SCHOOL OF PUBLIC HEALTH

EPIDEMIOLOGY

UNIVERSITY *of* WASHINGTON

Session 11:
Gene-Environment Interactions

What are gene-environment (GE) interactions?

“A different effect of an environmental exposure on disease risk in persons with different genotypes,"

or, alternatively,

"a different effect of a genotype on disease risk in persons with different environmental exposures.”

Ottman, Prev Med 1996

Why study GE interactions?

Why study GE interactions?

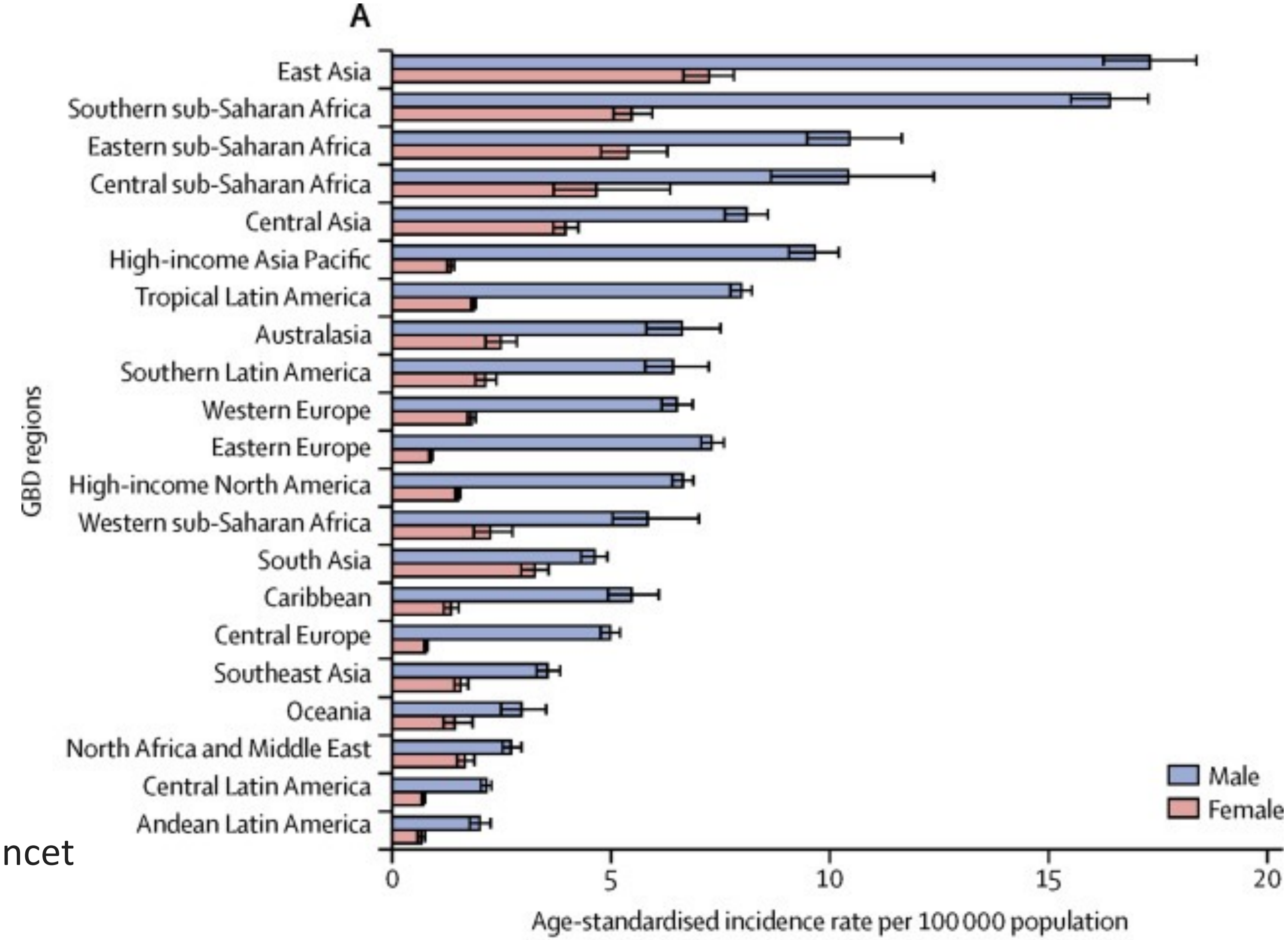
- **Gain insights about already known genes**
  - Information about effect in different strata might give insights into pathways and biology
- **Clinical Importance**
  - Disease prediction, pharmacogenetics
- **A tool in gene discovery**
  - A genetic variant is only associated with disease in exposed individuals
  - The environment risk factor is only associated with disease in those with the genetic variant
  - Incorporating GE interactions may boost power in association analysis

