (Causal variants(+/-) = 30%/0%; 20%/10%; 20%/0%; 13%/7%).

$$\mathbf{y}_i = 0.05\mathbf{X}_{1i} + 0.5\mathbf{X}_{2i} + \boldsymbol{\beta}_1\mathbf{G}_1 + \boldsymbol{\beta}_2\mathbf{G}_2 + \cdots + \boldsymbol{\beta}_k\mathbf{G}_k + \mathbf{e}_i$$

Results

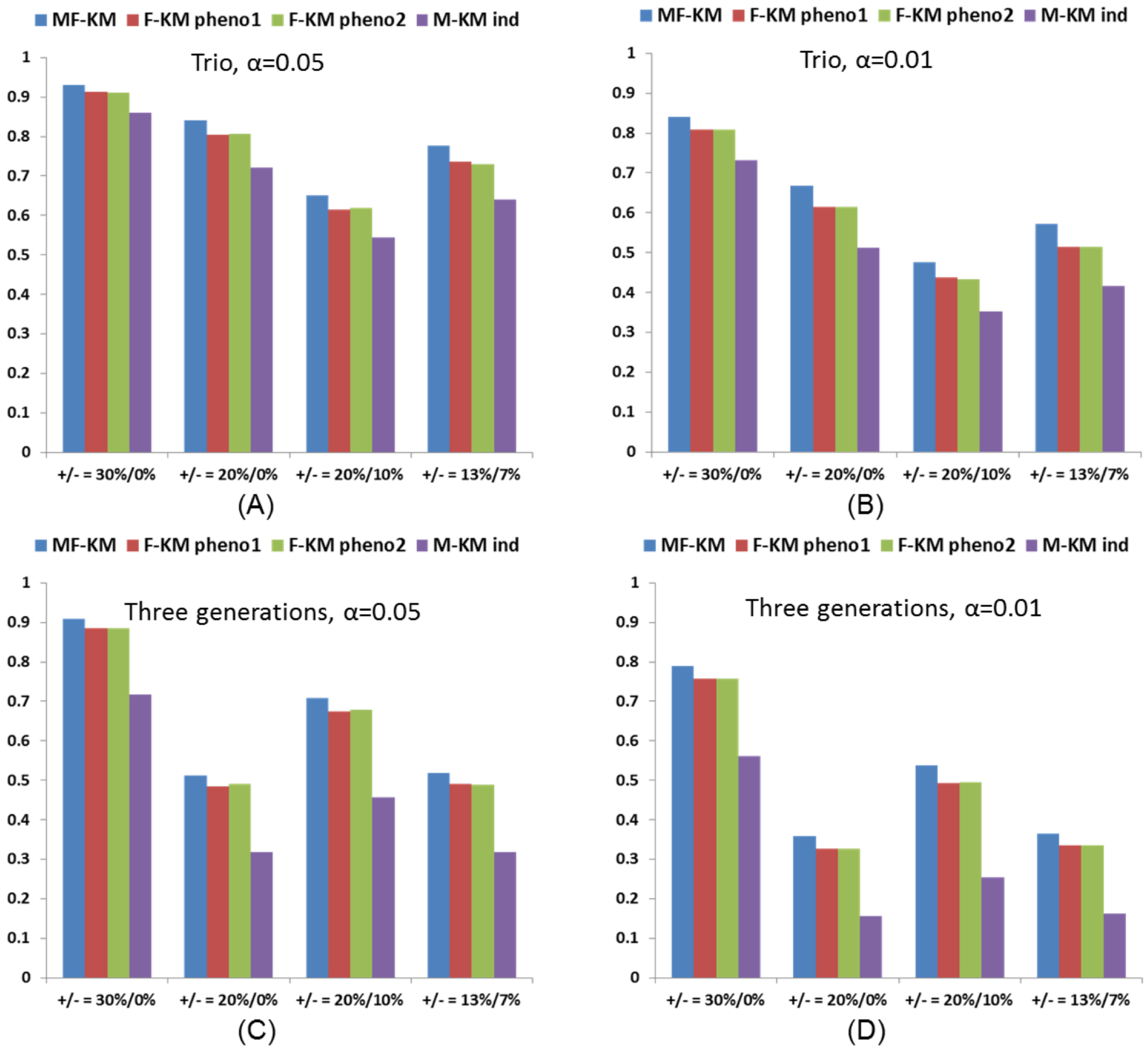
➢ Simulation of the Type I Error Rate:

