iGWAS: Integrative Genome-Wide Association Studies

Xihong Lin

Harvard Chan School of Public Health

Joint work with
Yen-Tsung Huang(Brown Univ) & Richard Barfield (Harvard)

Outline

Outline

Different Types of Genetic and Genomic Data

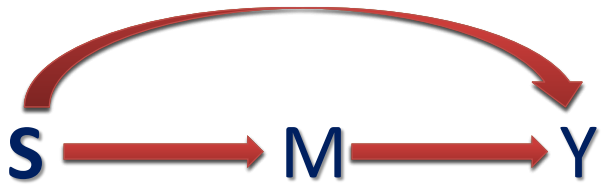
- Different types of genetic, genomic and environmental data are rapidly available:
  - GWAS
  - Genomics data, e.g., gene expressions, RNA sequencing, and DNA methylation data.
- Data complexities
  - Family studies (within-family correlation)
  - Missing data: a subset of GWAS subjects have expression/methylation data

Integrative GWAS (iGWAS)

- Integrate genetic (SNPs from GWAS/sequencing), genomic (gene expressions/DNA methylation) data, and environmental data to understand disease etiologic mechanism.
- Account for complexities in the data: within-family correlation and missing data.

Outline

Direct and Indirect Effects of a SNP Set

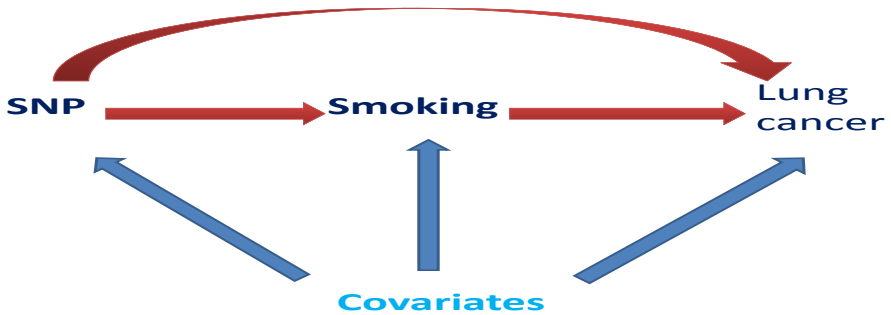


- Direct effect (DE) of a SNP set: The effect of a SNP set(S) independent of mediator (M), e.g., gene expression(G)/environment(E), on disease/trait (Y)
- Indirect effect (IE) of a SNP set: The effect of a SNP set (S) on disease (Y) mediated through mediator (M).
- Total effect (TE) of a SNP set =DE+IE

Mediation Analysis of SNP, Smoking and Lung Cancer (VanderWeele, et al, 2012)

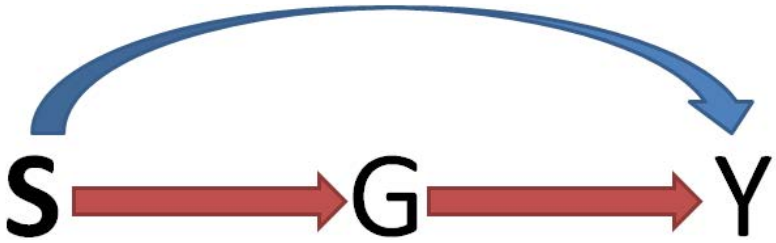
- Four cohorts: MGH, MD Anderson, IARC and Toronto (n=12,000)
- Two GWAS hit SNPs: rs8034191 and rs1051730, which are associated with both smoking and lung cancer.
- Question:
  Are the effects of these two SNPs on lung cancer mediated through smoking?

DE and IE of SNP rs8034191 on Lung Cancer Risk



	DE		IE		
	OR	p-value	OR	p-value	% Mediated
MGH	1.35	3.1×10^{-7}	1.01	0.15	3.6%
MDA	1.18	2.0×10^{-4}	1.01	0.13	6.8%
IARC	1.26	5.4×10^{-6}	1.00	0.95	0.2%
Toronto	1.33	0.04	1.00	0.87	0.4%
Meta-analysis	1.26	1.8×10^{-15}	1.01	0.09	3.2%

DE, IE and TE of a SNP set



- Models for gene expression and disease

$$g(\mu_i) = \mathbf{X_i}^T \alpha + \mathbf{S_i}^T \beta_S + G_i \beta_G + G_i \mathbf{S_i}^T \beta_I$$

$$G_i = \mathbf{X_i}^T \phi + \mathbf{S_i}^T \delta + \varepsilon_{Gi}$$

Direct effect (DE) and indirect effect (IE)

$$DE = (\mathbf{s_1} - \mathbf{s_0})^T [\beta_S + \beta_I (\mathbf{x}^T \phi + \mathbf{s_0}^T \delta + \beta_G \sigma_G^2)] + \frac{1}{2} \sigma_G^2 (\mathbf{s_1} + \mathbf{s_0})^T \beta_I (\mathbf{s_1} - \mathbf{s_0})^T \beta_I$$

$$IE = (\mathbf{s_1} - \mathbf{s_0})^T \delta (\beta_G + \mathbf{s_1}^T \beta_I)$$

