An Omnibus Test for Gene-Level Effects of Multi-Omics Data with Application to Childhood Asthma

Qi Yan, Wei Chen

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

December 22th, 2016

Motivation

- This study is motivated by the work (R01 HL117191 and R01 HL079966) on genetics, epigenetics and asthma in Puerto Rican children.

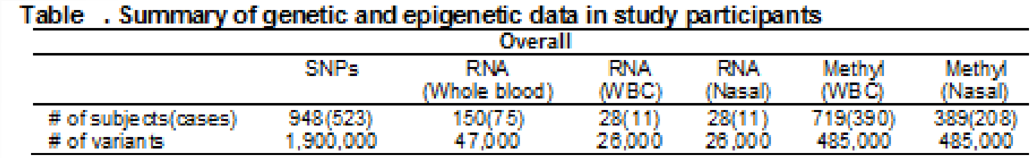


Table . Summary of genetic and epigenetic data in study participants

	Overall					
	SNPs	RNA (Whole blood)	RNA (WBC)	RNA (Nasal)	Methyl (WBC)	Methyl (Nasal)
# of subjects(cases)	948(523)	150(75)	28(11)	28(11)	719(390)	389(208)
# of variants	1,900,000	47,000	26,000	26,000	485,000	485,000

SNPs (0, 1, 2)

RNA expression (continuous)

DNA methylation (continuous)

- It has been reported that SNPs, DNA methylation and RNA expression are associated with childhood asthma individually. However, some genes may have weak individual effects that are hard to detect, but the joint effect is detectable.

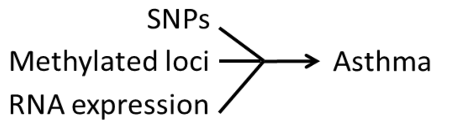


Figure 1

Aim

- To develop novel statistical approaches for testing the overall effect of SNPs, DNA methylation and RNA expression.

  - ➢ The developed omnibus (i.e., overall) tests help us identify additional genes that have weak but real effects of the three factors on asthma related phenotypes in Puerto Rican children. These genes could be missed in the one factor test.

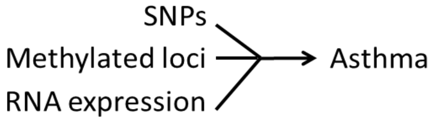


Figure 1

Methods

1. Test gene-level effects of SNPs, DNA methylation and RNA expression separately using Kernel Machine (KM) regression:

Use SNPs for illustration: let there be *n* subjects with *q* genetic variants. The $n \times 1$ vector of the continuous trait ***y*** follows a linear model:

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

when the phenotypes are binary, **y** follows a logistic model:

$$\text{logit}\, P(\mathbf{y} = \mathbf{1}) = \mathbf{X}\boldsymbol{\beta} + \mathbf{G}\boldsymbol{\gamma}$$

- **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),
- **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,

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

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

where **W** is a predefined $q \times q$ diagonal weight matrix for each variant

Methods

➢ 1. Test gene-level effects of SNPs, DNA methylation and RNA expression separately using Kernel Machine (KM) regression:

Following the same rationale as in the derivation of the SKAT score statistic [1], the test statistic is:

Continuous trait:

$$\mathrm{Q} = \left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\mathbf{GWG}'\left(\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)/\hat{\sigma}_E^2$$

Under H_0: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$

The estimates:

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

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

$$\mathbf{P_0} = \mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$$

Binary trait:

$$\mathrm{Q} = (\mathbf{y} - \widehat{\boldsymbol{\mu}})'\mathbf{GWG}'(\mathbf{y} - \widehat{\boldsymbol{\mu}})$$

Under H_0: $\text{logit}\, P(\mathbf{y} = \mathbf{1}) = \mathbf{X}\boldsymbol{\beta}$

The estimates:

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

$$\widehat{\boldsymbol{\mu}} = \text{logit}^{-1}\left(\mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

$$\widehat{\boldsymbol{\Sigma}} = diag\left(\widehat{\boldsymbol{\mu}} \cdot (1 - \widehat{\boldsymbol{\mu}})\right)$$

$$\mathbf{P_0} = \widehat{\boldsymbol{\Sigma}} - \widehat{\boldsymbol{\Sigma}}\mathbf{X}\left(\mathbf{X}'\widehat{\boldsymbol{\Sigma}}\mathbf{X}\right)^{-1}\mathbf{X}'\widehat{\boldsymbol{\Sigma}}$$

- 1. Wu MC, Lee S, Cai T, Li Y, Boehnke M, Lin X: Rare-variant association testing for sequencing data with the sequence kernel association test. American journal of human genetics 2011;89:82-93.

Methods

- **1. Test gene-level effects of SNPs, DNA methylation and RNA expression separately using Kernel Machine (KM) regression:**

Continuous trait:

$$Q = \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right)' \mathbf{GWG}' \left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}\right) / \hat{\sigma}_E^2$$

Binary trait:

$$Q = (\mathbf{y} - \hat{\boldsymbol{\mu}})' \mathbf{GWG}' (\mathbf{y} - \hat{\boldsymbol{\mu}})$$

The statistic Q is a quadratic form and follows a mixture of chi-square distributions under H_0. Thus,

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

where λ_i are the eigenvalues of the matrix $\mathbf{P_0}^{\frac{1}{2}} \mathbf{GWG}' \mathbf{P_0}^{\frac{1}{2}}$ for both continuous and binary traits. The *p*-values can be calculated by numerical algorithms, such as Davies’ method.

Analogously, the gene-level effects of DNA methylation and RNA expression can be tested by replacing **Gγ** with **Mρ** and **E**γ

Methods

➢ 2. Modified Fisher's method for combining gene-level effects of SNPs, DNA methylation and RNA expression:

- Aim to have one single *p*-value to represent the significance of a gene;
- Use Fisher's method to combine three *p*-values (SNPs, DNA methylation and RNA expression) to one, but *p*-values may not be independent;
- Consider a Satterthwaite approximation by approximating a scaled *T* statistic with a new chi-square distribution:

$$cT \approx \chi_v^2, \text{ where } c = \frac{v}{E(T)}, v = 2\frac{[E(T)]^2}{\mathrm{Var}(T)},$$

$$E(T) = E\left(-2\sum_{i=1}^{w} \ln(p_i)\right) = 2w,$$

$$\mathrm{Var}(T) = \mathrm{var}\left(-2\sum_{i=1}^{w} \ln(p_i)\right) = 4w + 2\sum_{i<j} \mathrm{cov}\left(-2\ln(p_i), -2\ln(p_j)\right)$$

- where *w* = 3 for SNPs, DNA methylation and RNA expression.
- The covariance part takes the correlations of *p*-values into account and can be empirically estimated by perturbations.

Methods

➢ 3. Optimal test for the gene-level effects of SNPs, DNA methylation and RNA expression using perturbations:

- If the disease risk only depends on SNPs and the model with SNPs, RNA expression and DNA methylation is used, then the testing power will lose.
- Since in reality we do not know the underlying true disease model, it is difficult to choose the correct model.
- Thus, it is desirable to develop a method accommodating all possible disease models to maximize the power.
- This can be achieved by using the minimum *p*-value of all possible models as a new test statistic.
- Then, perturbation can be used to calculate the final *p*-value.