Example: Esophageal cancer

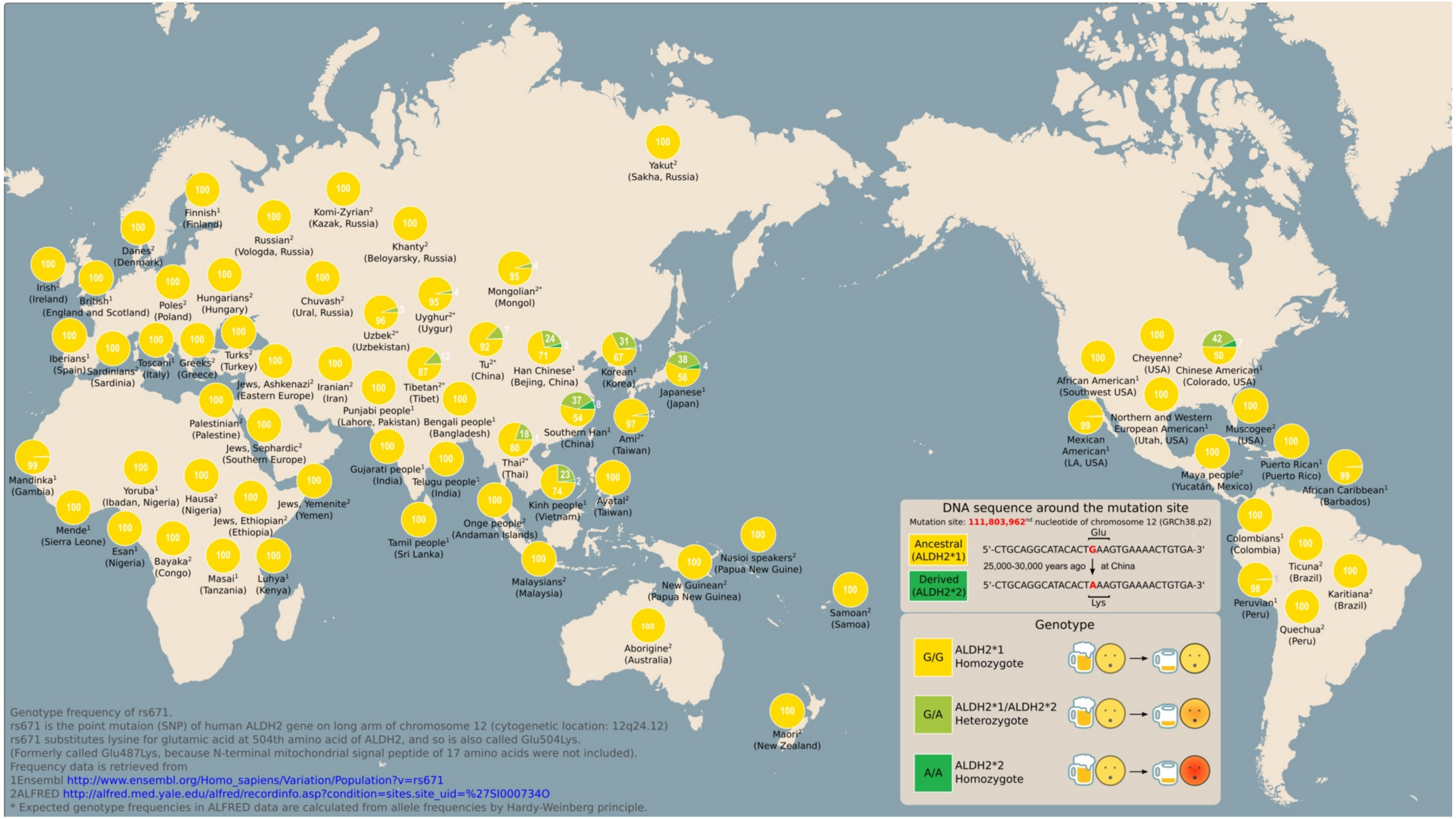
- > **Risk factors: alcohol intake, tobacco use, Barrett Syndrome, obesity**

t

GBD 2017 Oesophageal Cancer Collaborators. Lancet Gastroenterol Hepatol. 2020



Metabolism of alcohol involves the *ALDH* and *AHD* genes group
ALDH2 variation has been associated with alcohol flush reaction



DNA sequence around the mutation site

Mutation site: **111,803,962**nd nucleotide of chromosome 12 (GRCh38.p2)

Ancestral (ALDH2*1)	5'-CTGCAGGCATACACT**G**AAGTGAAAACTGTGA-3' (Glu)
	25,000-30,000 years ago ↓ at China
Derived (ALDH2*2)	5'-CTGCAGGCATACACT**A**AAGTGAAAACTGTGA-3' (Lys)

Genotype

G/G	ALDH2*1 Homozygote	→
G/A	ALDH2*1/ALDH2*2 Heterozygote	→
A/A	ALDH2*2 Homozygote	→

Genotype frequency of rs671.
rs671 is the point mutaion (SNP) of human ALDH2 gene on long arm of chromosome 12 (cytogenetic location: 12q24.12)
rs671 substitutes lysine for glutamic acid at 504th amino acid of ALDH2, and so is also called Glu504Lys.
(Formerly called Glu487Lys, because N-terminal mitochondrial signal peptide of 17 amino acids were not included).
Frequency data is retrieved from
1Ensembl http://www.ensembl.org/Homo_sapiens/Variation/Population?v=rs671
2ALFRED http://alfred.med.yale.edu/alfred/recordinfo.asp?condition=sites.site_uid=%27SI000734O
* Expected genotype frequencies in ALFRED data are calculated from allele frequencies by Hardy-Weinberg principle.

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Interaction between alcohol intake and *ADH1B* and *ALDH2* genotypes in esophageal squamous-cell carcinoma