Test for Direct, Indirect and Total Effects (DE, IE, TE)

- Recall: Model

$$g\{\mu_i\} = \mathbf{X_i}^T\alpha + \mathbf{S_i}^T\beta_S + G_i\beta_G + G_i\mathbf{S_i}^T\beta_I$$

- Assuming eQTL, hypothesis Tests of Interest:

$$\begin{aligned} H_0 : DE = 0 \quad & H_0 : \beta_S = 0, \beta_I = 0 \\ H_0 : IE = 0 \quad & H_0 : \beta_S = 0, \beta_G = 0 \\ H_0 : TE = 0 \quad & H_0 : \beta_S = 0, \beta_G = 0, \beta_I = 0 \end{aligned}$$

Variance Component Tests for DE, IE and TE

- As # of SNPs in gene might be large, assume

$$\beta_{\boldsymbol{S}} \sim F_1(\mathbf{0}, \tau_S I_p) \quad \text{and} \quad \beta_{\boldsymbol{I}} \sim F_2(\mathbf{0}, \tau_I I_p)$$

  where $F_1(\cdot)$ and $F_2(\cdot)$ are arbitrary distributions.
- Assuming an eQTL, null hypotheses of interest are equivalent to:

$$\begin{aligned} H_0 : DE = 0 : &\qquad H_0 : \tau_S = 0, \tau_I = 0 \\ H_0 : IE = 0 : &\qquad H_0 : \beta_G = 0, \tau_I = 0 \\ H_0 : TE = 0 : &\qquad H_0 : \beta_G = 0, \tau_S = 0, \tau_I = 0 \end{aligned}$$

Outline

iGWAS for Family Studies

- Need to account for within-family correlation.
- Idea:
  - Construct estimating equations for regression coefficients β_G and variance components τ_S and τ_I.
  - Test for DE, IE and TE using sandwich-based variance component "score type" estimating equation based tests.

Test statistic for Direct Effect (DE)

- Estimating equation based joint tests for fixed effects and variance components.
- Test statistics for Direct Effect (DE)

$$Q_{DE} = (\mathbf{Y} - \hat{\mu}_{DE})^T \boldsymbol{R}^{-1} (a_1 \mathbf{S}\mathbf{S}^T + a_3 \mathbf{C}\mathbf{C}^T) \boldsymbol{R}^{-1} (\mathbf{Y} - \hat{\mu}_{DE})$$

where $C_i = \mathbf{S}_i G_i$ and $\boldsymbol{R}^{-1}$ is a working correlation matrix.

$$g(\mu_{DE}) = \beta_0 + \mathbf{X}\beta_x + \beta_G G$$

and $a_1 = \sqrt{var(U_{\tau_S})}$ and $a_3 = \sqrt{var(U_{\tau_I})}$ are sandwich variance estimators.

Test for Indirect Effects(IE)

- Test statistic for IE:

$$Q_{IE} = (\mathbf{Y} - \hat{\mu}_{IE})^T \mathbf{R}^{-1} (a_2 \mathbf{G}\mathbf{G}^T + a_3 \mathbf{C}\mathbf{C}^T) \mathbf{R}^{-1} (\mathbf{Y} - \hat{\mu}_{IE})$$

  where $g(\mu_{IE}) = \beta_0 + \mathbf{X}\alpha + \mathbf{S}\beta_S$ and is fitted using ridge regression.

- Test statistic for TE:

$$Q_{TE} = (\mathbf{Y} - \hat{\mu}_0)^T \mathbf{R}^{-1} (a_1 \mathbf{S}\mathbf{S}^T + a_2 \mathbf{G}\mathbf{G}^T + a_3 \mathbf{C}\mathbf{C}^T) \mathbf{R}^{-1} (\mathbf{Y} - \hat{\mu}_0)$$

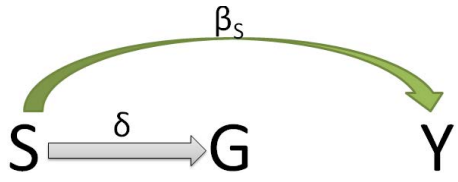
  where $g(\mu_0) = \beta_0 + \mathbf{X}\alpha$

Distributions of the Test Statistics

- All the test statistics take quadratic forms of Y asymptotically.
- Their null distributions are a mixture of chi-squares asymptotically and can be approximated using the Davies method.

