National Cancer Institute

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health

Exploiting Interactions for Enhanced Detection of Genetic and Environmental Risk-Factors for Complex Diseases

Nilanjan Chatterjee, PhD

Biostatistics Branch

Division of Cancer Epidemiology and Genetics

chattern@mail.nih.gov

Value of Assessing Statistical Interaction (Thompson, 1991)

- Understanding biology?
- **Enhanced detection of effects**
- Characterization of joint effects
- Targeting intervention

Outline

- Omnibus tests
- Strategies for improving power
- Selecting SNPs for replication following GWAS

Test for G in Presence of a Known Risk Factor E (G)

- G would be considered of “interest” if it is associated with D in any sub-group defined by E

- Null hypothesis of interest

$$\mathbf{H_0}: \beta_{\mathbf{G|E=0}} = 0 \quad \textbf{and} \quad \beta_{\mathbf{G|E=1}} = 0$$

- Alternatively

$$\mathbf{H_0}: \beta_{\mathbf{G|E=0}} = 0 \quad \textbf{and} \quad \theta = \beta_{\mathbf{G|E=1}} - \beta_{\mathbf{G|E=0}} = 0$$

  - Simultaneous test for main- and interaction- effect of G in a logistic model that includes a main effect of E

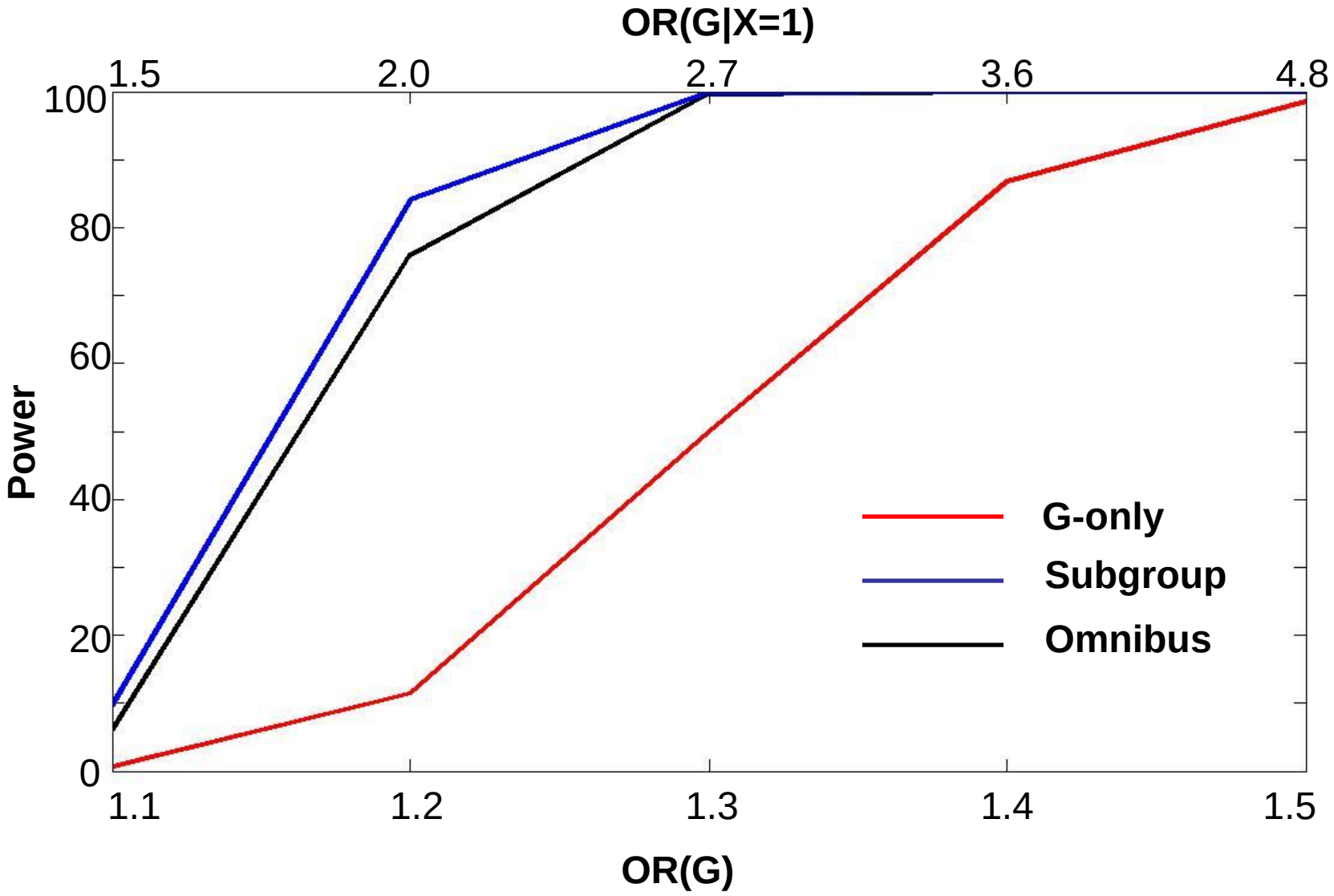
Three Tests for Detecting G

- G-only
  - $\beta_G^*=0$
  - 1 d.f
- Subgroup specific
  - $\beta_{G|E=1}=0$
  - 1 d.f
- Omnibus test
  - $\beta_{G|E=0}=0$ and $\beta_{G|E=1}=0$
  - 2 d.f

Effect of *NAT2* Acetylation and Smoking on Bladder Cancer (Garcia-Closas et al., Lancet, 2005)

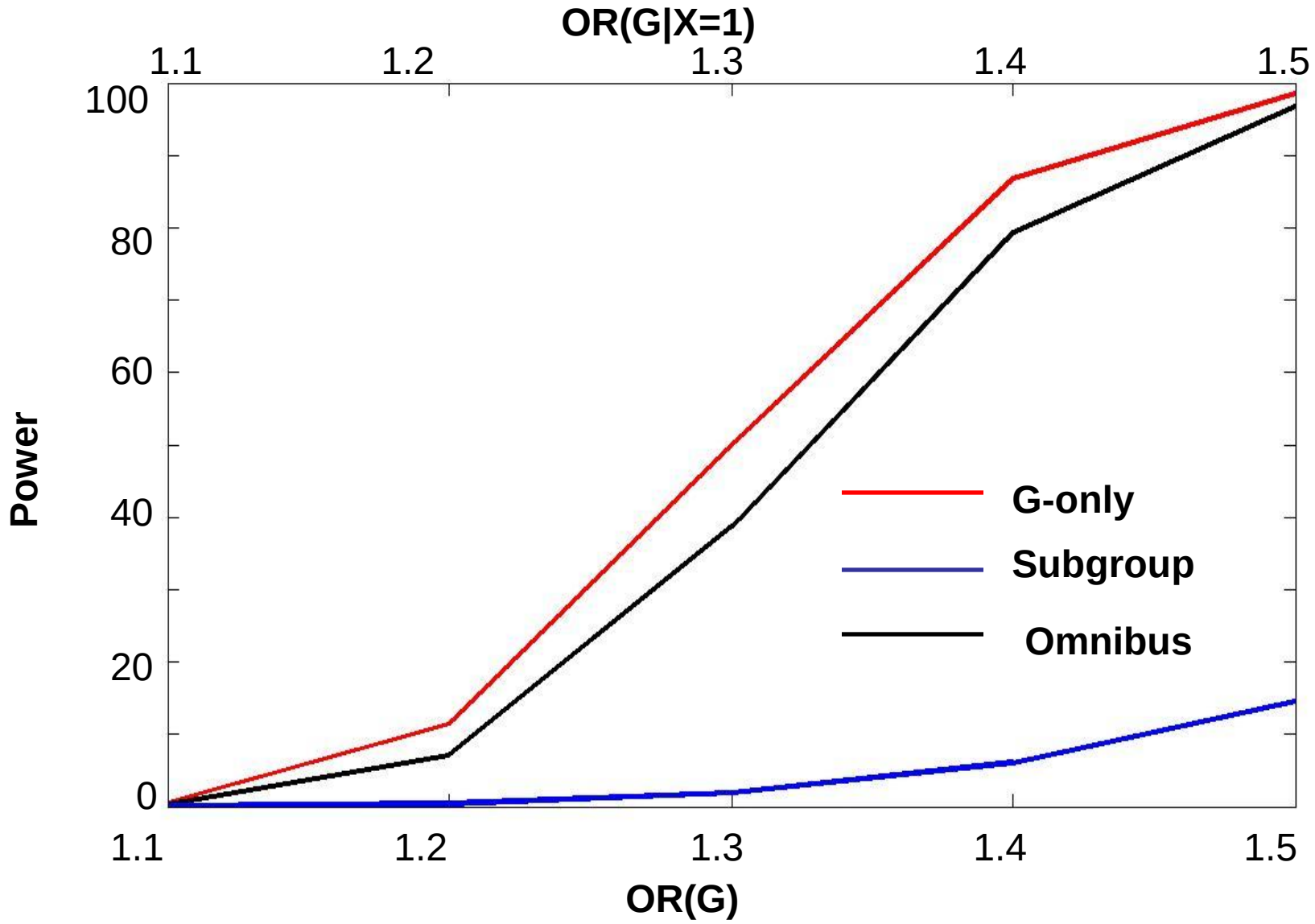
	Controls	Cases	OR	Chi-square (df)	P-value
Overall					
Rapid	493	406			
Slow	637	728	1.39	14.44 (1)	$1.45£\ 10^{-4}$
Non-smokers					
Rapid	131	66			
Slow	199	91	0.91	0.24 (1)	$6.23£\ 10^{-1}$
Smokers					
Rapid	362	340			
Slow	438	637	1.55	20.01 (1)	$7.72£\ 10^{-6}$
Omnibus				20.52(2)	$4..01£\ 10^{-5}$

