

Finding SNP Associations with a Secondary Phenotype in Genetic Association Studies

Huilin Li

12th IMS New Researchers' Conference, July 30, 2009

Joint work with Mitchell H. Gail, Sonja Berndt and Nilanjan Chatterjee

BB,DCEG, National Cancer Institute

Background

	$G = 0$		$G = 1$		
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	Total
$D = 0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D = 1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

- Rare disease
- Gene and secondary phenotype are dichotomous

- $P(X = 1|G) = \frac{\exp(\beta_0 + \beta_1 G)}{1 + \exp(\beta_0 + \beta_1 G)}$

$$\exp(\beta_1) = \text{OR}_{GX}$$

Background

$$P(D = 1|G, X) \doteq \exp(\mu + \gamma_1 G + \gamma_2 X + \gamma_{12} GX)$$

- If $\gamma_{12} = 0$

$$OR_{GX}^{CO} = OR_{GX}^{CA} = OR_{GX}$$

We can use cases in the association estimation.

* Weighted method

* Maximum likelihood estimation method (MLE), Lin and Zeng (2008)

- If $\gamma_{12} \neq 0$

$$OR_{GX}^{CO} = OR_{GX} \text{ but } OR_{GX}^{CA} = OR_{GX} \exp(\gamma_{12})$$

Can cases be used?

Data example

Colorectal adenoma case-control study

	NAT2=0		NAT2=1		Total
	Smoking=0	Smoking=1	Smoking=0	Smoking=1	
$D = 0$	255	317	10	23	605
$D = 1$	199	380	18	13	610

Moslehi and others (2006)



Log Odds ratio estimates

	$\hat{\beta}_1$	s.d.
control only	0.615	0.39
case only	-0.972	0.27
weighted	-0.207	0.27
MLE	-0.172	0.26

* $\gamma_{12} = -1.58$ p-value=0.0032

Methods

Control only: $\hat{\beta}_{1CO} = \log\left(\frac{r_{000}r_{011}}{r_{001}r_{010}}\right)$, $\hat{\sigma}_{CO}^2 = \sum_{g=0}^1 \sum_{x=0}^1 (1/r_{0gx})$

Case only: $\hat{\beta}_{1CA} = \log\left(\frac{r_{100}r_{111}}{r_{101}r_{110}}\right)$, $\hat{\sigma}_{CA}^2 = \sum_{g=0}^1 \sum_{x=0}^1 (1/r_{1gx})$

Weighted: $\hat{\beta}_{1W} = w_{cc}\hat{\beta}_{1CO} + (1 - w_{cc})\hat{\beta}_{1CA}$, where $w_{cc} = \hat{\sigma}_{CA}^2 / (\hat{\sigma}_{CA}^2 + \hat{\sigma}_{CO}^2)$. $\hat{\sigma}_W^2 = \hat{\sigma}_{CA}^2 \hat{\sigma}_{CO}^2 / (\hat{\sigma}_{CA}^2 + \hat{\sigma}_{CO}^2)$

MLE: Lin and Zeng (2008)

$$\prod_{i=1}^n P(G_i, X_i | D_i)$$
$$= \prod_{i=1}^n \left\{ \frac{P(D_i=1|G_i, X_i)P(X_i|G_i)P(G_i)}{P(D_i=1)} \right\}^{D_i} \left\{ \frac{P(D_i=0|G_i, X_i)P(X_i|G_i)P(G_i)}{P(D_i=0)} \right\}^{1-D_i}$$

Adaptively weighted method

- $$\hat{\beta}_{1AW} = \frac{\hat{\gamma}_{12}^2}{(\hat{\sigma}_{CO}^2 + \hat{\gamma}_{12}^2)} \hat{\beta}_{1CO} + \frac{\hat{\sigma}_{CO}^2}{(\hat{\sigma}_{CO}^2 + \hat{\gamma}_{12}^2)} \hat{\beta}_{1W}$$

where $\hat{\gamma}_{12}^2 = (\hat{\beta}_{1CA} - \hat{\beta}_{1CO})^2$.