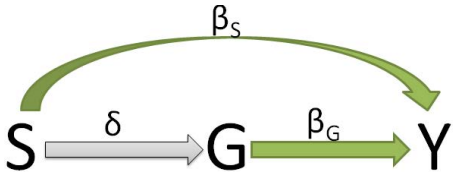
Omnibus Test (Focus on TE)

- SNP-only causal model



$$Q_S = m^{-1}(\mathbf{Y} - \mu_\mathbf{0})^T \mathbf{R}^{-1}(a_1 \mathbf{S}\mathbf{S}^T)\mathbf{R}^{-1}(\mathbf{Y} - \mu_\mathbf{0})$$

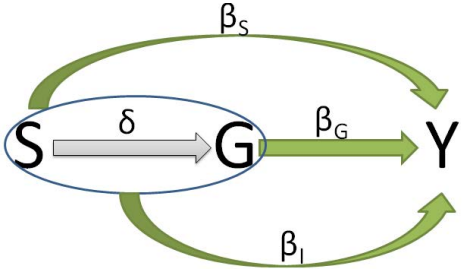
- SNP+G main effect only causal model:



$$Q_{SG} = m^{-1}(\mathbf{Y} - \mu_\mathbf{0})^T \mathbf{R}^{-1}(a_1 \mathbf{S}\mathbf{S}^T + a_2 \mathbf{G}\mathbf{G}^T)\mathbf{R}^{-1}(\mathbf{Y} - \mu_\mathbf{0})$$

Three Causal Models

- SNP+G interaction causal model



$$Q_{SGI} = m^{-1}(\mathbf{Y} - \mu_{\mathbf{0}})^T \mathbf{R}^{-1}(a_1 \mathbf{S}\mathbf{S}^T + a_2 \mathbf{G}\mathbf{G}^T + a_3 \mathbf{C}\mathbf{C}^T)\mathbf{R}^{-1}(\mathbf{Y} - \mu_{\mathbf{0}})$$

- Which model to use?

Omnibus Test for Different causal models

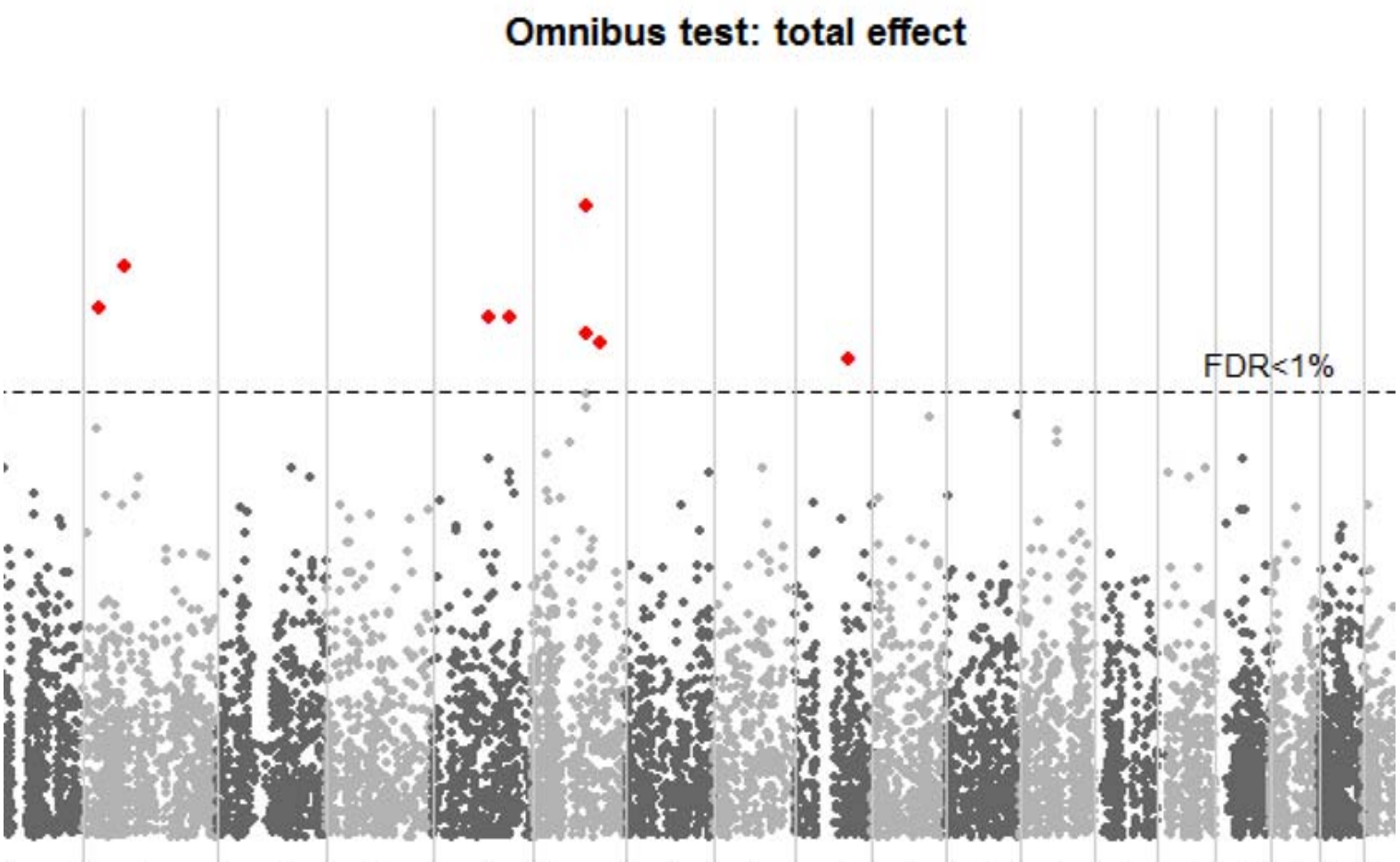
- **Omnibus test:** Test for the SNP set effect under each of three models and calculate the minimum p-value.
- The omnibus test can improve the test power without knowing the true model.
- The null distribution of the omnibus test is constructed using estimating equation based perturbation.