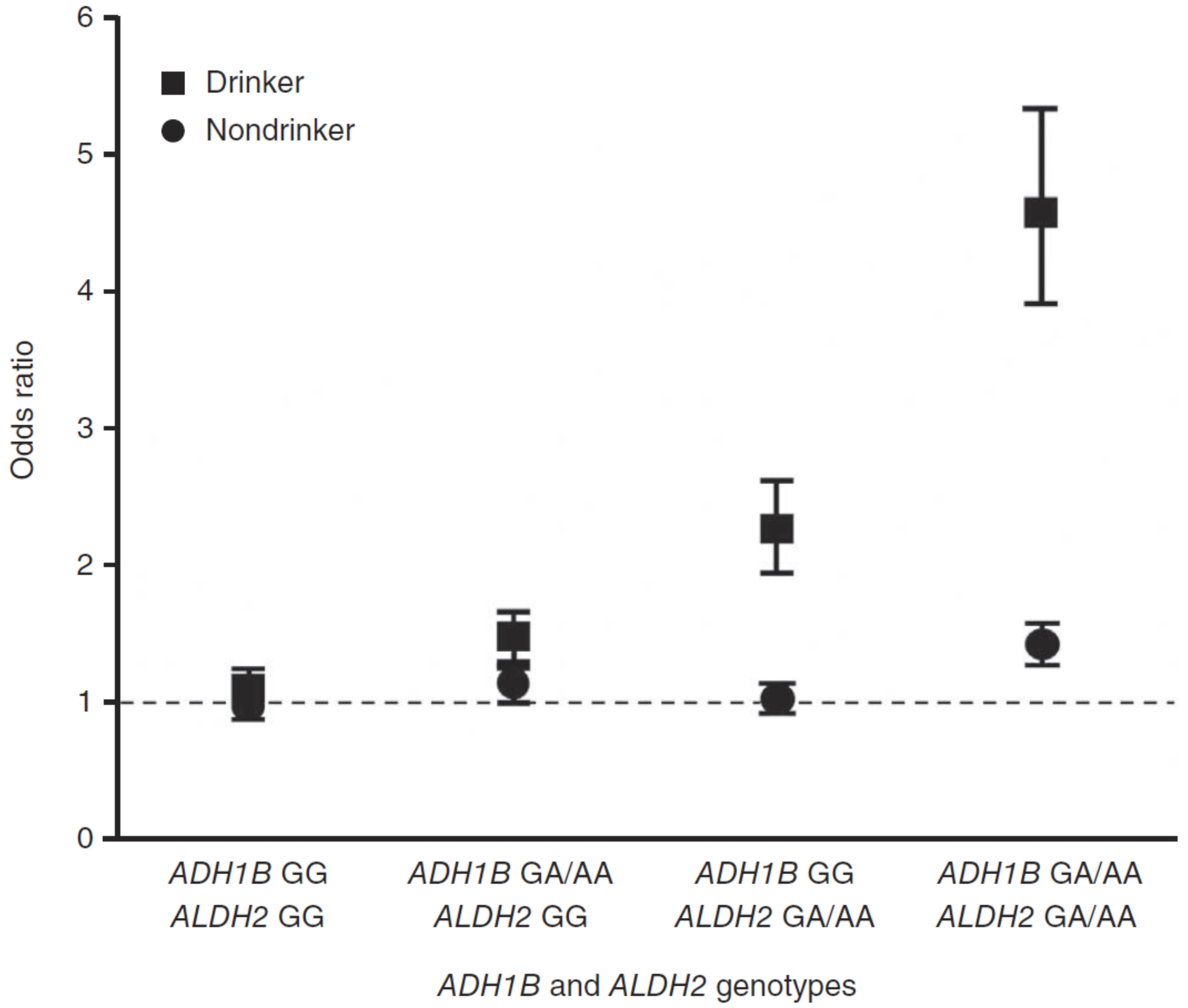


Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

Wu et al. (2012) Nat Genet

GE interactions and statistical power

> **Rule of thumb:**

You need four times as many individuals to detect an interaction effect compared to a main effect

The 4-by-2 table for case-control data

	Case	Control	OR	
G=0,E=0	N_{100}	N_{000}	1	Reference
G=1,E=0	N_{110}	N_{010}	$\frac{N_{110}N_{000}}{N_{010}N_{100}}$	Risk among unexposed carriers
G=0,E=1	N_{101}	N_{001}	$\frac{N_{101}N_{000}}{N_{001}N_{100}}$	Risk among exposed non-carriers
G=1,E=1	N_{111}	N_{011}	$\frac{N_{111}N_{000}}{N_{011}N_{100}}$	Risk among exposed carriers

Often when we talk about interaction, we talk about departure from the multiplicative scale

$$OR_{INT} = \frac{OR_{11}}{OR_{10}OR_{01}}$$

Interaction exists when observed effect of G & E together is not a simple function of their individual effects

H_0: $OR_{GE}=OR_G OR_E$ vs. H_A: $OR_{GE} \neq OR_G OR_E$

Breakout room activity

ADH1B, alcohol intake and esophageal cancer

ADH1B genotype	Cases	Controls	OR	
GG, non-drinker	1,618	2,187	1	Reference
GA+AA, non-drinker	1,211	1,440	??	Risk among unexposed carriers
GG, drinker	1,519	1,873	??	Risk among exposed non-carriers
GA+AA, drinker	1,348	1,299	??	Risk among exposed carriers

Wu, Nat Genet, 2012

1. **Calculate the stratum-specific odds ratios**
2. **Calculate the interaction odds ratio**

$$OR_{INT} = \frac{OR_{11}}{OR_{10}OR_{01}}$$

ADH1B genotype	Case	Control	OR	
GG, non-drinker	1,618	2,187	1	Reference
GA+AA, non-drinker	1,211	1,440	1.14	Risk among unexposed carriers
GG, drinker	1,519	1,873	1.10	Risk among exposed non-carriers
GA+AA, drinker	1,348	1,299	1.40	Risk among exposed carriers

$OR_{GE}=1.40/(1.10x1.14)=1.13$

Calculated estimates

ADH1B genotype	Case	Control	OR	
GG, non-drinker	1,618	2,187	1	Reference
GA+AA, non-drinker	1,211	1,440	**1.14**	Risk among unexposed carriers
GG, drinker	1,519	1,873	**1.10**	Risk among exposed non-carriers
GA+AA, drinker	1,348	1,299	**1.40**	Risk among exposed carriers

Estimates from the paper

ADH1B genotype	Case	Control	OR	
GG, non-drinker	1,618	2,187	1	Reference
GA+AA, non-drinker	1,211	1,440	**1.13**	Risk among unexposed carriers
GG, drinker	1,519	1,873	**1.15**	Risk among exposed non-carriers
GA+AA, drinker	1,348	1,299	**1.46**	Risk among exposed carriers

In practice, we often rely on regression models to test for GE interactions

$$\operatorname{logit} \mathrm{P}(D=1) = \beta + \beta_g G + \beta_e E + \beta_{ge} GE$$

$$Test : \beta_{ge} \neq 0$$

The joint 2-df interaction test

- A tool for SNP discovery
- Is a SNP associated with disease risk in any of the exposure sub-groups?