-

$$\widehat{Var}(\hat{\beta}^{AW}) = \hat{\sigma}_{CO}^2 \left[1 + \frac{\hat{\sigma}_{CO}^2 (1 - w_{cc}) \{(\hat{\beta}_{1CA} - \hat{\beta}_{1CO})^2 - \hat{\sigma}_{CO}^2\}}{\{\hat{\sigma}_{CO}^2 + (\hat{\beta}_{1CA} - \hat{\beta}_{1CO})^2\}^2} \right]^2 + \hat{\sigma}_{CA}^2 \left[\frac{\hat{\sigma}_{CO}^2 (1 - w_{cc}) \{(\hat{\beta}_{1CA} - \hat{\beta}_{1CO})^2 - \hat{\sigma}_{CO}^2\}}{\{\hat{\sigma}_{CO}^2 + (\hat{\beta}_{1CA} - \hat{\beta}_{1CO})^2\}^2} \right]^2.$$

Data example

Colorectal adenoma case-control study

	NAT2=0		NAT2=1		Total
	Smoking=0	Smoking=1	Smoking=0	Smoking=1	
$D = 0$	255	317	10	23	605
$D = 1$	199	380	18	13	610

Odds ratio estimates

	$\hat{\beta}_1$	s.d.
control only	0.615	0.39
Adaptively weighted	0.569	0.40
case only	-0.972	0.27
weighted	-0.207	0.27
MLE	-0.172	0.26

* $\gamma_{12} = -1.58$ p-value=0.0032

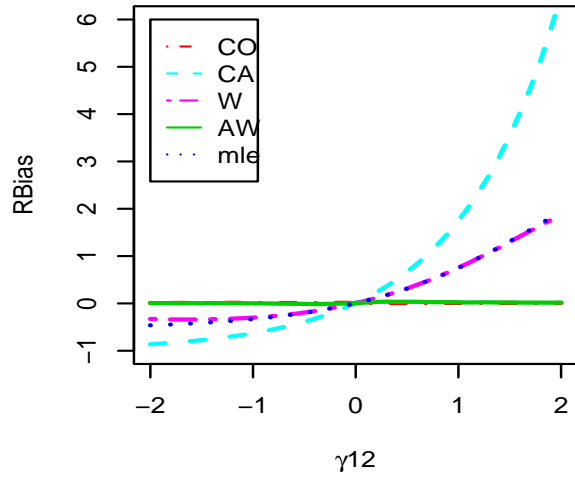
Simulations

- Application to a pre-selected SNP
 - * $P(G = 1) = P(X = 1) = 0.3$ and $OR_{GD} = OR_{XD} = 1$
 - * γ_{12} varies from -2 to 2
 - * Cell probability vectors: p_0 and p_1 can be determined
 - * r_0 and r_1 independently from the two multinomial distributions
 - * 10,000 simulated replications

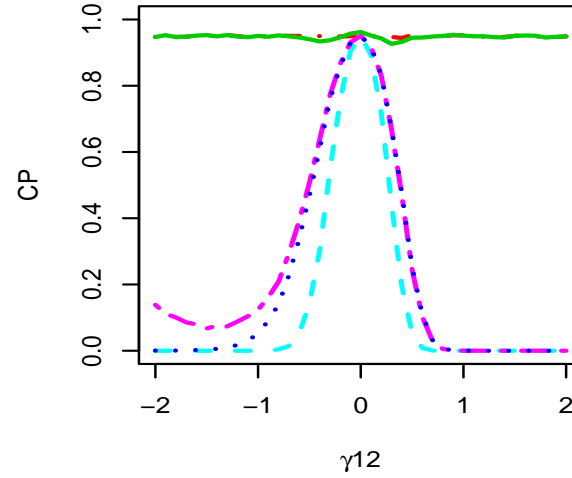
	$G = 0$		$G = 1$		
	$X = 0$	$X = 1$	$X = 0$	$X = 1$	Total
$D = 0$	r_{000}	r_{001}	r_{010}	r_{011}	n_0
$D = 1$	r_{100}	r_{101}	r_{110}	r_{111}	n_1