IGWAS Analysis of Asthama Data

- MRCA study: GWAS study of families of the British descents:
- Data: 378 subjects (266 cases and 112 controls).
- Control subjects were either siblings or parents of the cases.
- GWAS: 300K SNP
- Gene expression
- Gene-based analysis: 12,000 genes

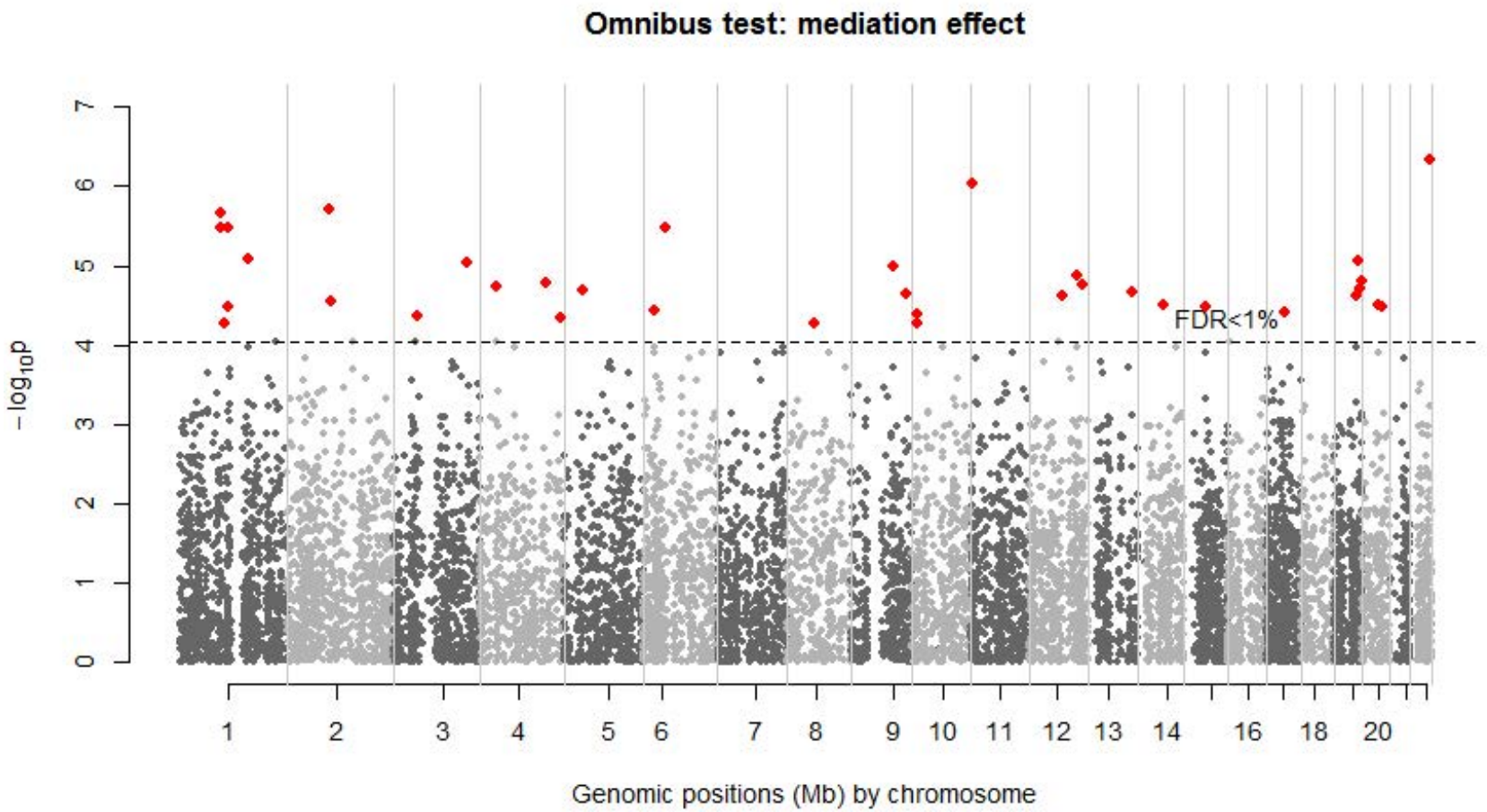
iGWAS results: Total Effect (TE)