		α=0.05	α=0.01	α=0.005	α=0.001
Trios	**MF-KM**	0.0497	0.0108	0.0051	0.0014
	F-KM pheno1	0.0511	0.0113	0.0057	0.0007
	F-KM pheno2	0.0473	0.0103	0.0051	0.0012
	M-KM	**0.0861**	**0.0211**	**0.0125**	**0.0031**
	M-KM ind	0.0497	0.0108	0.0047	0.0011
	Fisher F-KM	**0.0796**	**0.0285**	**0.0192**	**0.0072**
Three generations	**MF-KM**	0.0503	0.0105	0.0049	0.0010
	F-KM pheno1	0.0519	0.0104	0.0049	0.0010
	F-KM pheno2	0.0496	0.0104	0.0051	0.0010
	M-KM	**0.1270**	**0.0384**	**0.0222**	**0.0062**
	M-KM ind	0.0495	0.0094	0.0051	0.0011
	Fisher F-KM	**0.0830**	**0.0292**	**0.0200**	**0.0078**



Results

➢ Statistical Power Comparison:



(A)

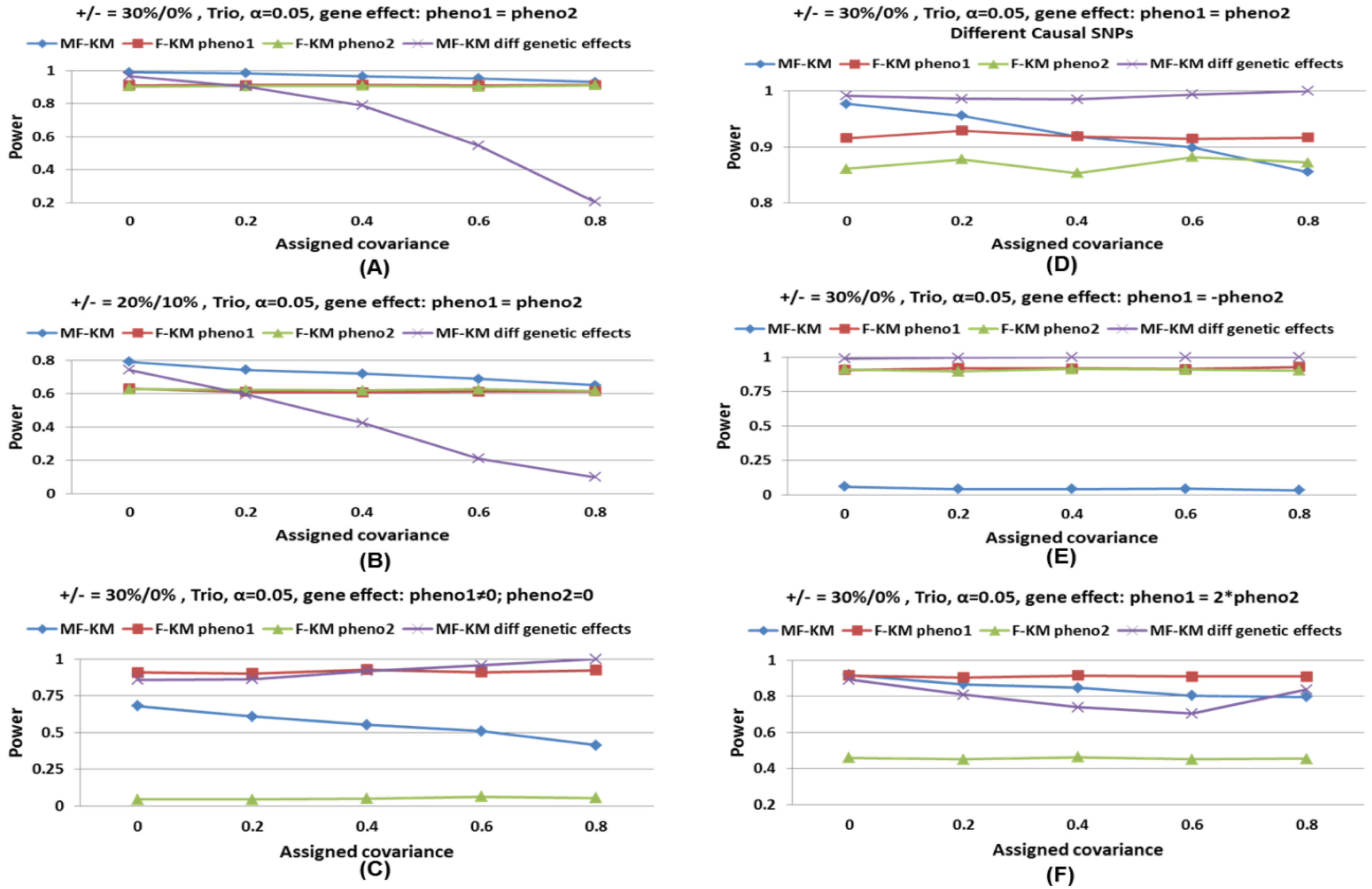
(B)