Single SNP estimation

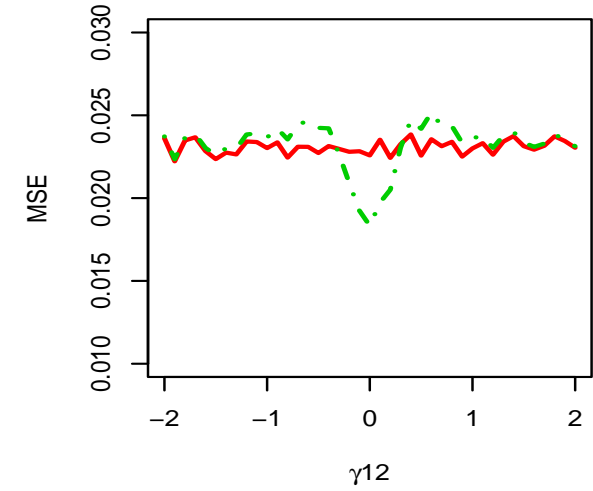
$\beta_1 = 0, n_1 = n_0 = 1000$



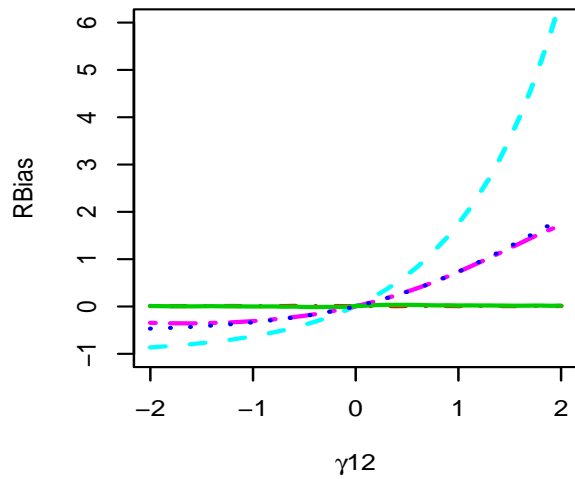
$\beta_1 = 0, n_1 = n_0 = 1000$



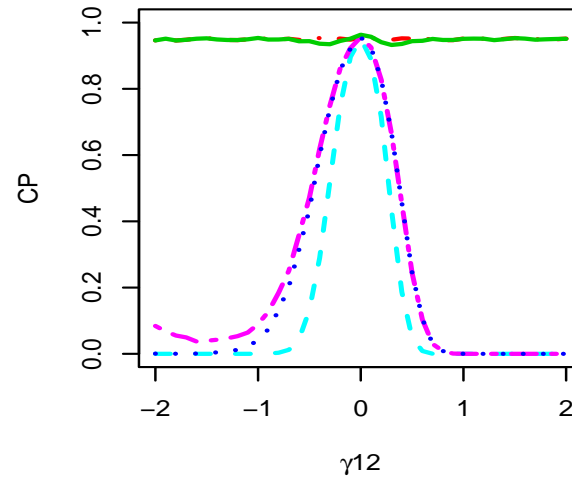
$\beta_1 = 0, n_1 = n_0 = 1000$



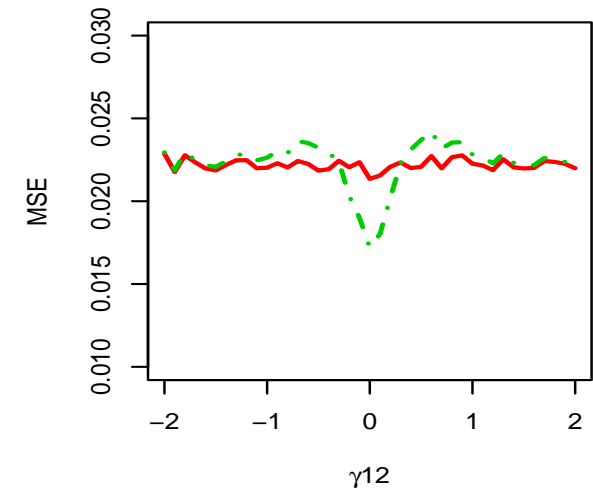
$\beta_1 = 0.25, n_1 = n_0 = 1000$



$\beta_1 = 0.25, n_1 = n_0 = 1000$

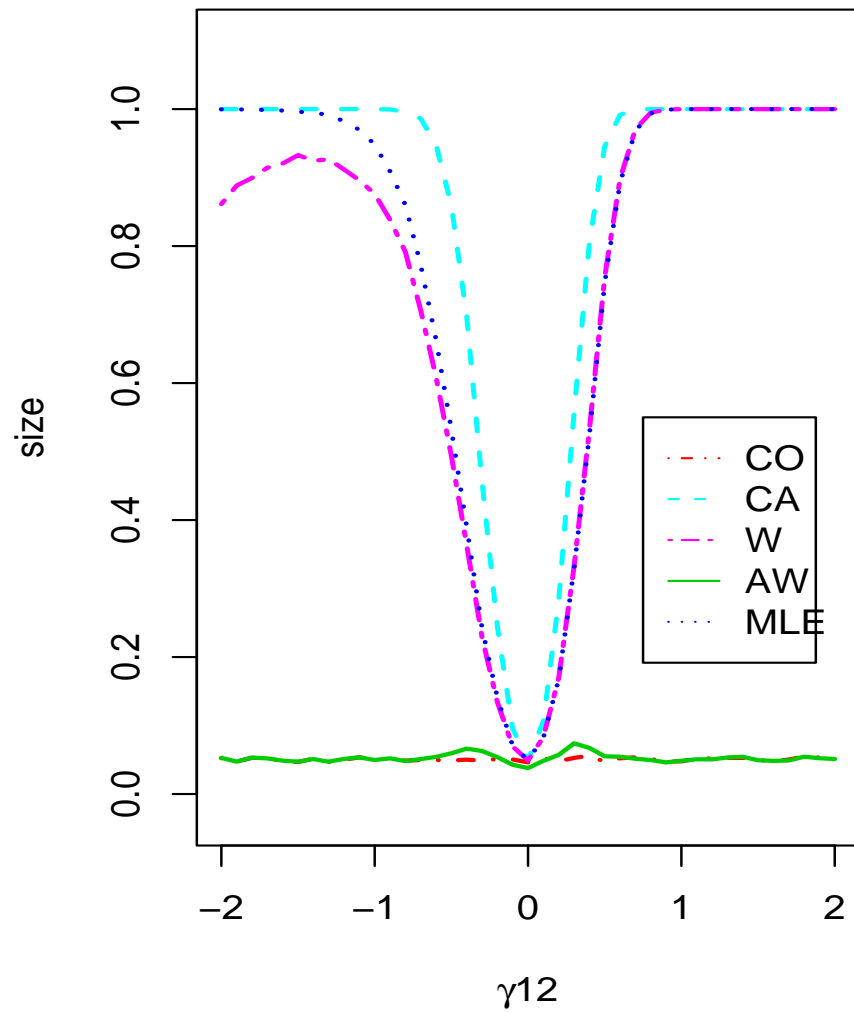


$\beta_1 = 0.25, n_1 = n_0 = 1000$

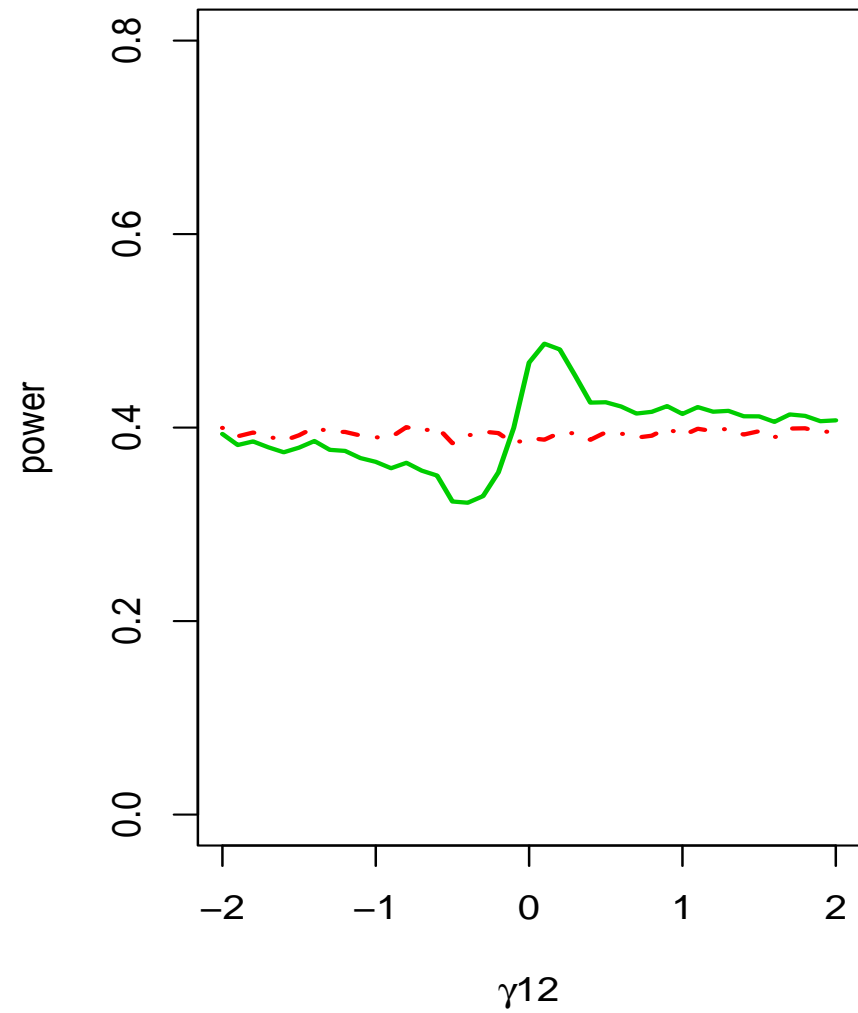


Selected SNP hypothesis testing

$\alpha=0.05$, $\beta_1 = 0$, $n_1=n_0=1000$



$\alpha=0.05$, $\beta_1 = 0.25$, $n_1=n_0=1000$