- 8 genes with FDR<1%



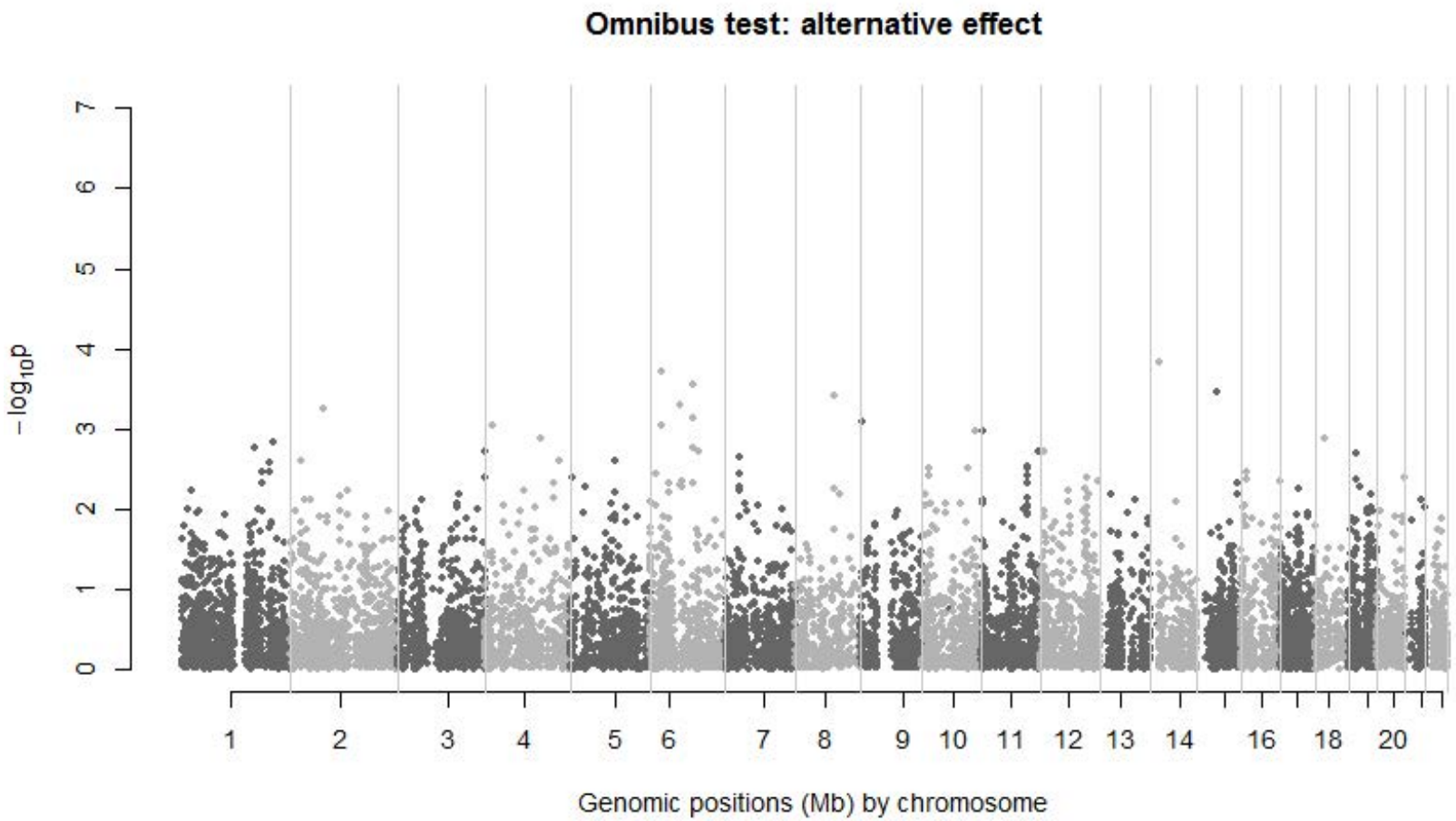
iGWAS results: Mediation Effect/Indirect Effect (ME/IE)