(C)

(D)

Results

- ➢ Statistical Power Comparison:



+/- = 30%/0% , Trio, α=0.05, gene effect: pheno1 = pheno2

(A)

+/- = 20%/10% , Trio, α=0.05, gene effect: pheno1 = pheno2

(B)

+/- = 30%/0% , Trio, α=0.05, gene effect: pheno1≠0; pheno2=0

(C)

+/- = 30%/0% , Trio, α=0.05, gene effect: pheno1 = pheno2
Different Causal SNPs

(D)

+/- = 30%/0% , Trio, α=0.05, gene effect: pheno1 = -pheno2

(E)

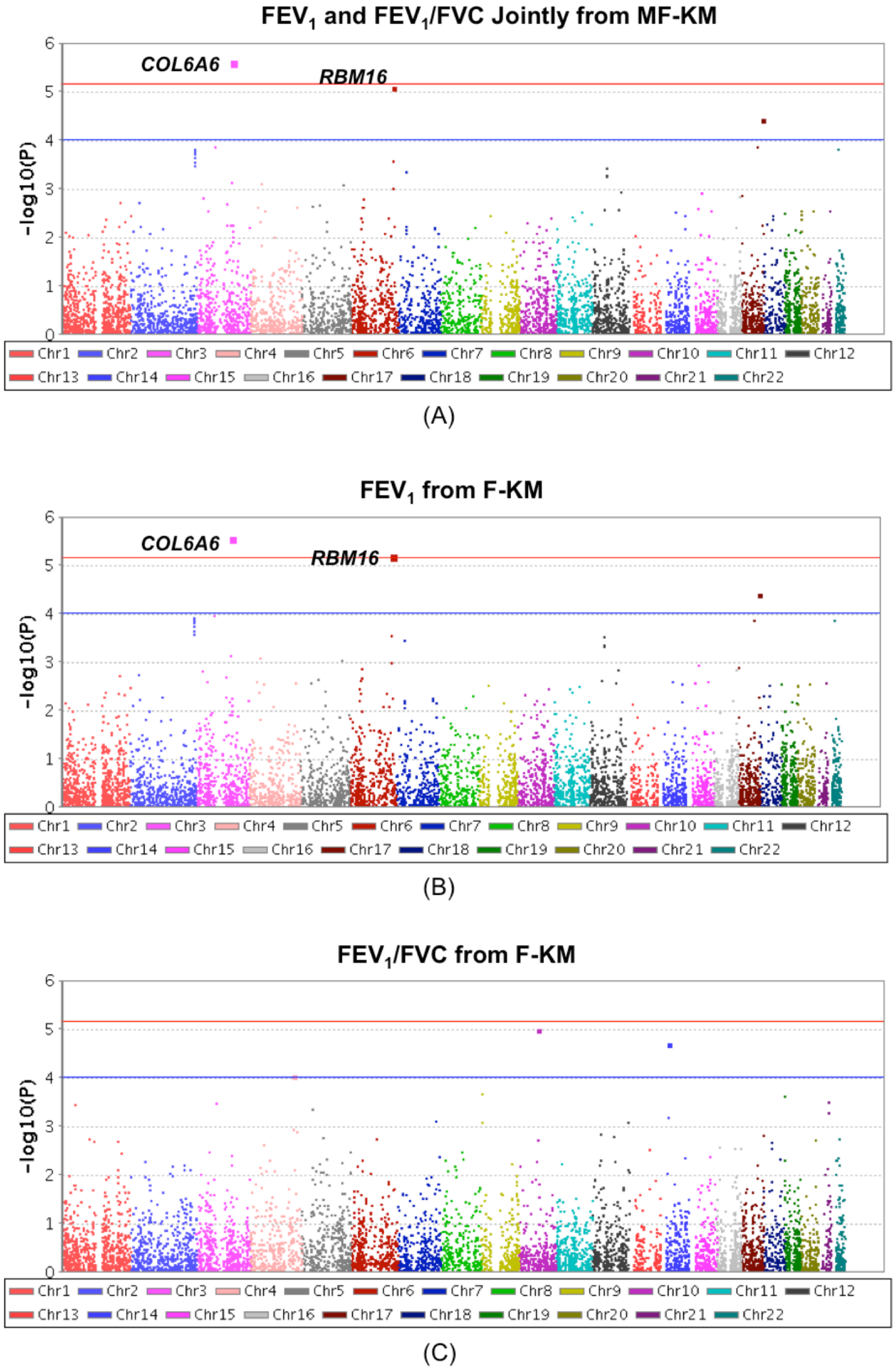
+/- = 30%/0% , Trio, α=0.05, gene effect: pheno1 = 2*pheno2

(F)

Results

➢ Analysis of Genome Wide Lung Function Data:

- 579 subjects, including 316 samples from 13 families;
- 658,502 SNPs were genotyped, where 67,121 are rare variants (MAF<0.05);
- Assigned rare variants to a gene if they are located within a 5kb flank;
- 7,064 genes were used in the analysis;
- Carried out gene-based genome wide association tests of the correlated lung function phenotypes FEV1 (Forced Expiratory Volume in One Second) and FEV1/FVC (Forced Vital Capacity) ratio using MF-KM adjusted for age, gender and height.