OR(G|X=0)=1.0



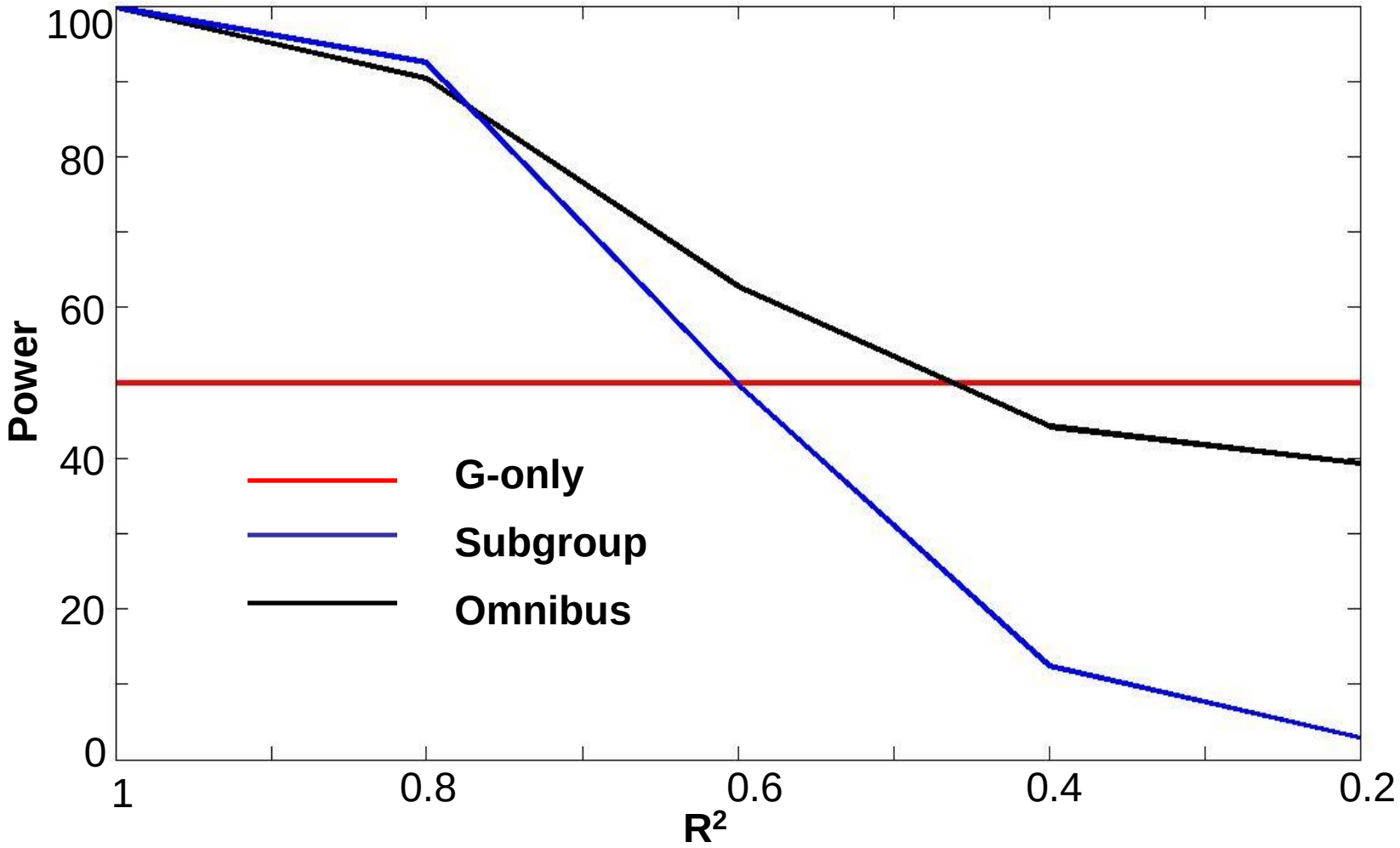
α=0.0001,P(G=1)=0.3,P(X=1)=0.2, OR(X)=1.3

OR(G|X=0)=OR(G|X=1)=OR(G)



α=0.0001,P(G=1)=0.3,P(X=1)=0.2, **OR(G|X=1)=OR(G|X=0)=OR(G),**OR(X)=1.3

OR(G|X=0)=1, but X is Misclassified



$\alpha=10^{-4}$,P(G=1)=0.3,P(X=1)=0.2, OR(X)=1.3, **OR(G)=1.3**, **OR(G|X=1)=2.7**

Increasing Power

- Power of omnibus test can be improved by increasing the precision of the interaction parameter

- Strategies for efficiency gain

  - Stratified sampling
    - If E is already available in a cohort, one can collect G on a case-control sample selected based on E

  - Reducing d.f.
    - Chatterjee et al., AJHG, 2006
    - Chapman and Clayton, Genetic Epi, 2007