- 36 genes with FDR<1%

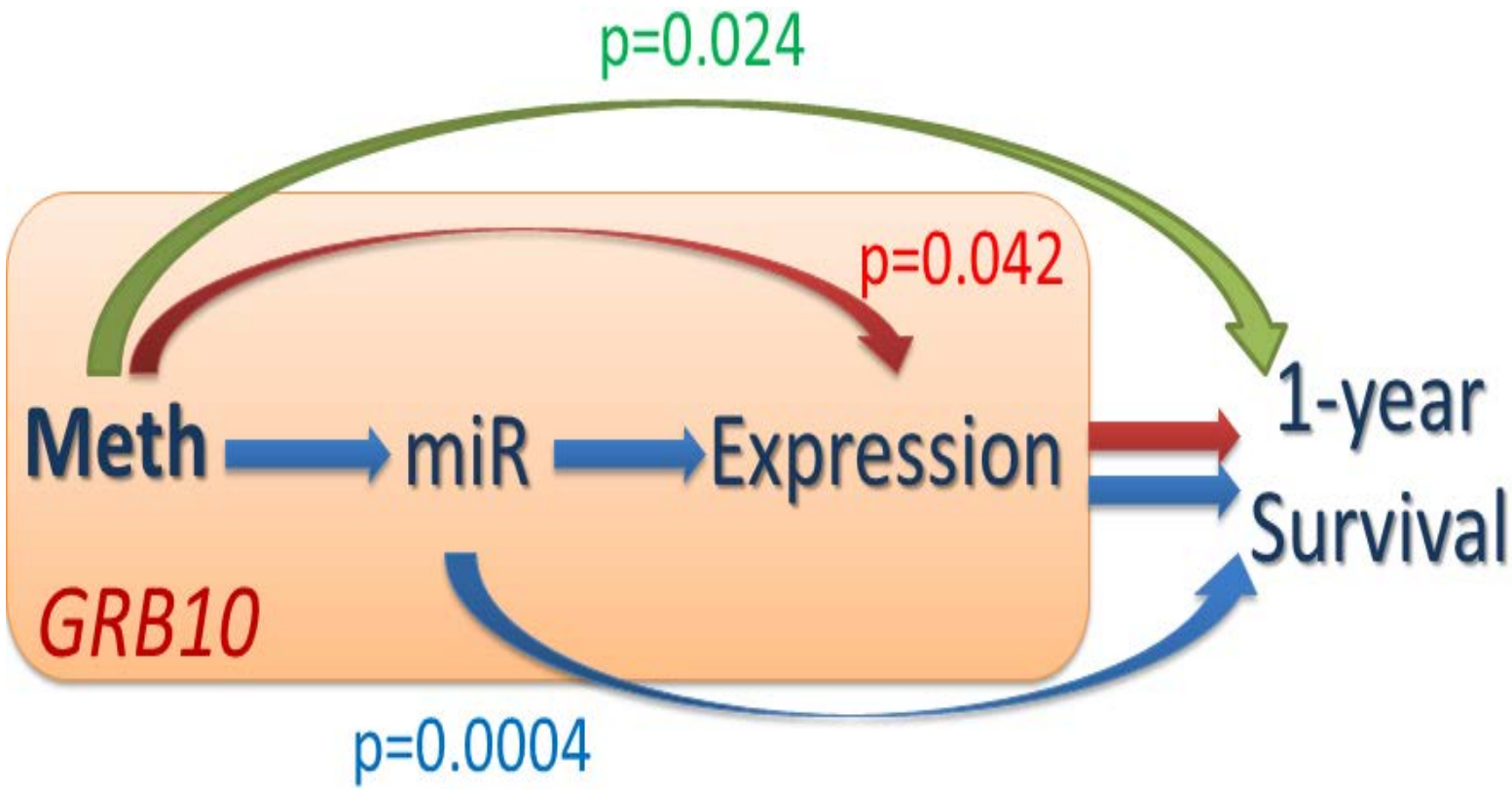


iGWAS results: Alternative Effect/Direct Effect (AE/IE)