- Compare “main effect of E only” model to “main effects plus interaction” model

Null model: $logit\ P(D=1)=\beta+\beta_e E$

Alternative model: $logit\ P(D=1)=\beta+\beta_e E+\beta_g G+\beta_{ge} GE$

Compare -2 log L_{null} +2 log L_{alt} to chi-square 2 d.f.

Kraft et al, Hum Hered. 2007

Case-Only Analysis

Based on genotype-exposure table in CASES

	Carrier	Non-carrier
Exposed	N_{11}	N_{12}
Unexposed	N_{21}	N_{22}

Genotypic odds ratios for exposure from this table are equal to interaction relative risks **only if genotypes and exposure are not correlated in general population**.

Assume that G and E are independent in the source population:

An association between G and E among the cases indicates a departure from a multiplicative odds model. (i.e., regress E on G in cases—if there is an association, there is an "interaction.")

Can be much more powerful than traditional logistic regression analysis!

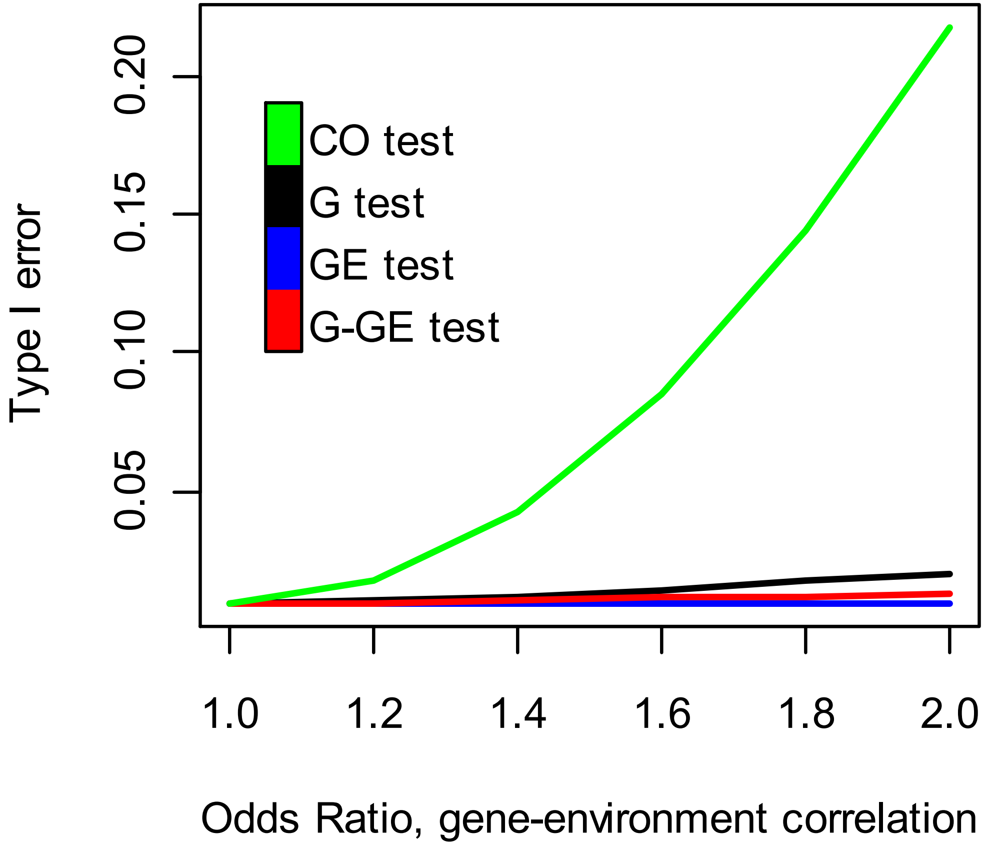
Piegorsch et al, Stat Med 1994

Does this mean I can throw away all my controls (and decrease genotyping cost)?

- The increase in power is due to your assumption that G and E are independent of each other (which you can test in your controls)

- Controls allow for estimation main effects for G and E and will also allow for calculating stratum-specific ORs

What if G and E are (positively) correlated?



Type I error rates as a function of GE dependence.

Sensitivity= 0.6

Specificity = 0.9

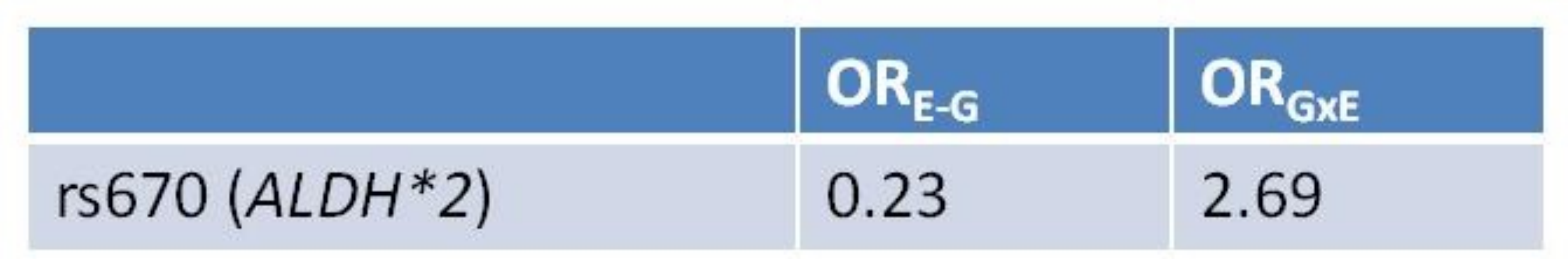
OR(E)= 1.6

Lindstrom et al, Hum Hered. 2009

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

What if there is a negative correlation between G and E?

