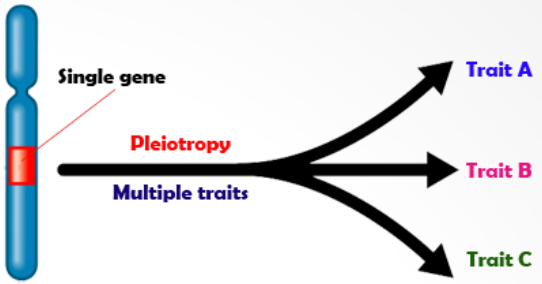
Sequence Kernel Association Test for Multivariate Quantitative Phenotypes in Family Samples

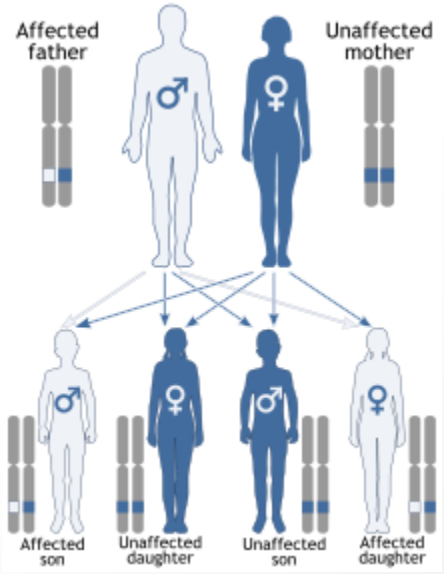
Qi Yan

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

Motivation



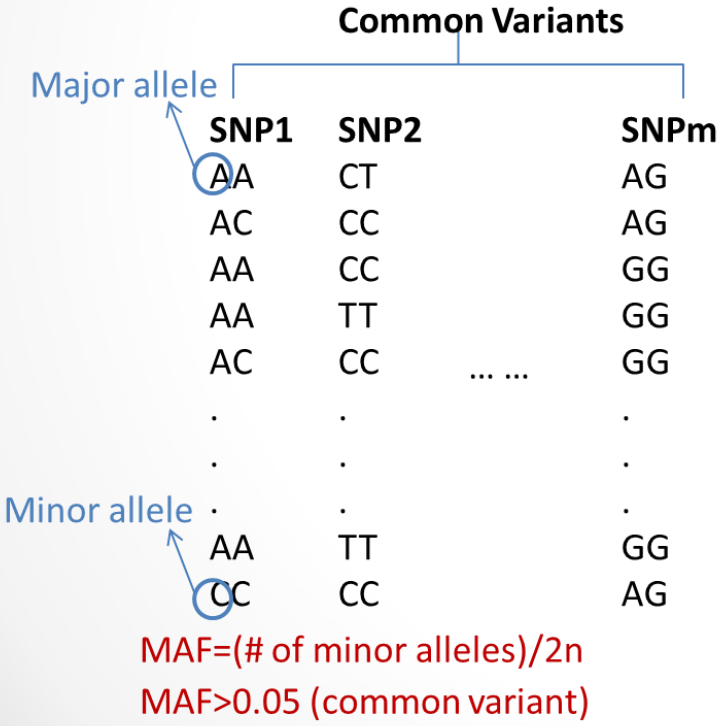
- Phenotypes:
  - Genetic studies have been conducted to collect multiple correlated phenotypes for one complex disease. Jointly modeling multiple phenotypes can improve the statistical power [Sivakumaran S, et al. AJHG. 2011];
  - Family based designs have been widely used [Spielman RS, et al. AJHG.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