- 0 gene with FDR<1%



Methylation, MicroRNA, Expression Effects on Survival of Glioblastoma Multiforme

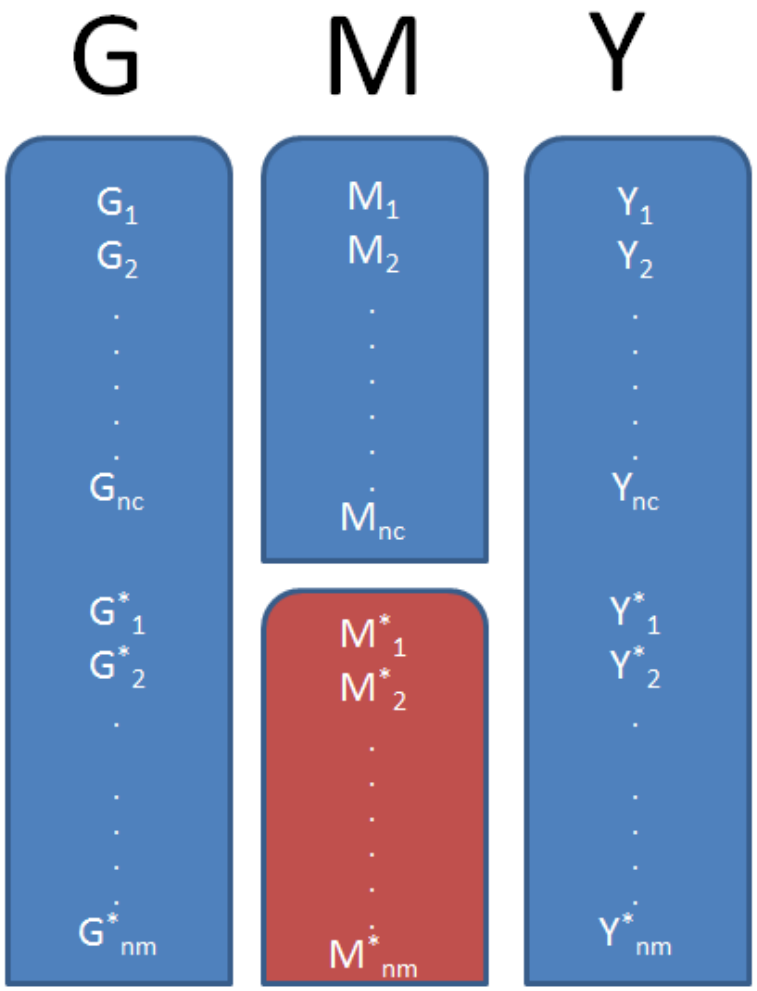


Outline

Mediation Analysis In the Presence of Missing Data on the Mediator

- Genomic data (DNA methylation or Gene expression) are available only for a subset of subjects in a GWAS study.
- Subjects with only GWAS and phenotype data still contribute information to direct and indirect effects.
- Objective: Perform medication analysis in the presence of missing data the of the mediator.

Data Break Down



Estimation

- Assume the mediator is missing at random.
- Use the EM algorithm to estimate model parameters.

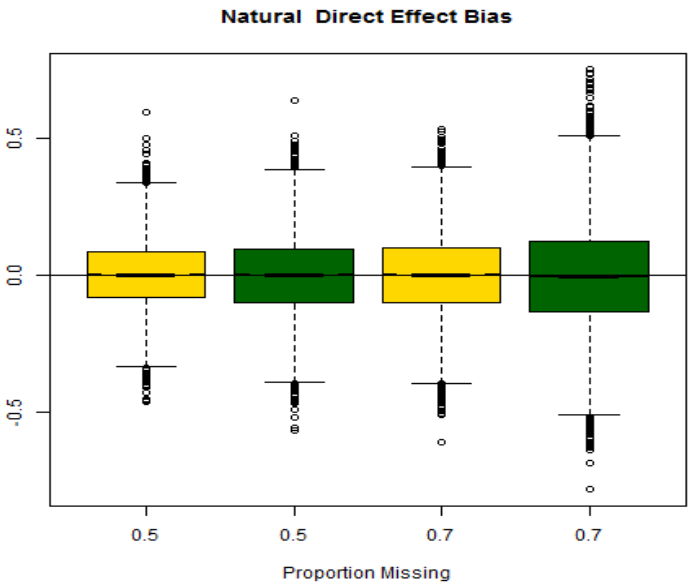
Likelihood and Estimation Using the EM Algorithm

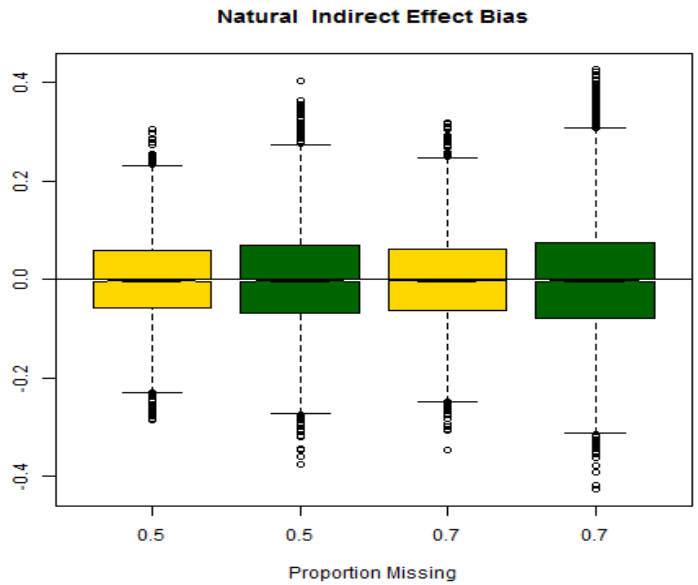
- Two types of individuals:
  - Those with complete data (n_c)
    - X_i, G_i, M_i, Y_i
  - Those with incomplete data(n_m)
    - X_i, G_i, Y_i
- The loglikelihood

$$\sum_{i=1}^{n_c} \ell_i(Y_i, M_i | X_i, G_i) + \sum_{j=1}^{n_m} \ell_j(Y_j | X_j, G_j)$$

- Estimation of model parameters proceed with the EM algorithm.