  - **Exploiting assumption of G-E independence**

Exploiting Independence: The Case-Only Estimator and Extensions

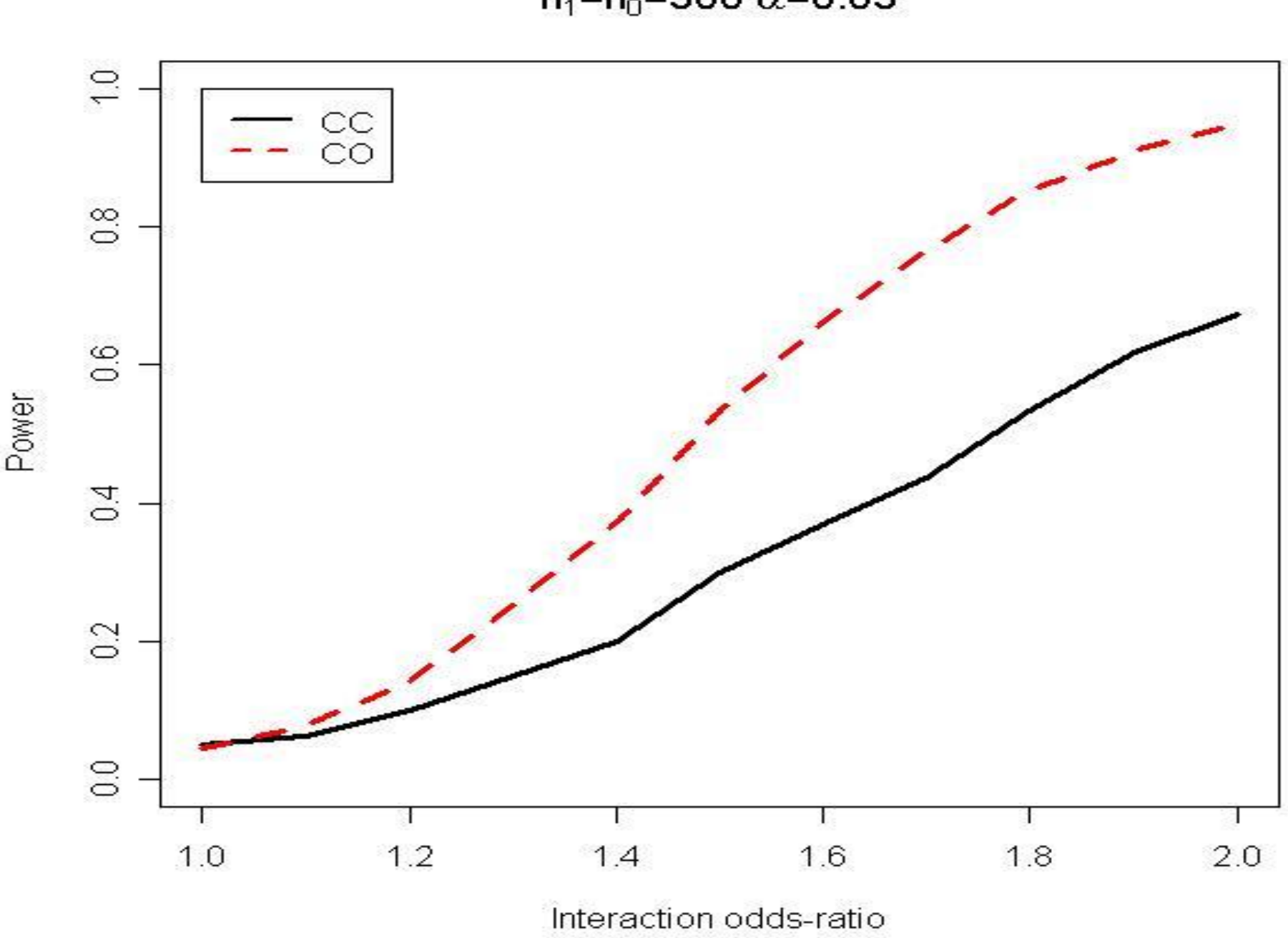
- Piegorsch et al., Stat Med, 1994

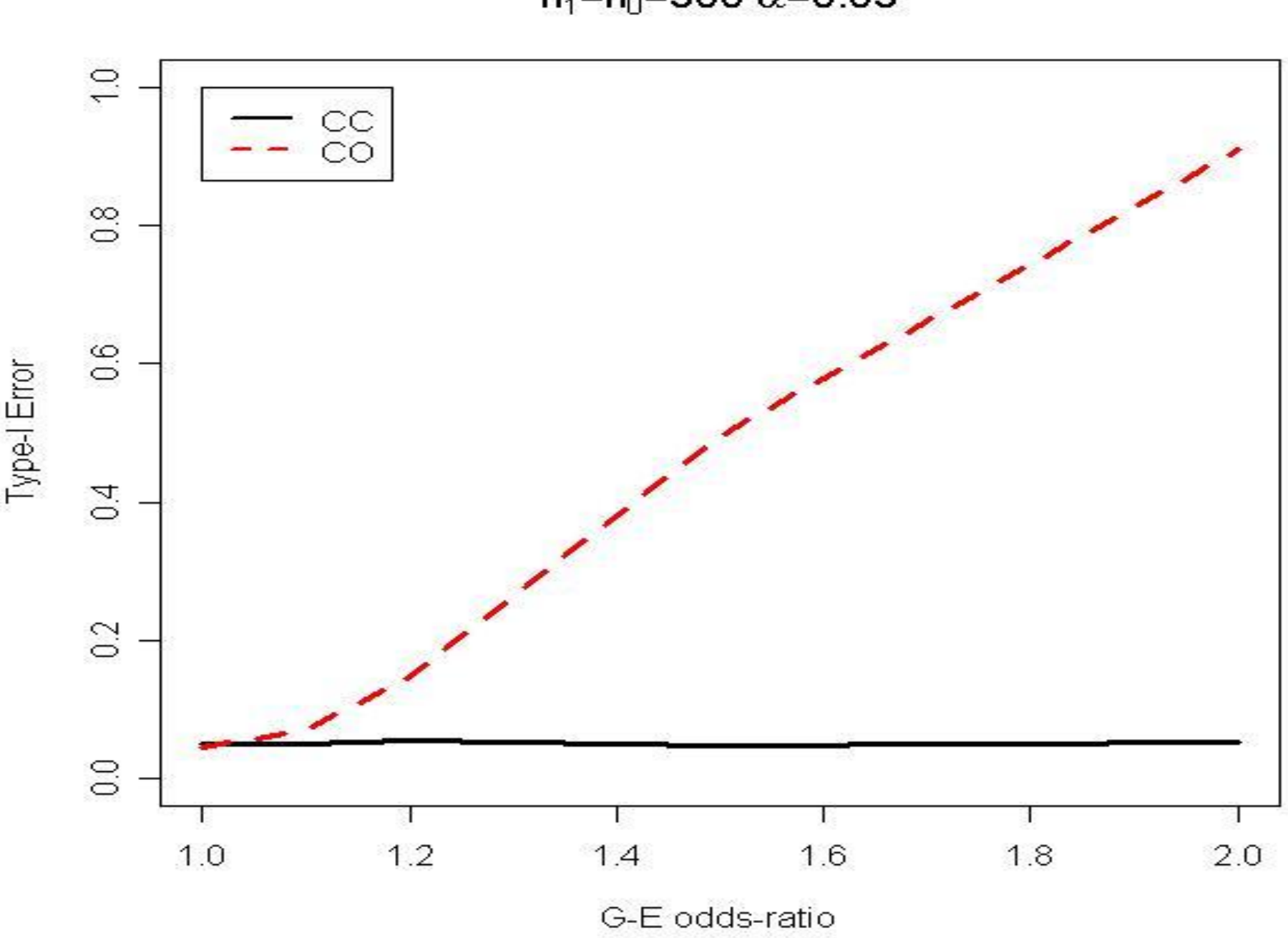
$$\frac{\mathrm{OR}(\mathrm{G},\mathrm{E}|\mathrm{D}=1)}{\mathrm{OR}(\mathrm{G},\mathrm{E}|\mathrm{D}=0)} \approx \mathrm{OR}(\mathrm{G},\mathrm{E}|\mathrm{D}=1)$$

  - More efficient than that obtained from logistic regression analysis

- Inference for a general logistic regression model under the independence assumption
  - Umbach and Weinberg, *Stat Med* 1997; Chatterjee and Carroll, *Biometrika* 2005;

- Sensitivity to independence assumption


$n_1=n_0=500$ $\alpha=0.05$
CC
CO
Power
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
Interaction odds-ratio


$n_1=n_0=500$ $\alpha=0.05$
CC
CO
Type-I Error
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
G-E odds-ratio

EB Estimator

(Mukherjee and Chatterjee, Biometrics, 2008)

$$\hat{\beta}^{EB} = \frac{\hat{\sigma}^2_{CC}}{(\hat{\tau}^2 + \hat{\sigma}^2_{CC})}\hat{\beta}^{CO} + \frac{\hat{\tau}^2}{(\hat{\tau}^2 + \hat{\sigma}^2_{CC})}\hat{\beta}^{CC}.$$

where

$$\begin{aligned}
\hat{\beta}_{CC} &= \text{Case-control estimator} \\
\hat{\beta}_{CO} &= \text{Case-only estimator} \\
\hat{\tau} &= \hat{\beta}_{CC} - \hat{\beta}_{CO} \\
\hat{\sigma}^2_{CC} &= \text{Var}(\hat{\beta}_{CC})
\end{aligned}$$