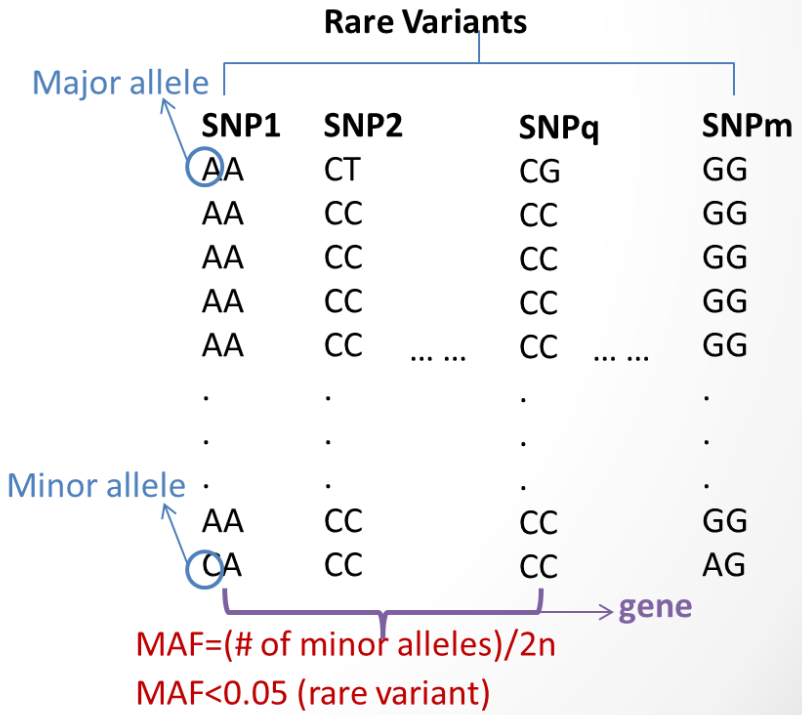


- http://dragonflyissuesinevolution13.wikia.com/wiki/Pleiotropy
  http://en.wikipedia.org/wiki/Hereditary_hemorrhagic_telangiectasia

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).





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 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 $\varepsilon \sim N(0, \ \sigma_E^2 I)$

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

Under the null hypothesis,

$$y = X\beta + h + \delta + \varepsilon$$

continuous phenotypes

covariates

random effect of correlated phenotypes

random effect of familial correlation

random error

For example, one family (father, mother and child), and two correlated phenotypes

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & & & & \\ \sigma_{12} & \sigma_2^2 & & & & \\ & & \sigma_1^2 & \sigma_{12} & & \\ & & \sigma_{12} & \sigma_2^2 & & \\ & & & & \sigma_1^2 & \sigma_{12} \\ & & & & \sigma_{12} & \sigma_2^2 \end{bmatrix} + \sigma_G^2 \begin{bmatrix} 1 & 1 & 0 & 0 & 0.5 & 0.5 \\ 1 & 1 & 0 & 0 & 0.5 & 0.5 \\ 0 & 0 & 1 & 1 & 0.5 & 0.5 \\ 0 & 0 & 1 & 1 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 & 1 \\ 0.5 & 0.5 & 0.5 & 0.5 & 1 & 1 \end{bmatrix} + \sigma_E^2 \begin{bmatrix} 1 & & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & & 1 \end{bmatrix}$$

$Var(h)$ $\qquad$ $Var(\delta)$ $\qquad$ $Var(\varepsilon)$

➢ Simulation Studies

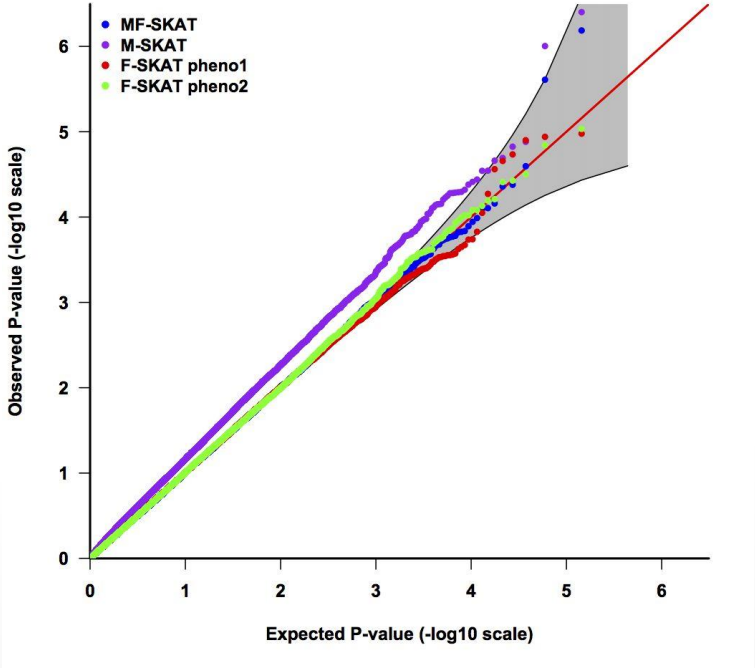
- Genotypes:
  - One set of genotypes = 300 trios ×30 rare variants;
  - Total = 100 sets of genotypes.