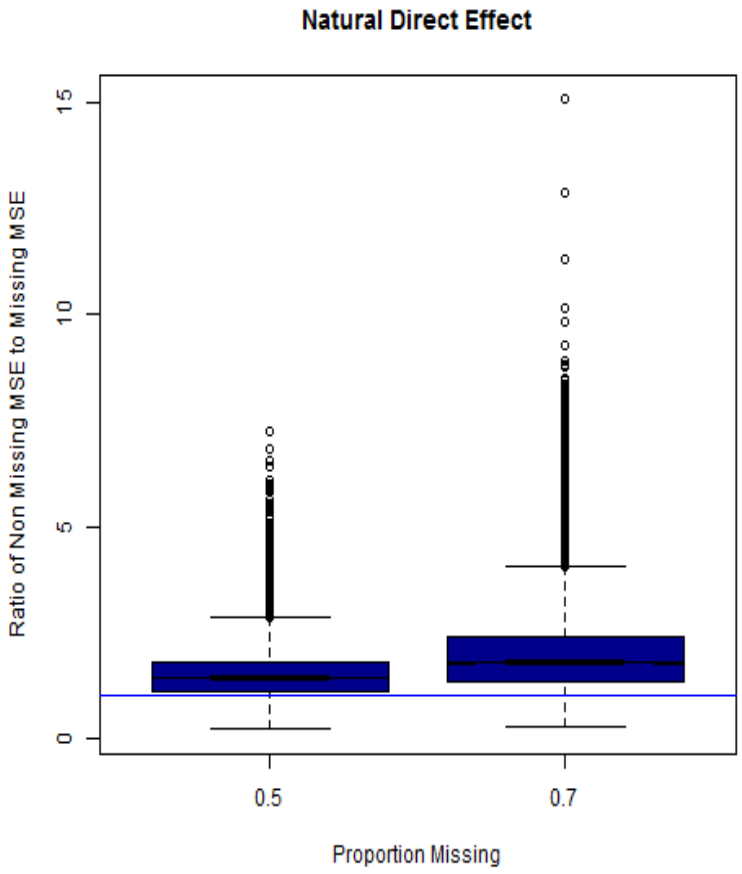
Simulation results: Unbiasness of point estimates

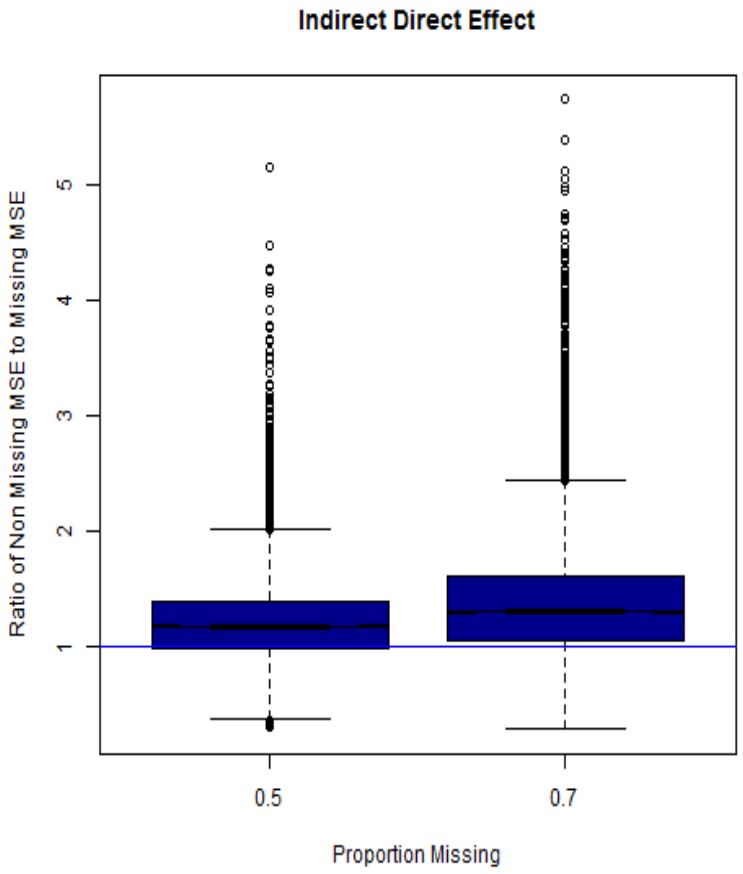




- Yellow represents using all data
- green represents using just individuals with complete data

Simulation results: Gain in efficiency





- MSE of estimates using complete data divided by MSE using all data. 10000 iterations

Discussions

- Causal mediation analysis provides an attractive framework for integrative analysis of genetic (SNPs), genomic (gene expressions/DNA methylation) and environmental data to understand the causal pathways.
- Mediation analysis for family data
- Accounting for family relatedness ensures correct inference
- Medication analysis when some of mediators are missing.
- Analysis is more challenging for discrete phenotypes

References

- VanderWeele, et al, American Journal of Epidemiology, 2012
- Huang, et al, Annals of Applied Statistics, 2014
- Huang, et al, Genetic Epidemiology, 2015.