Variance Estimation

$$\widehat{V}_A(\hat{\beta}_{EB}) \approx \hat{\sigma}^2_{CO} + \left(\frac{\hat{\tau} 2(\hat{\tau}^2 + 3\hat{\sigma}^2_{CC})}{(\hat{\sigma}^2_{CC} + \hat{\tau}^2)^2} \right)^2 \hat{\sigma}^2_{\hat{\tau}}.$$

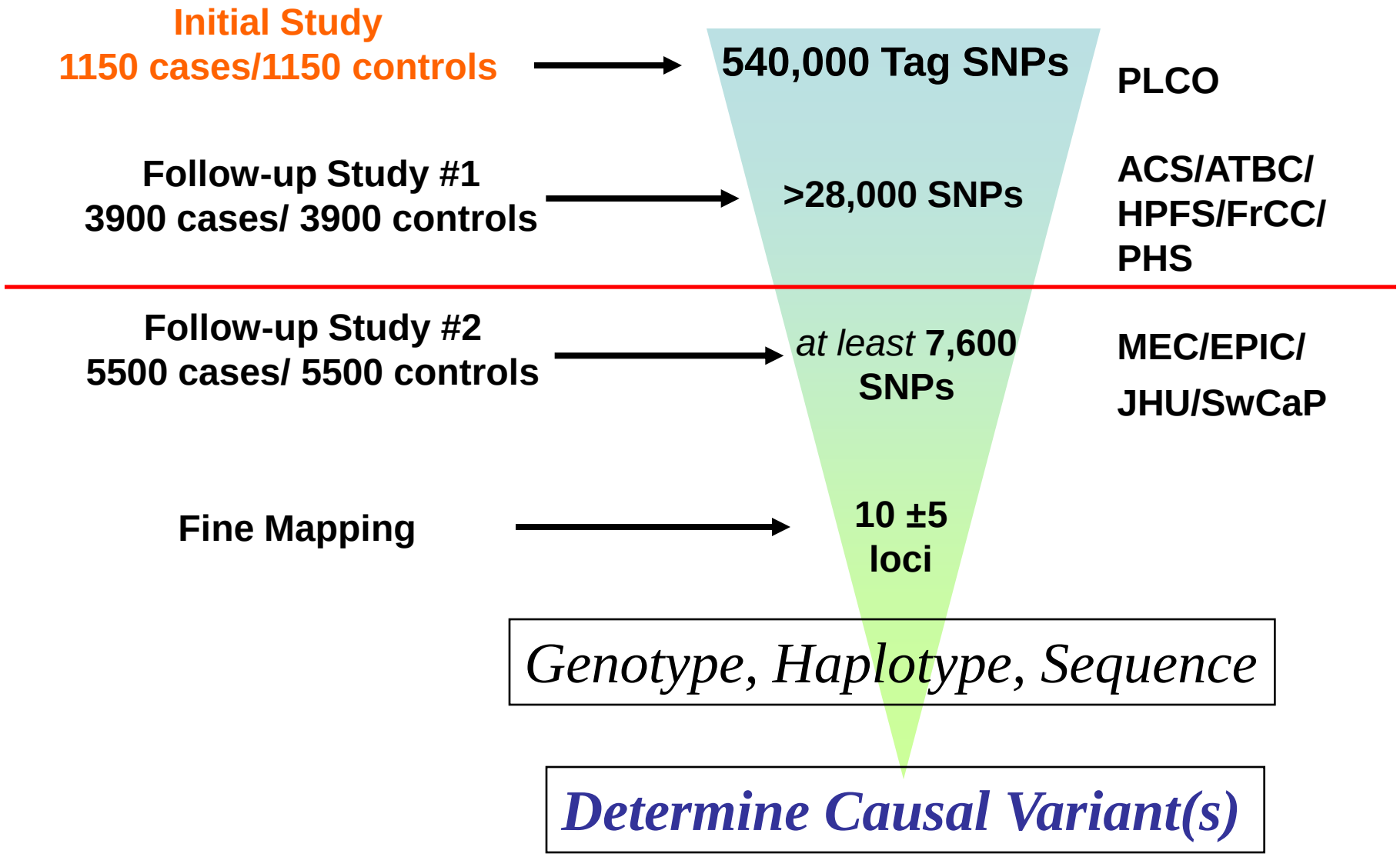
Mukherjee and Chatterjee, Biometrics, 2008
Chen, Chatterjee and Carroll, Submitted

Integrated Type-I Error/Power

		Case-Control	Case-Only	Two-stage	EB
®=0.05, N1=N0=500	Type-I Error	0.050	0.070	0.072	0.042
	Power (MI=1.5)	0.289	0.528	0.522	0.408
®=0.005, N1=N0=1000	Type-I Error	0.004	0.021	0.013	0.004
	Power (MI=1.5)	0.204	0.524	0.510	0.356

Mukherjee et al., Genetic Epidemiol, 2008

General Strategy for Prostate GWAS



Conditional Search

- Searching for association conditional on known genetic or/and environmental risk factors of a disease

- Conditioning factors
  - Known (or strongly suspected) candidate genes
  - Initial hits from a GWAS
  - Established environmental risk-factors such as smoking

Search for Susceptibility SNPs for PrCA conditional on “Confirmed” Genes

“Confirmed Genes”

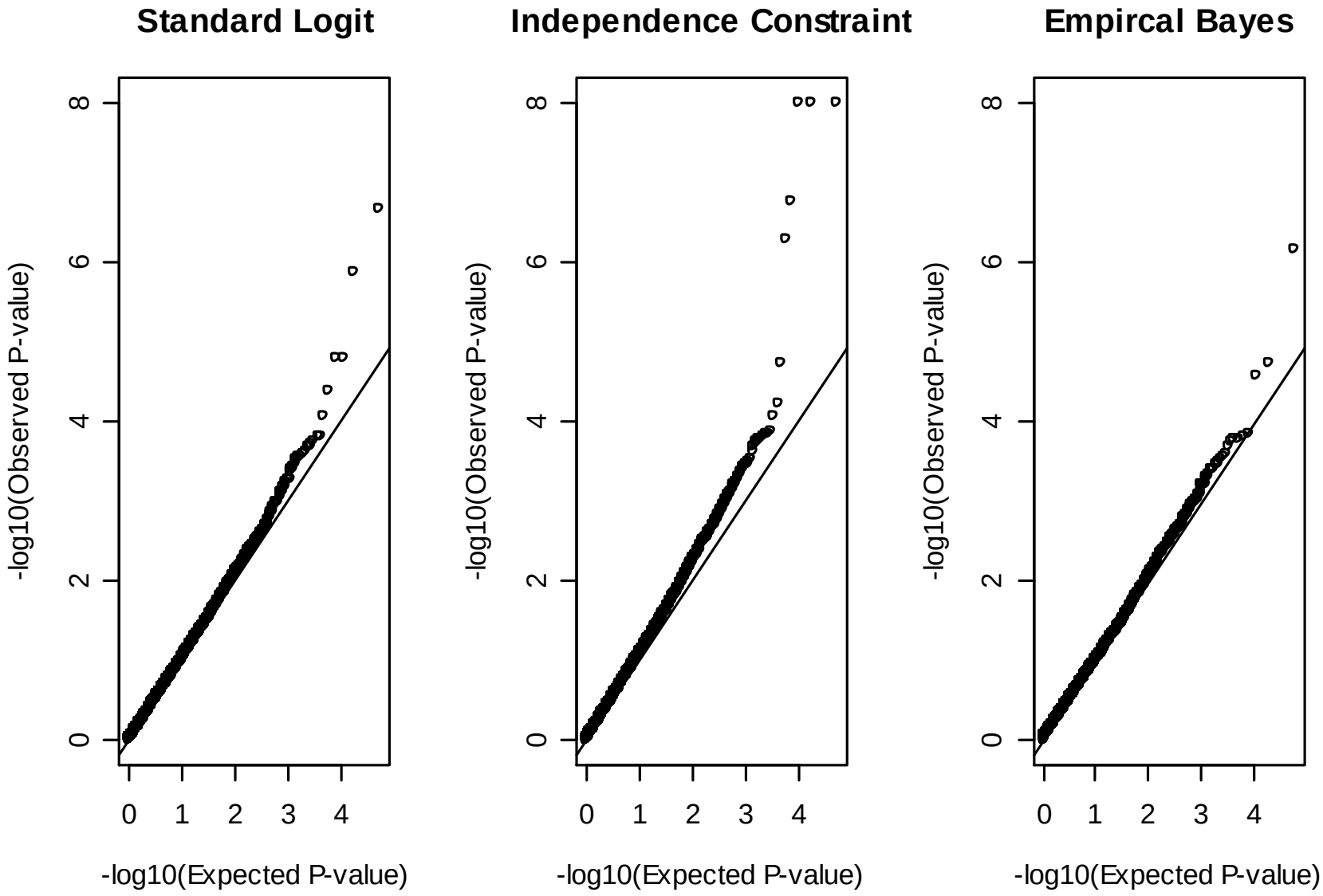
Gene/Region, Chr	Near or In Gene?	Biology
8q24, 8	Neither	
CTBP2 10	In	Two protein products; One is a transcription repressor; Associated with decreased PTEN (tumor suppressor)
DAB2IP 9	In	Tumor suppressor gene; Inactivated in multiple cancers; Association seen with aggressive cases
EHBP1, 2	In	Endocytic trafficking
HNF1B,17	In	Transcription factor; Marker for epithelial ovarian cancer
JAZF1 7	In	Zinc finger protein is transcription repressor; Associated with endometrial stromal tumors
KLK-2&3, 19	Near	Serine proteases; Strong association with PSA levels
MSMB 10	Near	Immunoglobulin binding protein; Synthesized by prostate epithelial cells
MYEOV 11	Near	Normal levels barely detectable; Over-expressed in cancers (myeloma)