Methods

➢ 3. Optimal test for the gene-level effects of SNPs, DNA methylation and RNA expression using perturbations:

The intuition behind the perturbation:

For continuous phenotypes, with large *n*, under H_0 the$(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/\hat{\sigma}_{\boldsymbol{E}}$ is approximately standard normal. Then each $\mathrm{Q}_d=(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})'\mathbf{GWG}'(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/\hat{\sigma}_E^2$ is essentially comprised of a vector of standard normal variables sandwiching a square matrix. The vectors of normal values are the same across all $\mathrm{Q}_1,\cdots\ \mathrm{Q}_k$. Thus, we can perturb each Q_d by replacing$(\mathbf{y}-\mathbf{X}\widehat{\boldsymbol{\beta}})/\hat{\sigma}_{\boldsymbol{E}}$ with a new, common vector of normal values to generate new score statistics.

Methods

3. Optimal test for the gene-level effects of SNPs, DNA methylation and RNA expression using perturbations:

Perturbation:

1. Calculate the *p*-values for SNPs (G), DNA methylation (M) and RNA expression (E) separately (i.e., $p_G^{(0)}$, $p_M^{(0)}$, and $p_E^{(0)}$) by KM regression. Observed individual *p*-values

2. For *l* ∈ *G*, *M* and *E*, compute $\boldsymbol{\Lambda}_l = diag\left(\lambda_{l,1}, \cdots, \lambda_{l,ml}\right)$, and $\mathbf{V}_l = \left[\boldsymbol{v}_{l,1}, \cdots, \boldsymbol{v}_{l,ml}\right]$ where $\lambda_{l,1} \geq \lambda_{l,2} \geq \cdots \geq \lambda_{l,ml}$ are the ml positive eigenvalues of $\mathbf{P_0}^{\frac{1}{2}}\mathbf{GWG}'\mathbf{P_0}^{\frac{1}{2}}$ with corresponding eigenvectors $\boldsymbol{v}_{l,1}, \cdots, \boldsymbol{v}_{l,ml}$.

3. Generate $\boldsymbol{r}^{(b)} = \left[r_1^{(b)}, \cdots, r_n^{(b)}\right]'$ with each $r_j^{(b)} \sim N(0,1)$.

4. For *l* ∈ *G*, *M* and *E*, rotate $\boldsymbol{r}^{(b)}$ using the eigenvectors to generate $\boldsymbol{r}_l^{(b)} = \mathbf{V}_l'\boldsymbol{r}^{(b)}$.

5. Compute $\mathrm{Q}_l^{(b)} = \boldsymbol{r}_l^{(b)'}\boldsymbol{\Lambda}_l\boldsymbol{r}_l^{(b)}$ for each *l* and obtain a corresponding *p*-value, $p_l^{(b)}$.

6. Repeat (3)-(5) *B* times to obtain $p_G^{(1)}, p_G^{(2)}, \cdots, p_G^{(B)}$, $p_M^{(1)}, p_M^{(2)}, \cdots, p_M^{(B)}$ and $p_E^{(1)}, p_E^{(2)}, \cdots, p_E^{(B)}$ for some large number *B*. Perturbation individual *p*-values

7. Calculate the covariance between p_G, p_M and p_E by using $p_G^{(b)}$, $p_M^{(b)}$, and $p_E^{(b)}$ for *b* ∈ 0, 1,…, *B*.

8. Calculate the joint *p*-values of SNPs, DNA methylation and RNA expression (i.e., for *b* ∈ 0, 1,…, *B*, $p_{GM}^{(b)}$, $p_{GE}^{(b)}$, $p_{ME}^{(b)}$, and $p_{GME}^{(b)}$) by modified Fisher's method. The *p*-values for all possible disease models

9. For *l** ∈ *G*, *M*, *E*, *GM*, *GE*, *ME*, and *GME*; *b* ∈ 0, 1,…, *B*, set $p^{(b)} = \min_{1 \leq l^* \leq L^*} p_{l^*}^{(b)}$.