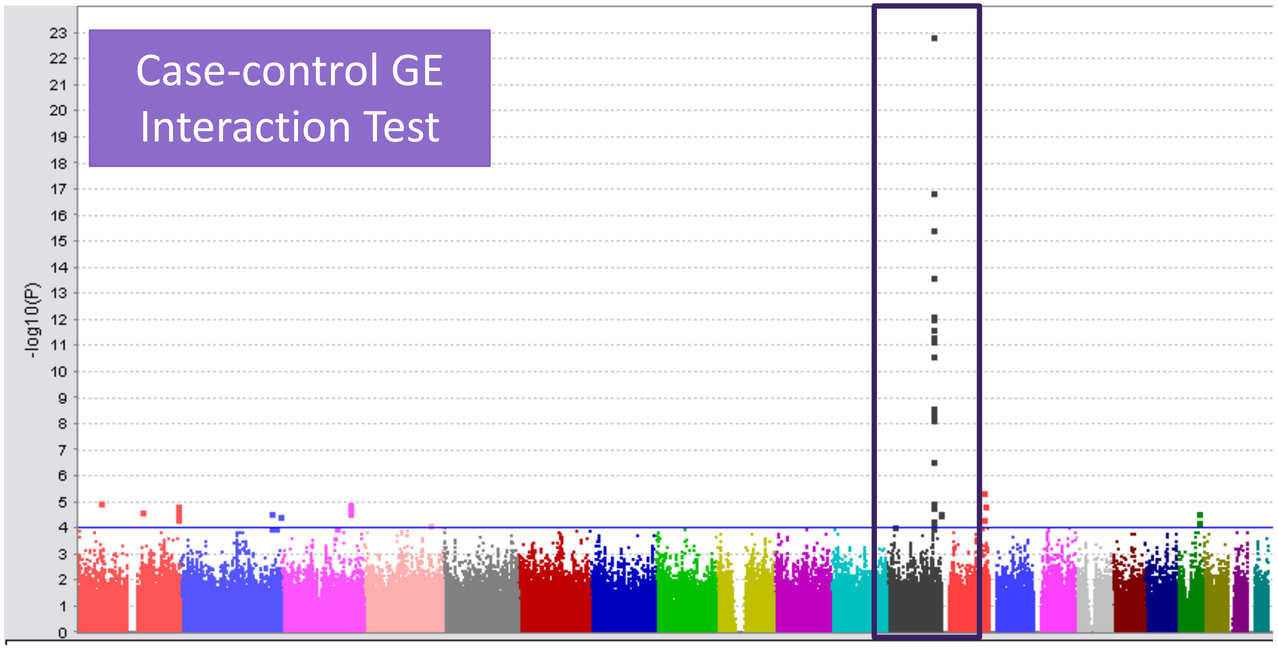
ALDH2, alcohol intake and esophageal cancer



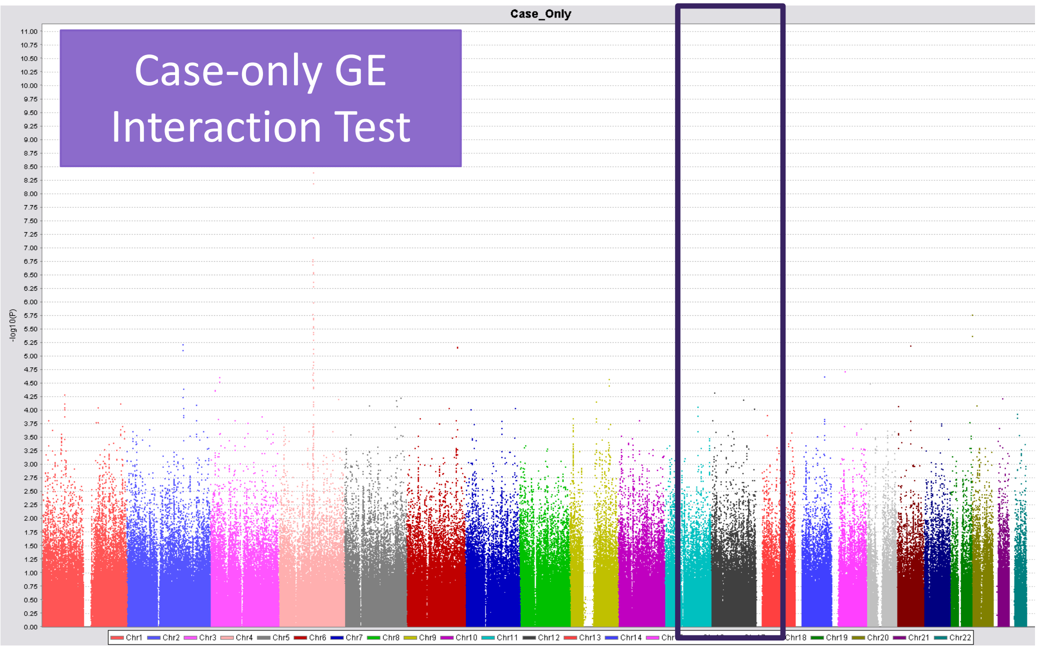
	$OR_{E\text{-}G}$	OR_{GxE}
rs670 (*ALDH*2*)	0.23	2.69