7 associated loci in CGEMS Prostate Cancer

Region	p-value	Risk Allele Freq.	Odds ratios	
			Heterozygotes	Homozygotes
8q24 (loc1)	$6.7\ 10^{-16}$	0.1	1.49 (1.34-1.64)	1.83 (1.32-2.53)
10q11	$8.7\ 10^{-14}$	0.38	1.20 (1.10-1.31)	1.61 (1.42-1.81)
8q24 (loc2)	$4.7\ 10^{-13}$	0.50	1.13 (1.02-1.26)	1.46 (1.30-1.64)
17q21	$1.5\ 10^{-10}$	0.52	1.25 (1.13-1.34)	1.47 (1.31-1.65)
11q13	$4.1\ 10^{-10}$	0.50	1.18 (1.08-1.28)	1.48 (1.27-1.74)
10q26	$1.7\ 10^{-7}$	0.25	1.14 (0.94-1.38)	1.40 (1.16-1.69)
7p15	$3.2\ 10^{-7}$	0.76	1.18 (1.07-1.31)	1.54 (1.37-1.73)

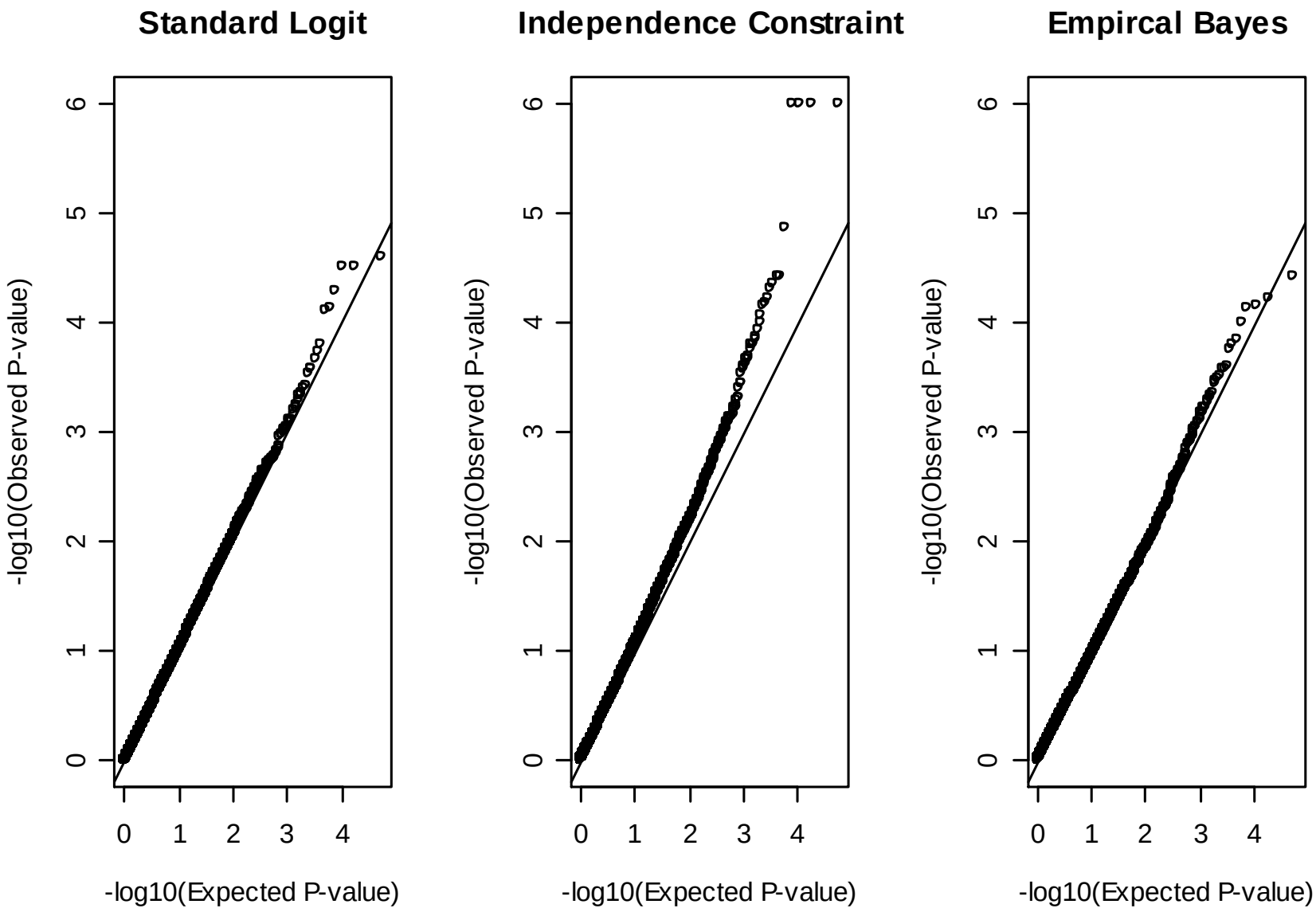
MSMB: Omnibus Wald Test