10. The final p-value for significance is estimated as

$$p = B^{-1}\sum_{b=1}^{B} I\left(p^{(b)} \leq p^{(0)}\right)$$

Methods

➢ Simulation Studies

We simulated two main settings:

1. G, M and E are independent with each other:

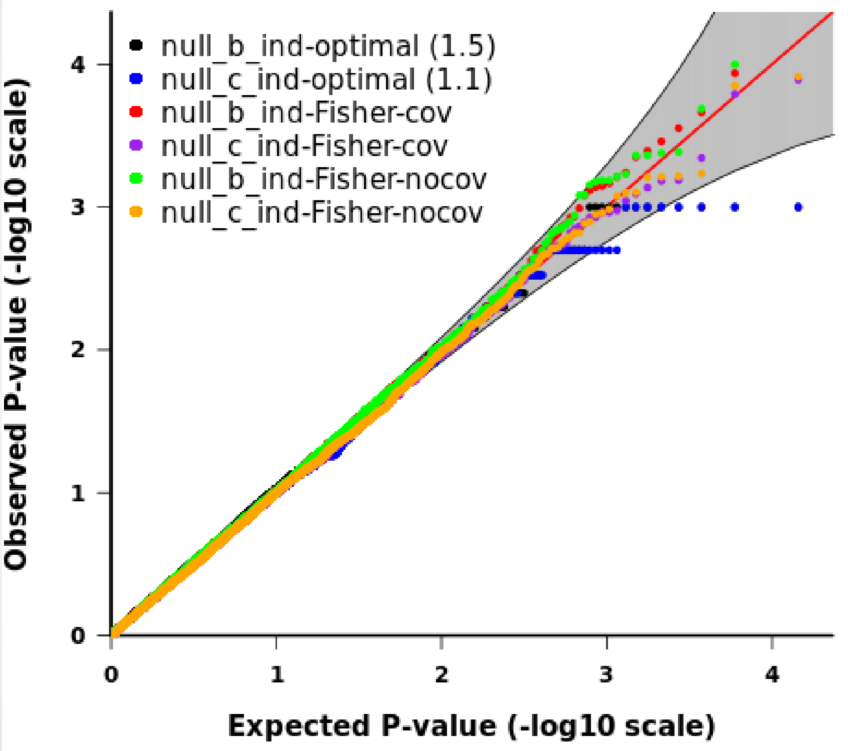
(1.1) continuous traits and no causal factors (denoted as null_c_ind),
(1.2) continuous traits and G is causal (denoted as causal_G_c_ind),
(1.3) continuous traits and G and M are causal (denoted as causal_GM_c_ind),
(1.4) continuous traits and G, M and E are causal (denoted as causal_GME_c_ind),
(1.5) binary traits and no causal factors (denoted as null_b_ind),
(1.6) binary traits and G is causal (denoted as causal_G_b_ind),
(1.7) binary traits and G and M are causal (denoted as causal_GM_b_ind),
(1.8) binary traits and G, M and E are causal (denoted as causal_GME_b_ind)