The risk allele is associated with a decreased risk of heavy drinking in the general population, and an increase in the effect of alcohol on esophageal cancer risk

ALDH2, alcohol intake and esophageal cancer



Case-control GE Interaction Test



Case-only GE Interaction Test

Courtesy of Chen Wu

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Empirical Bayes Estimator

- If G and E are independent: Use the case-only test.
- If G and E are not independent: Use the classical 1-df GE interaction test in a case-control setting

- Trade-off between bias and efficiency:

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

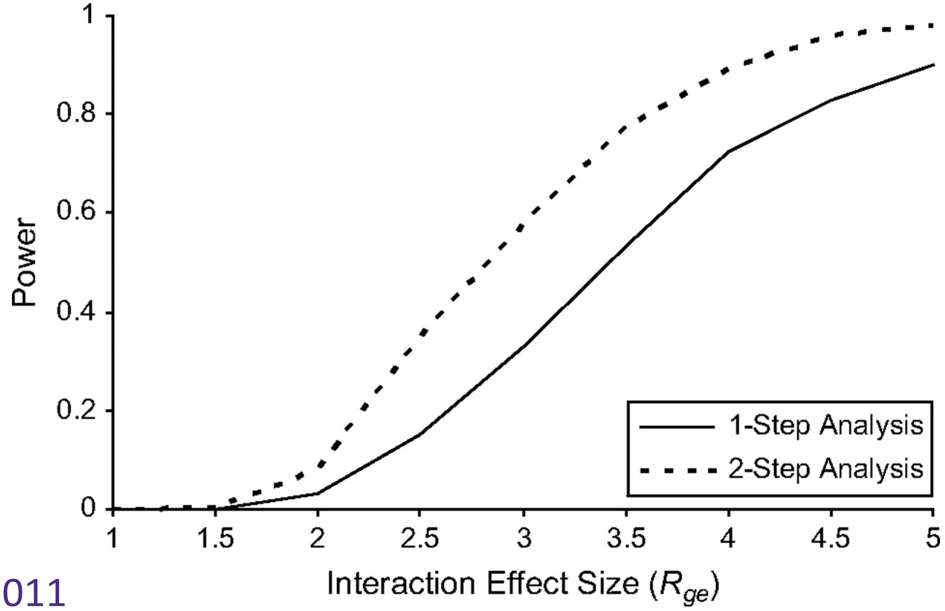
- $\hat{\tau}^2$ is an estimate of the G-E dependence

Mukherjee & Chatterjee, Biometrika 2007

Genome-wide GE Interaction analysis: 2-step approaches

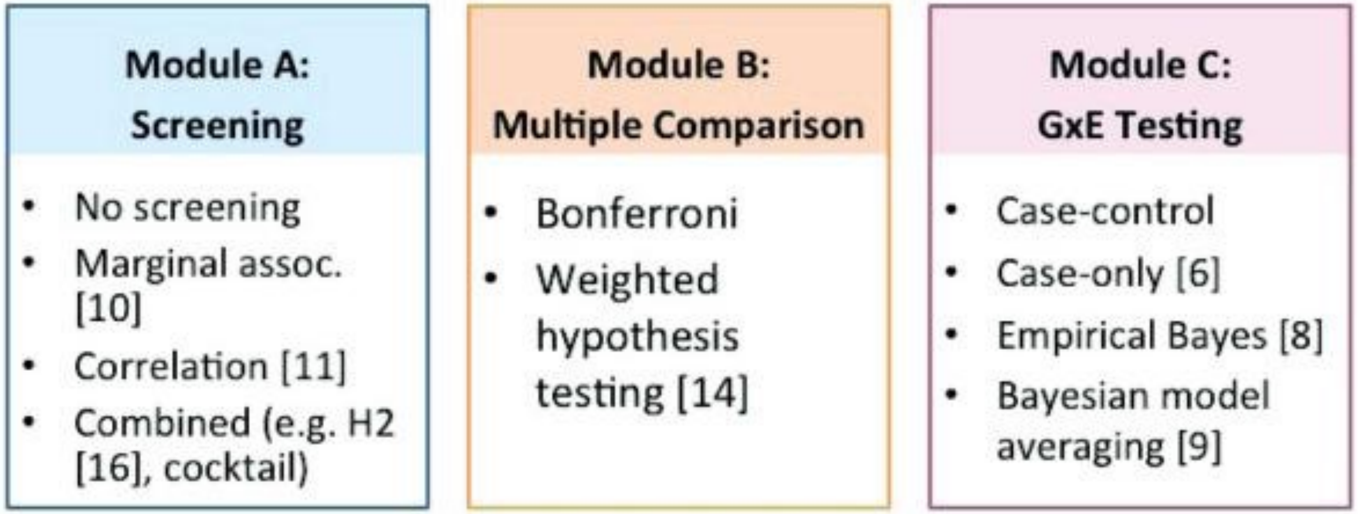
1. Test for G-E dependence and/or associations between the SNP and your outcome in your entire dataset. Select SNPs with p<α_1

2. Take m SNPs from stage 1 and perform traditional GE interaction tests in a case-control setting (1 df). All SNPs with p<α/m are declared significant



Murcray, Am J Epi 2009, Genet Epi 2011

Modular approach for genome-wide GE interactions



Module A: Screening

- No screening
- Marginal assoc. [10]
- Correlation [11]
- Combined (e.g. H2 [16], cocktail)

Module B: Multiple Comparison

- Bonferroni
- Weighted hypothesis testing [14]

Module C: GxE Testing

- Case-control
- Case-only [6]
- Empirical Bayes [8]
- Bayesian model averaging [9]

Be careful with mix-and-matching methods across modules!