Asymptotic Genome-Wide Size and Power

- γ_{12} has a mixture distribution: 99% SNPs have $\gamma_{12} = 0$, and the other 1% are generated from $N(0, (\log(2)/2)^2)$
- $\alpha = 0.05/500,000 = 10^{-7}$

$n_0 = n_1$	$\beta_1 = 0$; Size			$\beta_1 = .25$; Power		
	1,000	5,000	10,000	1,000	5,000	10,000
Control only, W_{CO}	0.049	0.049	0.049	0.000	0.060	0.504
Adaptively weighted, W_{AW}	0.049	0.049	0.049	0.002	0.500	0.982
Case only, W_{CA}	1.000	1.000	1.000	0.001	0.064	0.505
Weighted, W_W	1.000	1.000	1.000	0.002	0.504	0.984
MLE, W_{MLE}	1.000	1.000	1.000	0.002	0.504	0.984

Conclusions

- If we know that $\gamma_{12} = 0$, MLE and weighted methods are both much more efficient than control only method.
- To guard against possibility $\gamma_{12} \neq 0$, use control only or adaptively weighted method
 - * For estimation for a pre-selected SNP, adaptively weighted method may reduce MSE compare to control only method
 - * For whole genome scan, the adaptively weighted method can gain major power over control only method by considering the likely scenario of gene-secondary phenotype interaction on the disease rate in the underlying population.

THANK YOU!