FEV$_1$ and FEV$_1$/FVC Jointly from MF-KM

(A)

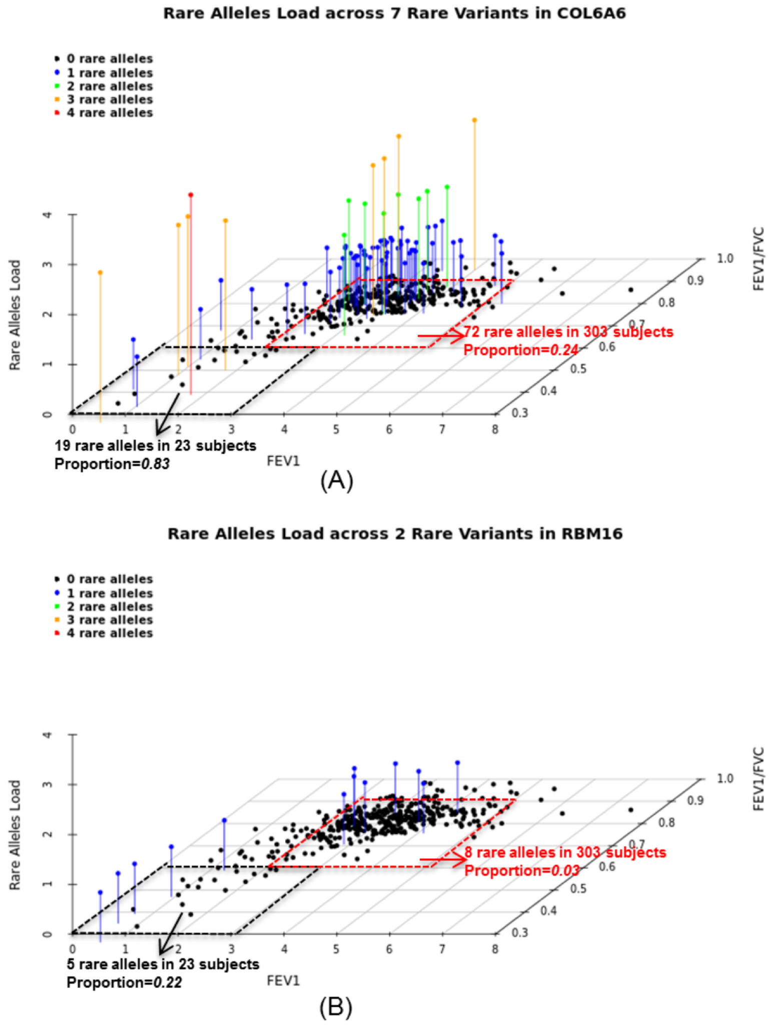
FEV$_1$ from F-KM

(B)

FEV$_1$/FVC from F-KM

(C)

Two genes



Rare Alleles Load across 7 Rare Variants in COL6A6

72 rare alleles in 303 subjects
Proportion=*0.24*

19 rare alleles in 23 subjects
Proportion=*0.83*

(A)

Rare Alleles Load across 2 Rare Variants in RBM16

8 rare alleles in 303 subjects
Proportion=*0.03*

5 rare alleles in 23 subjects
Proportion=*0.22*

(B)

Summary

- Developed the MF-KM statistic using a linear mixed model framework to analyze multivariate data with quantitative traits in family-based studies.

- MF-KM retains the correct Type I error rate, and achieves the best power performance.

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

Acknowledgements

- **University of Pittsburgh**

  Dr. Wei Chen
  Dr. Juan C. Celedón
  Dr. Daniel E. Weeks

- **University of Alabama at Birmingham**

  Dr. Nianjun Liu
  Dr. Hemant K. Tiwari

- **Vanderbilt University**

  Dr. Bingshan Li

- **National Taiwan University**

  Dr. Wan-Yu Lin

- **Tulane University**

  Dr. Xiang-Yang Lou

- **University of Chicago**

  Dr. Guimin Gao

- **Convatec Inc**

  Dr. Xiaojing Wang