You can use any of the case-control, the case-only, or the EB test to test GE if you used marginal association test for screening, but only the case-control test you used correlation for screening.

(This is to make sure the different modules are independent of each other so that you will maintain valid Type I error rates)

Hsu et al, Genetic Epi, 2012

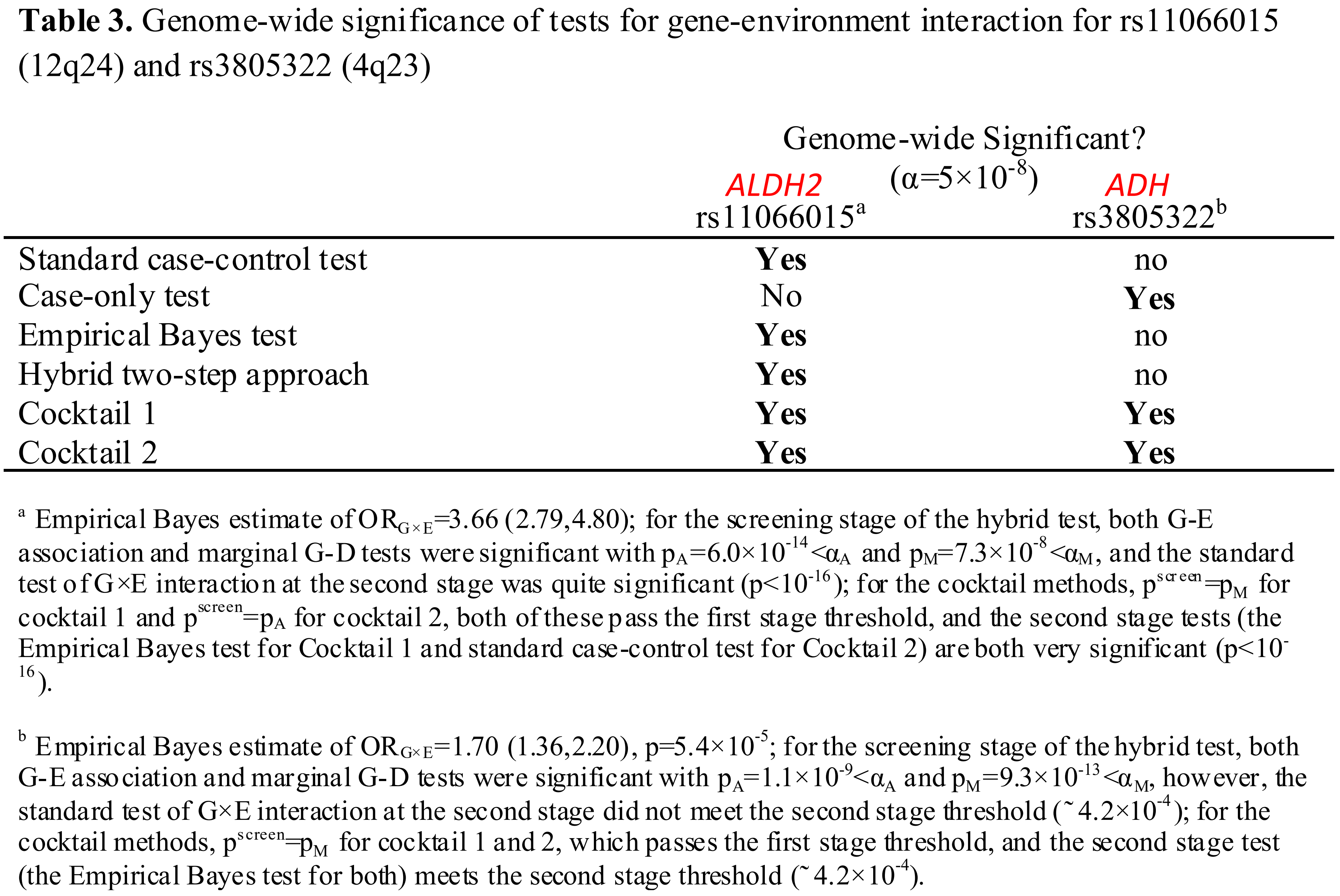


Table 3. Genome-wide significance of tests for gene-environment interaction for rs11066015 (12q24) and rs3805322 (4q23)

	Genome-wide Significant? ($\alpha=5\times10^{-8}$)	
	ALDH2 rs11066015[a]	*ADH* rs3805322[b]
Standard case-control test	**Yes**	no
Case-only test	No	**Yes**
Empirical Bayes test	**Yes**	no
Hybrid two-step approach	**Yes**	no
Cocktail 1	**Yes**	**Yes**
Cocktail 2	**Yes**	**Yes**

[a] Empirical Bayes estimate of $OR_{G\times E}=3.66$ (2.79,4.80); for the screening stage of the hybrid test, both G-E association and marginal G-D tests were significant with $p_A=6.0\times10^{-14}<\alpha_A$ and $p_M=7.3\times10^{-8}<\alpha_M$, and the standard test of G×E interaction at the second stage was quite significant ($p<10^{-16}$); for the cocktail methods, $p^{screen}=p_M$ for cocktail 1 and $p^{screen}=p_A$ for cocktail 2, both of these pass the first stage threshold, and the second stage tests (the Empirical Bayes test for Cocktail 1 and standard case-control test for Cocktail 2) are both very significant ($p<10^{-16}$).