- Phenotypes:
  - Type I error rate: 1000 sets of phenotypes for each set of genotypes (independent);
  - Power: 1000 sets of phenotypes for each set of genotypes (Causal variants(+/-) = 30%/0%; 20%/10%; 20%/0%; 13%/7%).

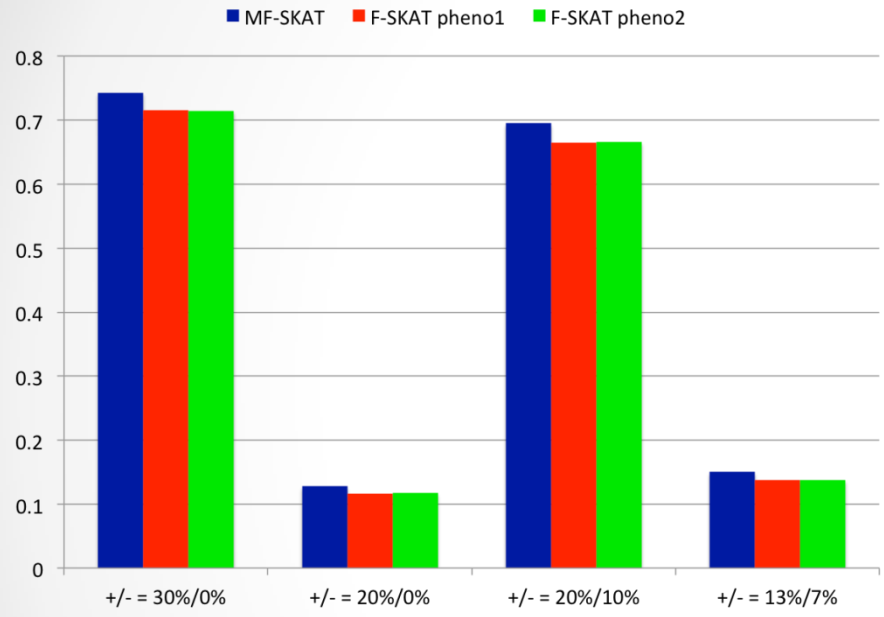
Results

➢ Simulation of the Type I Error Rate:

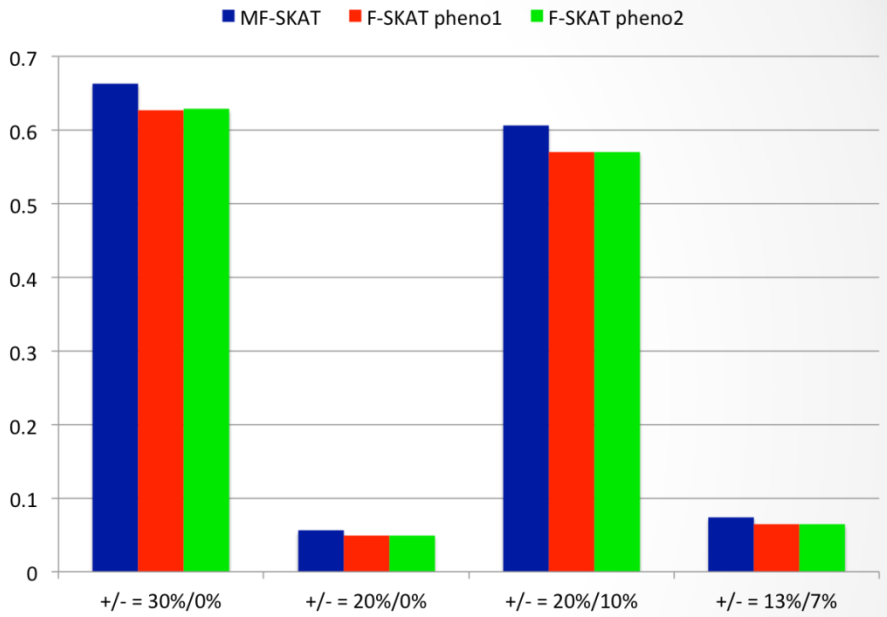
	α=0.05	α=0.01	α=0.005	α=0.001
MF-KM	0.05011	0.01013	0.00500	0.00107
M-KM	0.07481	0.01781	0.00923	0.00213
F-KM pheno1	0.05093	0.01011	0.00505	0.00092
F-KM pheno2	0.05114	0.00991	0.00512	0.00109