Results exclude SNPs within 500k base pairs of MSMB locus



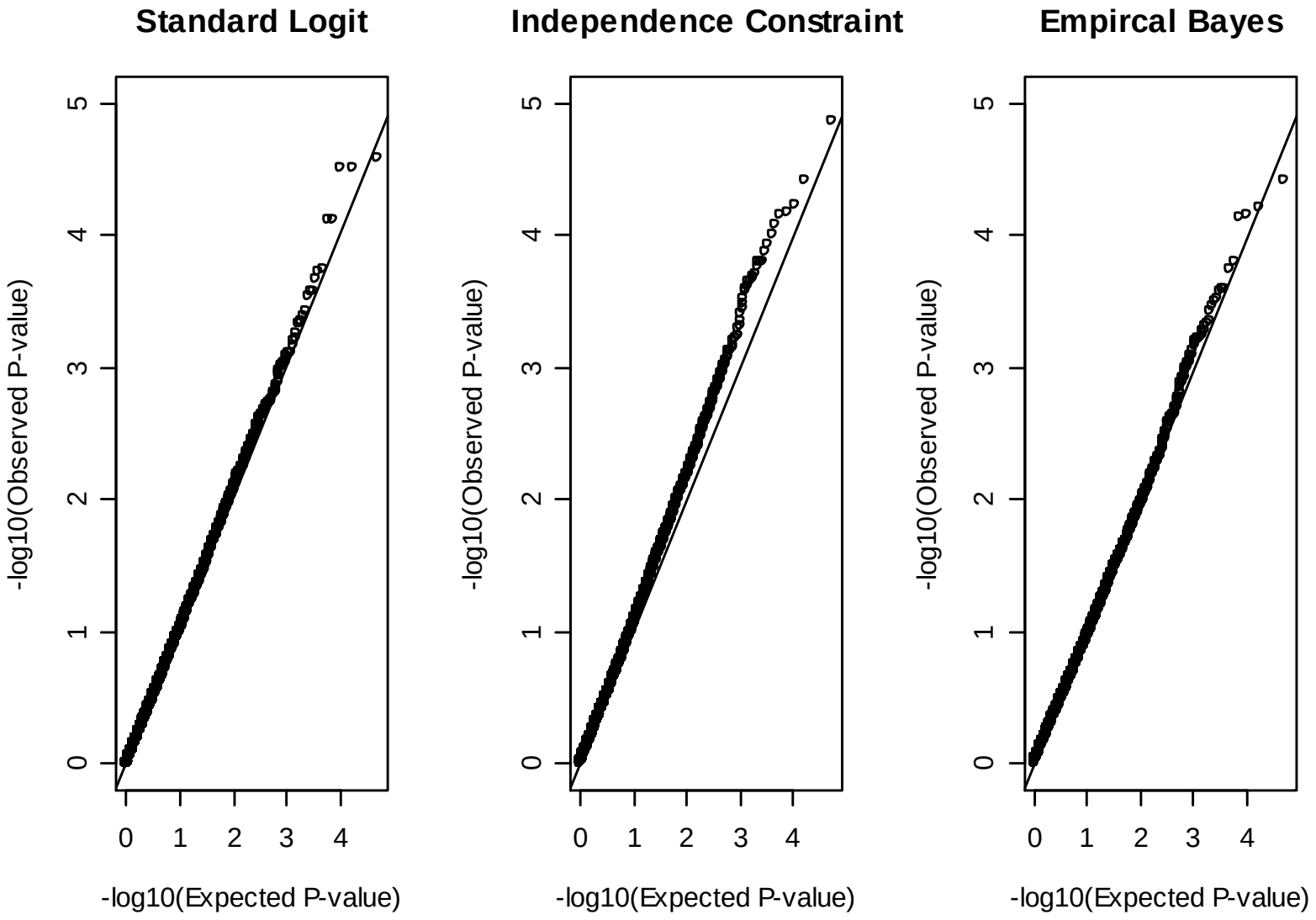
MSMB: Wald Test for Interaction

Results exclude SNPs within 500k base pairs of MSMB locus



MSMB: Wald Test for Interaction

Results exclude SNPs on MSMB chromosome



Summary Statistics

Genome Scan	**Wald Test P-values**
Conditional	Omnibus, Interaction
Main Effects	Marginal

- "Interaction Hit" criteria
  - Omnibus p-value ≤1.0E-3
  - Marginal p-value ≥1.0E-2

Scientific Results I

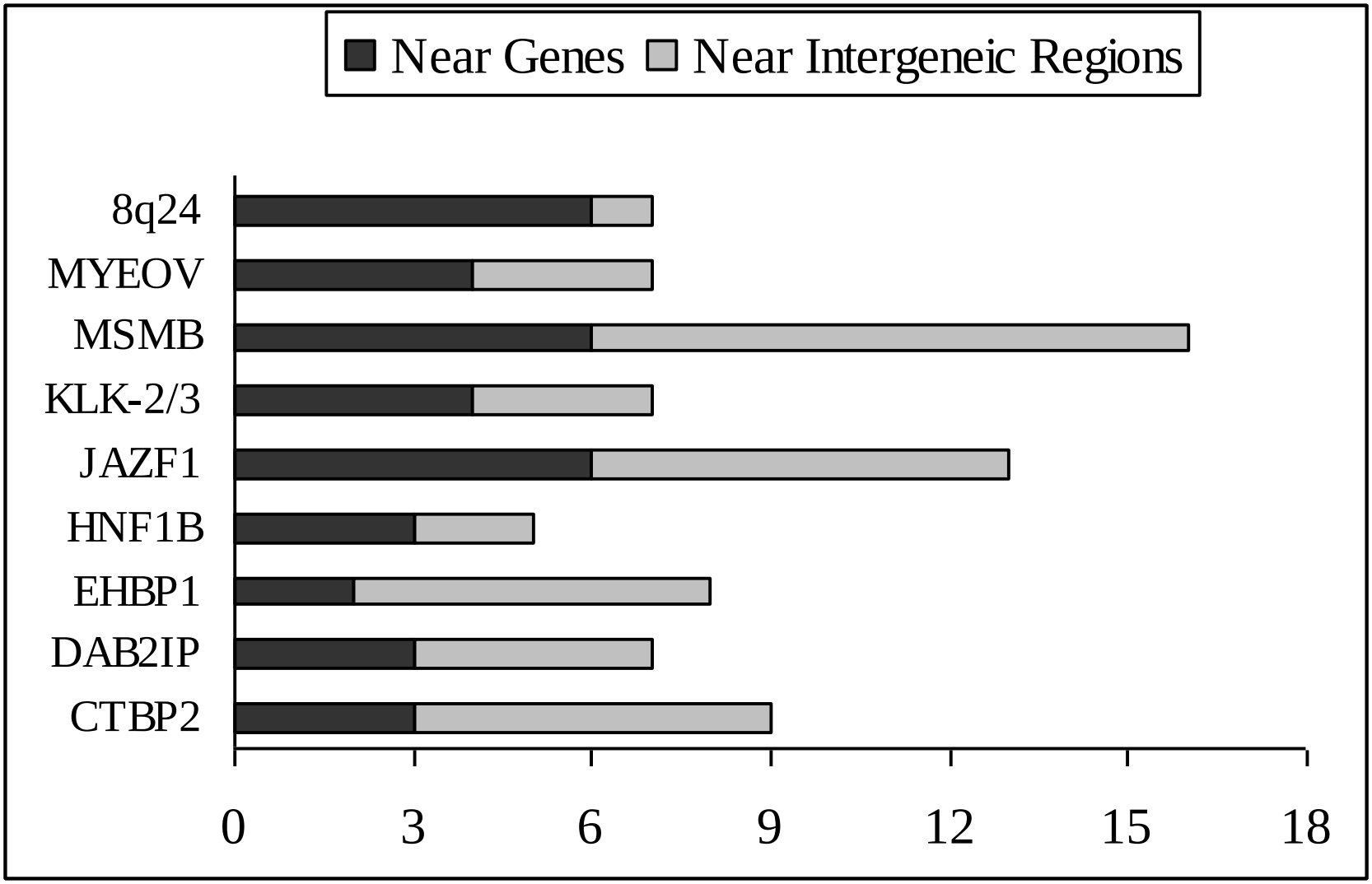


Table 1: Interaction hits identified through nine conditional genome scans of ~27k SNPs.