2. G and E are correlated, but independent with M:

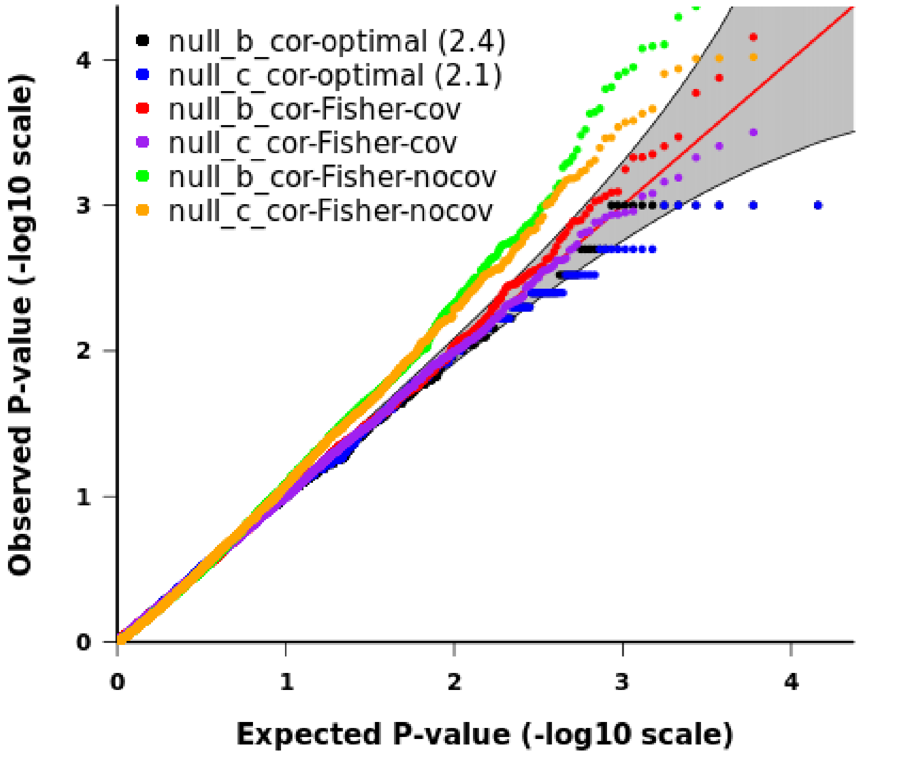
(2.1) continuous traits and no causal factors (denoted as null_c_cor),
(2.2) continuous traits and G is causal (denoted as causal_G_c_cor),
(2.3) continuous traits and G and M are causal (denoted as causal_GM_c_cor. When the causal SNPs in G are correlated with E, G and M causal is similar to G, E and M causal),
(2.4) binary traits and no causal factors (denoted as null_b_cor),
(2.5) binary traits and G is causal (denoted as causal_G_b_cor),
(2.6) binary traits and G and M are causal (denoted as causal_GM_b_cor).

Results

➢ Simulation of the Type I Error Rate:



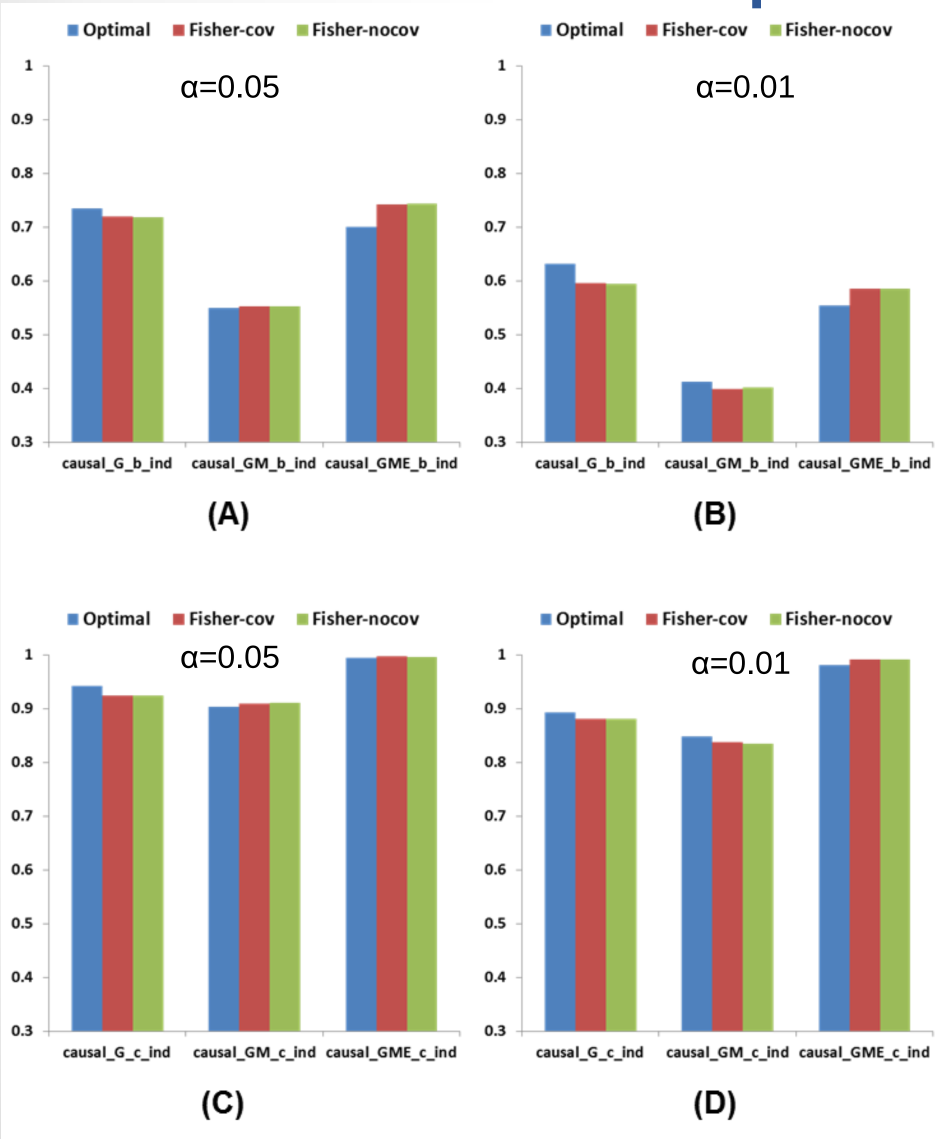
Samples with independent G, M and E



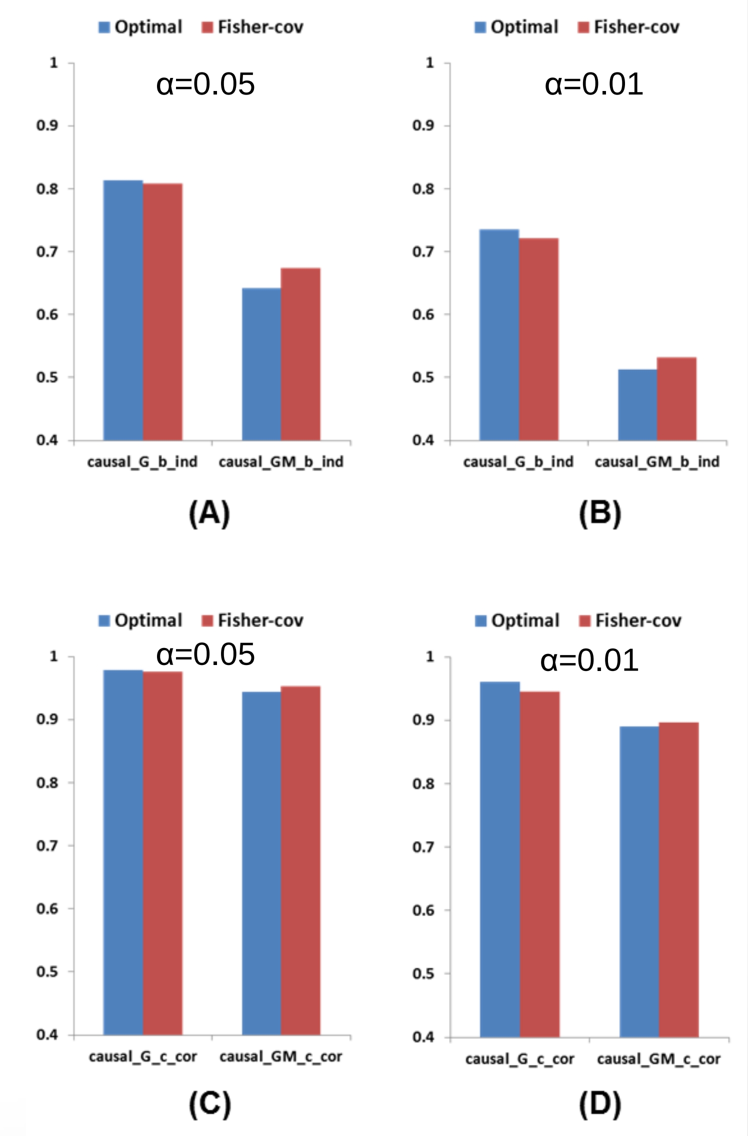
Samples with G and E correlated

Results

➢ Statistical Power Comparison:



Samples with independent G, M and E

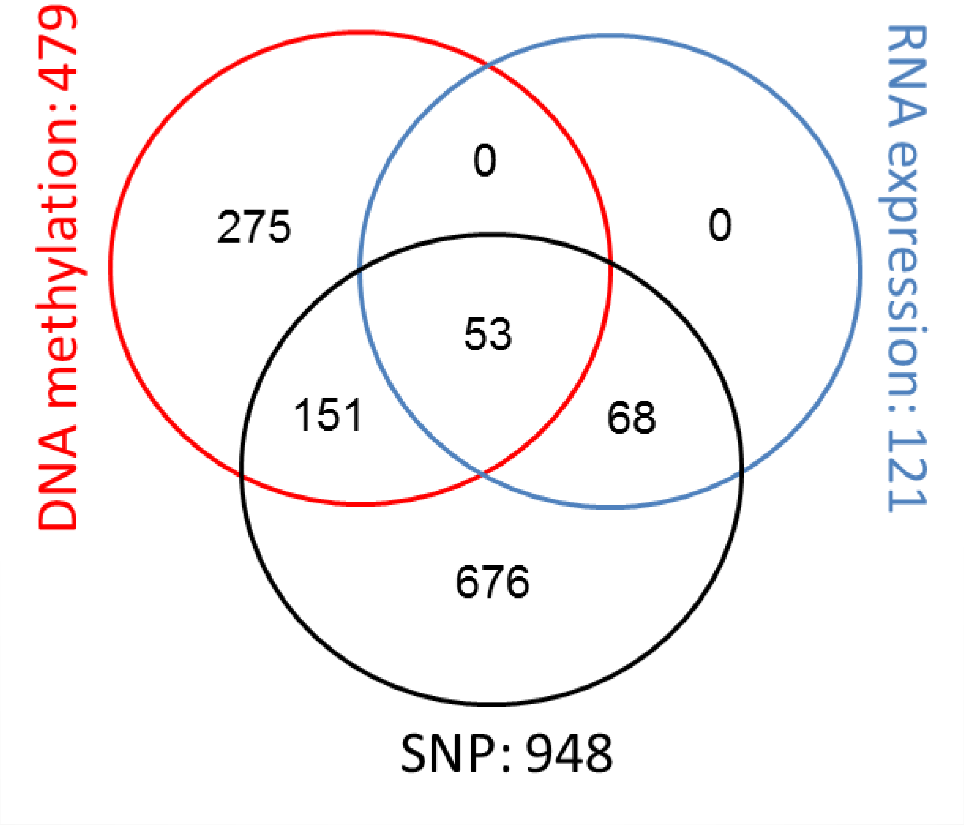


Samples with G and E correlated

Results

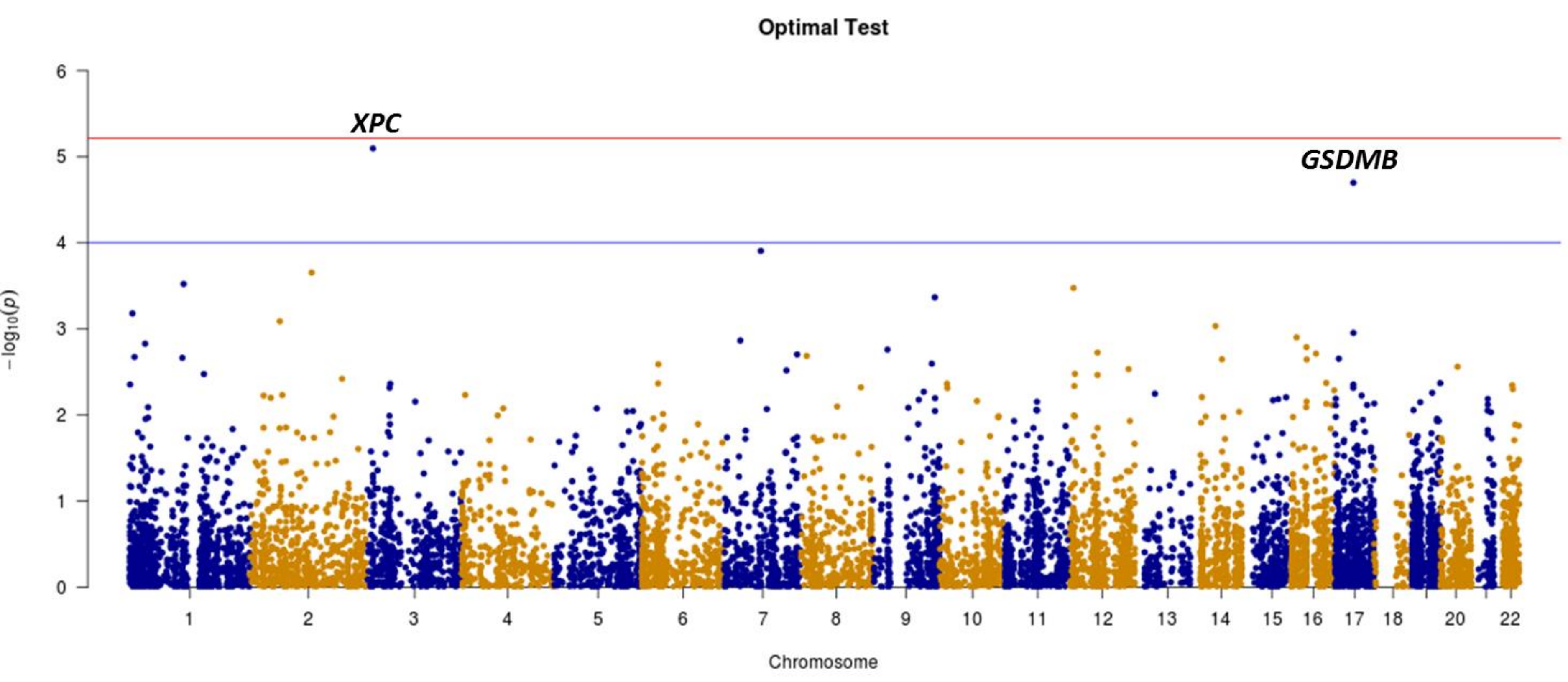
➢ Analysis of Genome Wide Childhood Asthma Data:

- In the final analysis, we used 1,223 subjects;
- We used the 8,254 genes with all SNP, DNA methylation and RNA expression information



Results

➢ Analysis of Genome Wide Childhood Asthma Data:



- The *GSDMB* gene could be served as a positive control in asthma genetic studies. In the optimal test, the significance of *GSDMB* ($P = 2\times10^{-5}$) was mainly driven by the genetic effect ($P = 7.54\times10^{-6}$).

- The *XPC* gene ($P = 8\times10^{-6}$) was suggestively associated with asthma driven by the methylation effect ($P = 3.14\times10^{-6}$). The *XPC* gene was reported to play an important role in lung carcinogenesis and air pollution induced pathogenesis of the inflammatory disease bronchitis.

Summary

- Developed an approach to test the overall gene effects from multiple omics data using a modified Fisher’s method.

- Further extended this approach to consider all possible disease models using perturbation.