➢ Statistical Power Comparison:



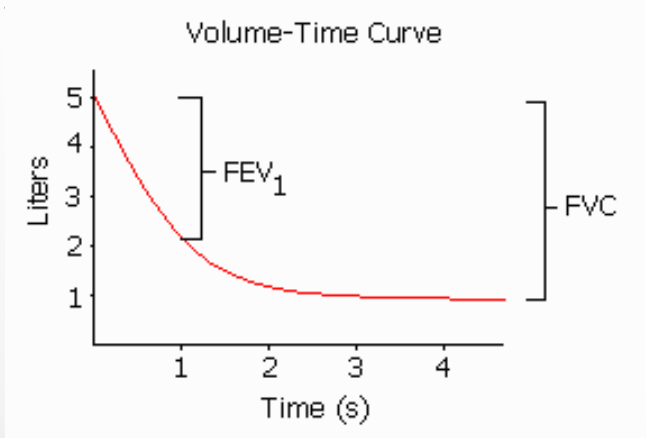
α=0.05



α=0.01

➢ Analysis of Genome Wide Lung Function Data:

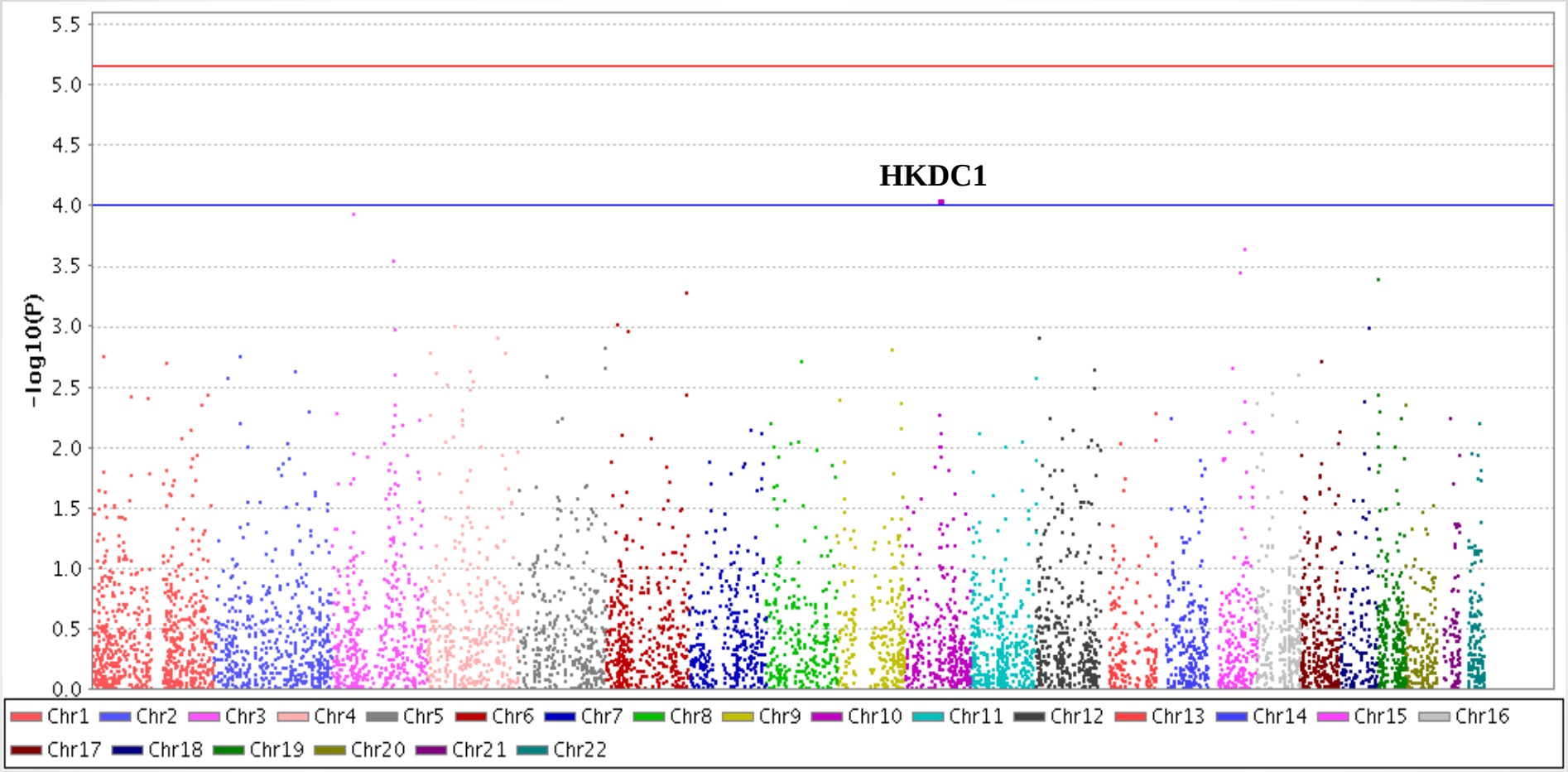
- 579 subjects, including 316 samples from 13 families;
- 658,502 SNPs were genotyped, where 67,121 rare variants (MAF<0.05);
- Assigned rare variants to a gene if they are located within a 5kb flank;
- 7,064 genes were used in the analysis;
- Analyzed the association between the correlated FEV1 & FVC and each gene using MF-KM adjusted for age, gender and height