Summary

- So far the world looks very flat
  - multiplicative/additive
- Possible reasons
  - The world is multiplicative
  - Sample size is not large enough and effects are modest
  - Not accounting for more complex interactions
- Simple approaches to exploring interaction using pathways and network information is needed
- Replication is must

Summary

- Incorporating interaction into test of association can substantially improve power of detecting underlying risk-factors with non-multiplicative effects, but
- Tests need to be carefully constructed so that they have robust power under multiplicative effects
- Low R^2 between the measured and causal factors can negate advantage of interaction-based tests
  - Effects “look” close to multiplicative
- Exploiting natural assumptions of gene-gene and gene-environment independence can give a big boost in power
  - Caution is needed to protect against large-scale false positives
  - EB is a promising solution

Thanks

- **CGEMS team**
- **Raymond Carroll, *Texas A&M University***
- **Julia Ciampa, *NCI and Oxford***
- **Yi-Hau Chen, *Academia Sinica***
- **Bhramar Mukherjee, *University of Michigaan***
- **Sholom Wacholder, *NCI***
- **Bill Wheeler, *Information Management System***

Why Model Multiplicative Interaction?

- Under multiplicative model there is no benefit of using E to study G and vice versa (assuming G-E independence)

$$
\begin{aligned}
L &= \prod_{dge} p_{dge}^{n_{dge}} \\
&= \prod_{dg} p_{dg+}^{n_{dg+}} \times \prod_{de} p_{d+e}^{n_{d+e}}
\end{aligned}
$$

- Dupis et al, *Genetics* 1995

Type-I Error/Power

	OR_{GE}	Case-Control	Case-Only	Two-stage	EB
Type-I Error	1.0	0.05	0.05	0.07	0.04
	1.1	0.05	0.08	0.09	0.05
	1.2	0.05	0.14	0.15	0.07
	1.5	0.04	0.50	0.28	0.08
	2.0	0.05	0.91	0.11	0.06
Power (MI=1.5)	1.0	0.29	0.53	0.52	0.41
	1.1	0.30	0.70	0.66	0.50
	1.2	0.29	0.84	0.72	0.51
	1.5	0.29	0.98	0.54	0.45
	2.0	0.30	1.00	0.32	0.40

Mukherjee et al., Genetic Epidemiol, In revision

Setting - I

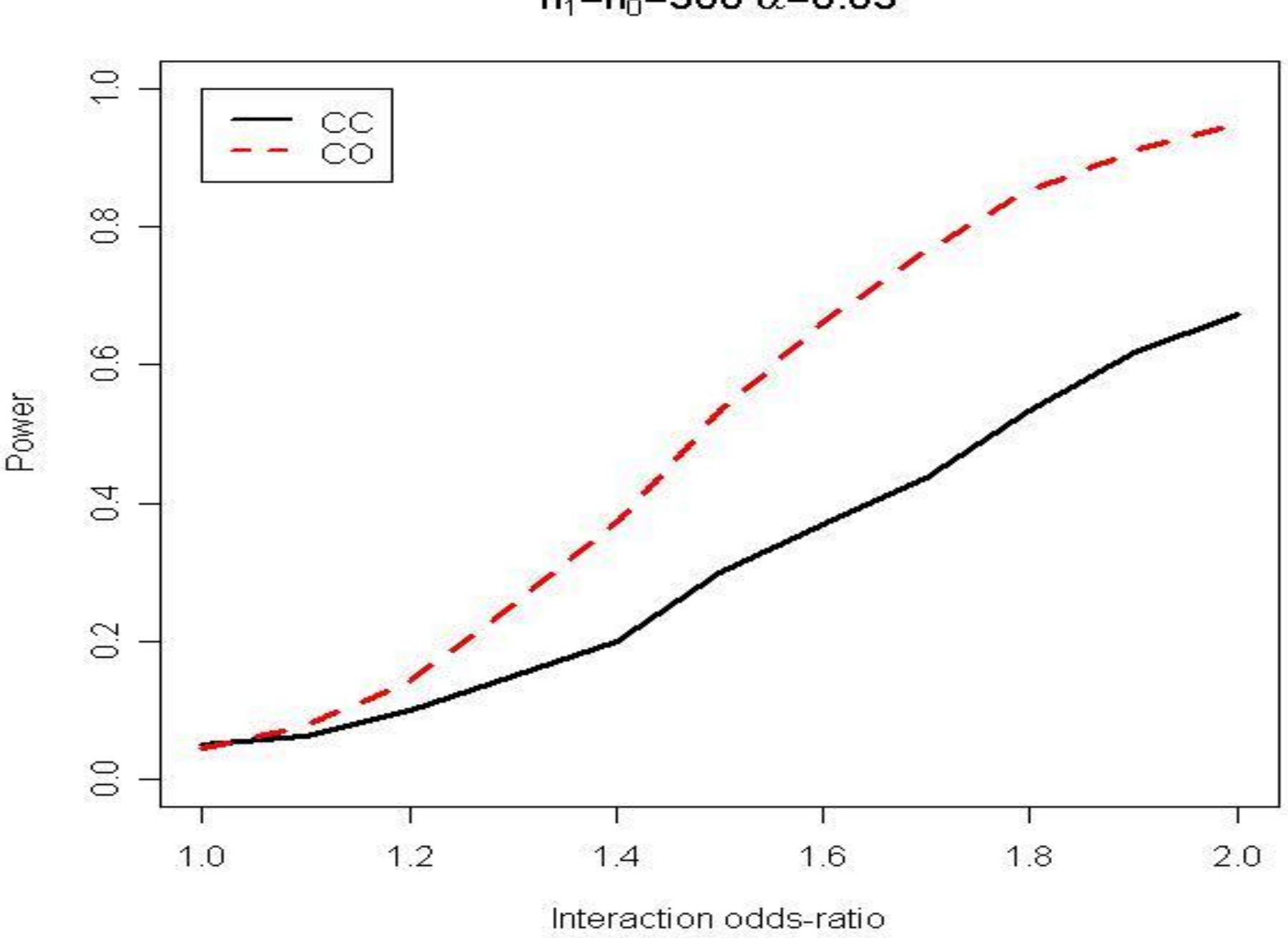
- Pr(G=1)=0.3
- Pr(E=1)=0.3
- OR_{GE}= Odds-ratio between G and E in disease-free subjects
- $N_1=N_0=500$
- ®=0.05
- Power evaluate at the alternative MI=1.5