FEV1: forced expiratory volume in 1st second;
FVC: forced vital capacity.

In this data, cor(FEV1, FVC) = 0.95

https://meded.ucsd.edu/isp/1998/asthma/html/spirexp.html



Results of MF-SKAT on lung function analysis. We tested the association between 7,064 genes in which they have SNPs with MAF < 0.05 and the correlated phenotypes, FEV1 and FVC.

➢ Analysis of Dental Caries Data (dbGaP):

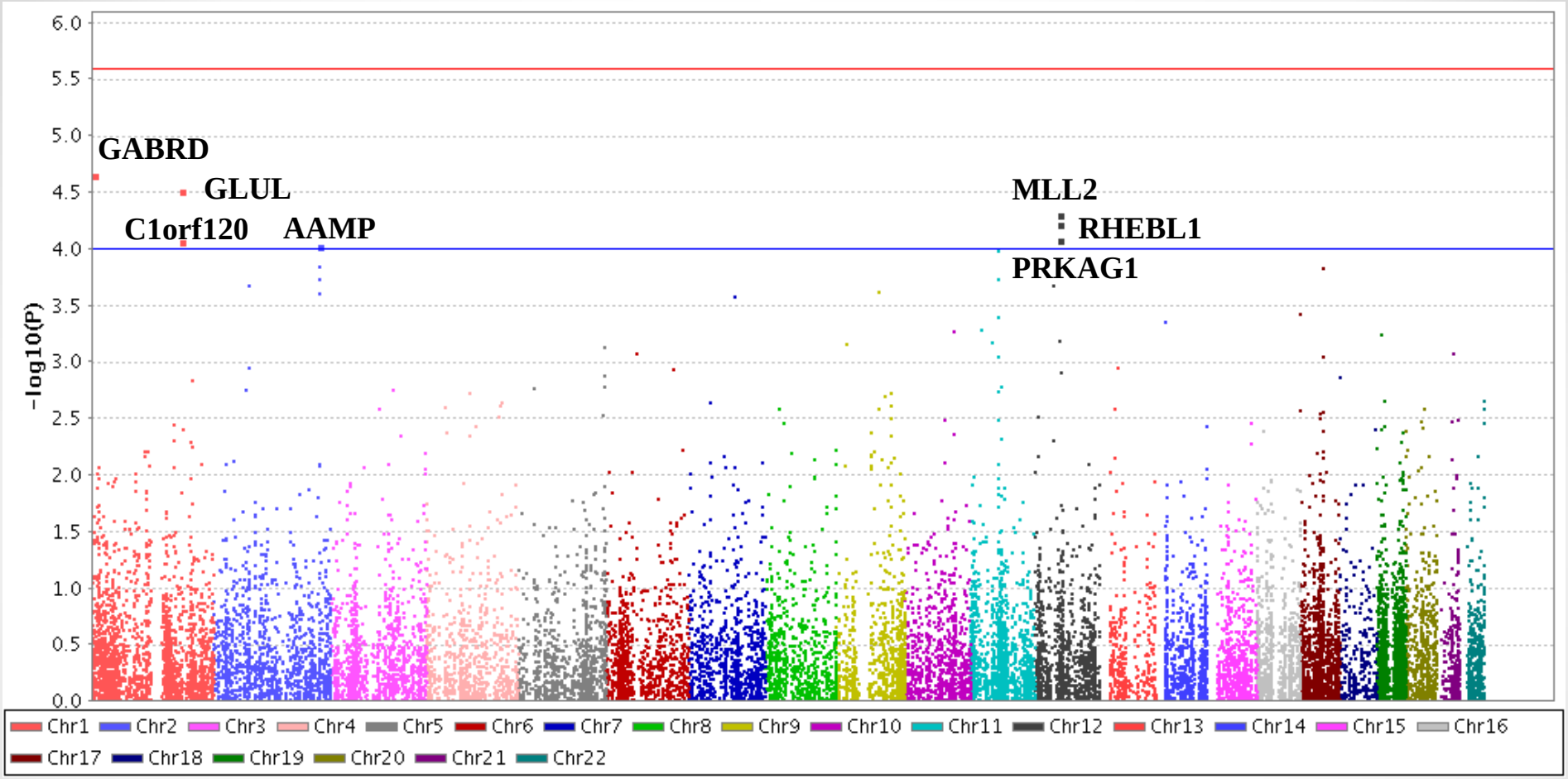
- 4,016 subjects from 1,874 families;
- 16,219,283 imputed SNPs, where 9,769,821 rare variants (MAF<0.05);
- Assigned rare variants to a gene if they are located within a 5kb flank;
- 19,564 genes were used in the analysis;
- Analyzed the association with Decayed, Missing due to Decay, and Filled tooth surfaces simultaneously considering their correlation and controlling for age and gender.

In this data, cor(Decayed surfaces (DS), Surfaces missing due to decay (MS), Filled tooth surfaces (FS)) =

	DS	FS	MS
DS	1	0.02	0.21
FS	0.02	1	0.05
MS	0.21	0.05	1

DMFS	DS	FS	MS
7	0	7	0
4	3	1	0
10	10	0	0
17	2	15	0
49	29	0	20

Commonly used phenotype, DMFS=DS+MS+FS



Results of MF-SKAT on dental caries analysis. We tested the association between 19,564 genes in which they have SNPs with MAF < 0.05 and the correlated phenotypes, Decayed, Missing due to Decay, and Filled tooth surfaces.

Summary

- Implement MF-SKAT for testing the association of rare variants in family samples, which simultaneously considers correlated phenotypes.
- MF-SKAT retains the correct Type I error rate, and achieves the best power performance.
- Observe potential important genes associated with lung function and dental caries.
- The software will be available.

Acknowledgements

- **University of Pittsburgh**

  Dr. Wei Chen
  Dr. Juan Celedon
  Dr. Daniel Weeks

- **University of Alabama at Birmingham**

  Dr. Nianjun Liu

- **GlaxoSmithKline**

  Dr. Xiaojing Wang

- **Vanderbilt University**

  Dr. Bingshan Li