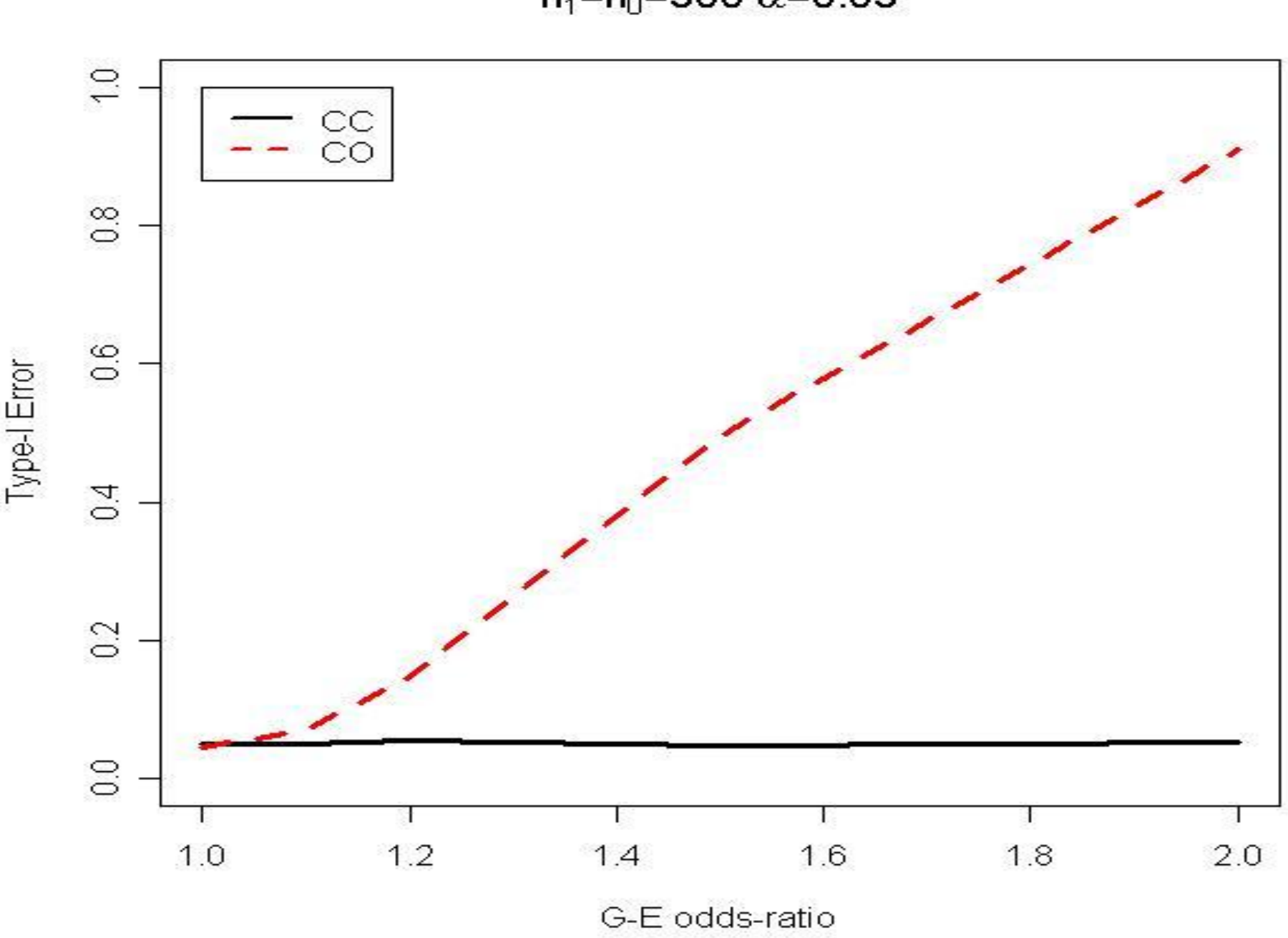
Setting - II

- Large scale association studies involve many possible G-E combinations
  - Independence assumption will be satisfied for most
  - but not all

- Assume
  - OR_{GE}=1 for 80% of the combinations
  - Distributed as LN(0,{log(1.5)/2}^{2}) for the rest

- Evaluate average Type-I error/Power


$n_1=n_0=500$ $\alpha=0.05$
CC
CO
Power
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
Interaction odds-ratio

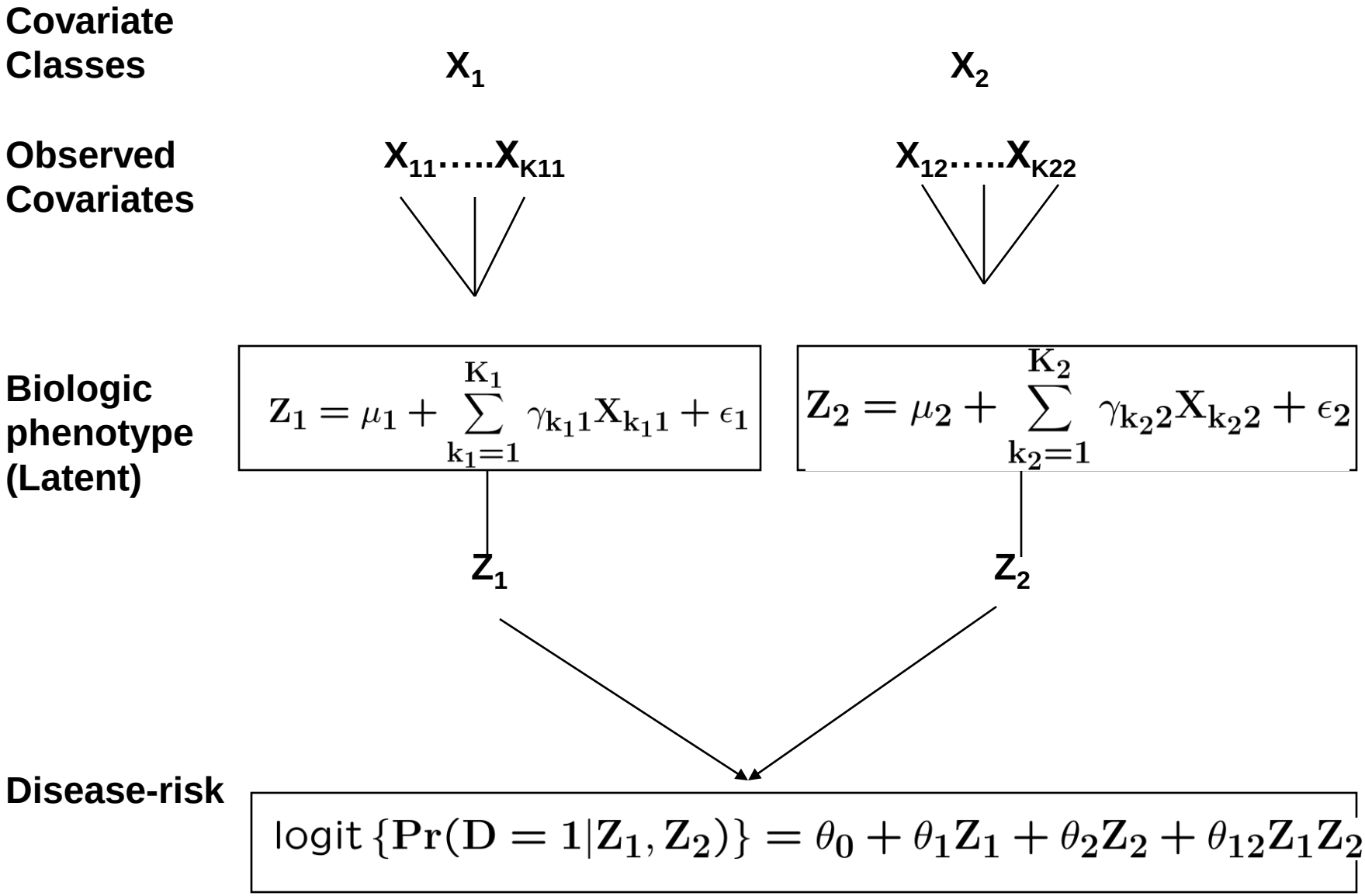

$n_1=n_0=500$ $\alpha=0.05$
CC
CO
Type-I Error
0.0
0.2
0.4
0.6
0.8
1.0
1.0
1.2
1.4
1.6
1.8
2.0
G-E odds-ratio

Operationally…

- CGEMS conditional Scan based on 1 d.f model for interaction for 8q24
    - Multiple (up to seven reported) susceptibility SNPs in the same region
    - Define a score for the 8q24 region based on the linear predictor from a logistic regression fit that only includes the main effects of the susceptibility SNPs
    - Model interaction of each SNP in the genome with the 8q24 score instead of the individual SNPs
- Asymptotic null distribution is non-standard, but can be generated using simple re-sampling method
- Permutation-based re-sampling can be also used under the assumption of G-E independence

Reducing degrees-of-freedom

A Conceptual Framework



Tests of Association in Tukey's model

$$\textbf{logit}\,\{\Pr(D=1|X_1,X_2)\}$$
$$\approx\ \alpha+\sum_{k_1=1}^{K_1}\beta_{1k_1}X_{1k_1}+\sum_{k_2=1}^{K_2}\beta_{2k_2}X_{2k_2}+\theta\sum_{k_1=1}^{K_1}\sum_{k_2=1}^{K_2}\beta_{1k_1}\beta_{2k_2}X_{1k_1}X_{2k_2},$$

$$\mathbf{H_{01}}:\beta_1\equiv(\beta_{11},\beta_{12},\ldots,\beta_{1K_1})=\mathbf{0}$$

- Captures both main and interaction effects
- Score test
  - Chatterjee et al., AJHG, 2006
  - Chapman and Clayton, Genetic Epi, 2007