[b] Empirical Bayes estimate of $OR_{G\times E}=1.70$ (1.36,2.20), $p=5.4\times10^{-5}$; for the screening stage of the hybrid test, both G-E association and marginal G-D tests were significant with $p_A=1.1\times10^{-9}<\alpha_A$ and $p_M=9.3\times10^{-13}<\alpha_M$, however, the standard test of G×E interaction at the second stage did not meet the second stage threshold ($\tilde{}4.2\times10^{-4}$); for the cocktail methods, $p^{screen}=p_M$ for cocktail 1 and 2, which passes the first stage threshold, and the second stage test (the Empirical Bayes test for both) meets the second stage threshold ($\tilde{}4.2\times10^{-4}$).

Wu et al, Genet Epi 2014

GE interaction tests for set-based approaches

> Look at the interaction between E and some combination of markers

> Particularly useful for rare variants

> Groups SNPs by pathways, genes, genome-wide significant SNPs, etc
  - SBERIA (Jiao et al, Genetic Epi 2013)
  - eSBERIA and coSBERIA (Jiao et al, Genetic Epi 2015)
  - GESAT (Lin et al, Biostatistics 2013)
  - iSKAT (Lin et al, Biometrics 2016)
  - MiSTi (Su, Biostatistics 2017)

GE interaction tests for continuous phenotypes

- Classical approach:

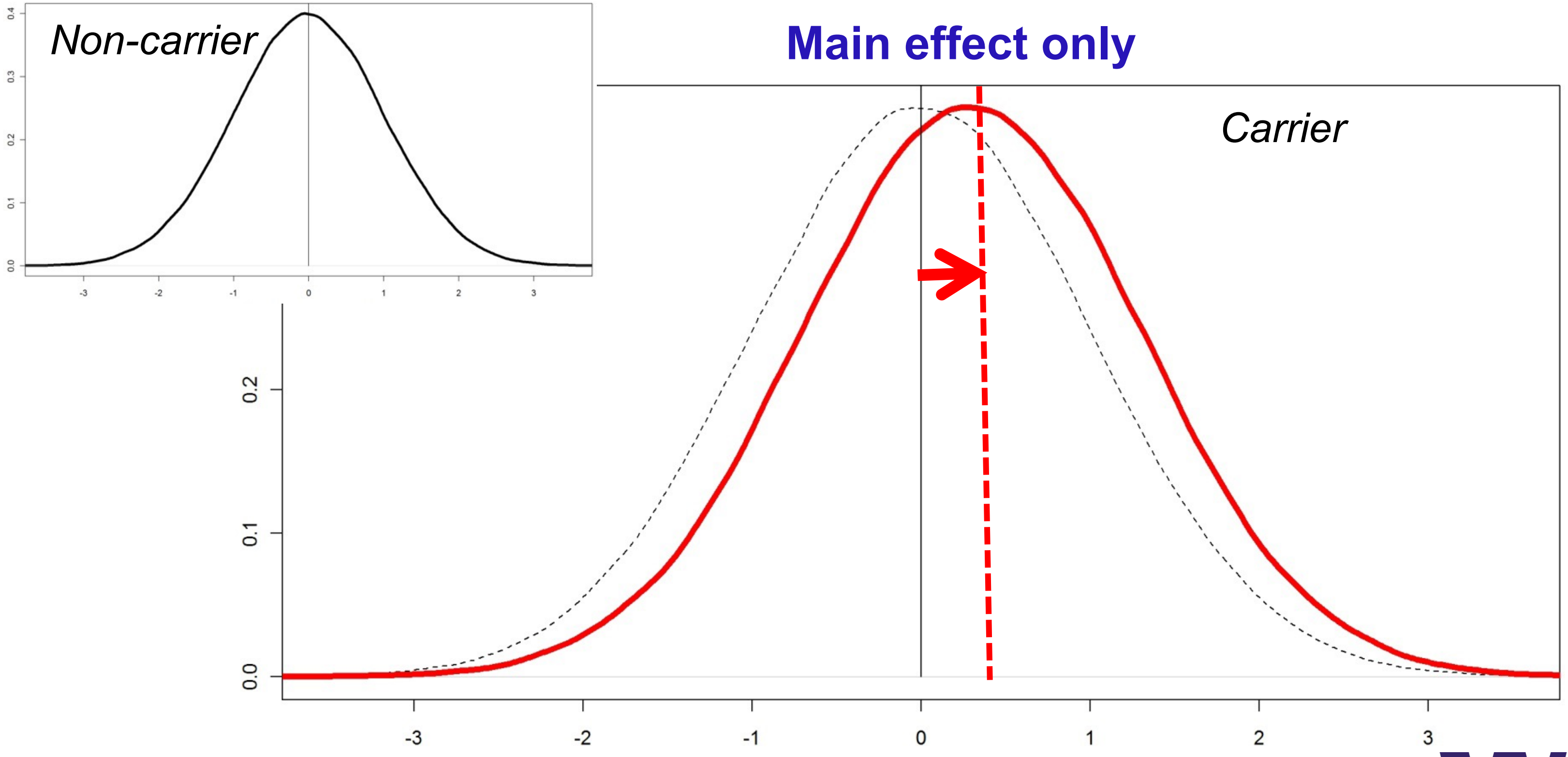
$$Y = b_0 + b_g\ G + b_e\ E + b_{ge}\ GE \quad \text{(linear regression)}$$

- Alternative approach:
  - Step 1: Look at the distribution of the trait across genotype classes. Move forward SNPs with evidence of unequal distribution across genotypes. Don’t need E.
  - Step 2: Conduct classic linear regression on SNPs selected in step 1.

Pare et al. PLoS Genetics 2010
Aschard et al. Genetc Epi 2013

GE interaction for continuous phenotypes (ii)

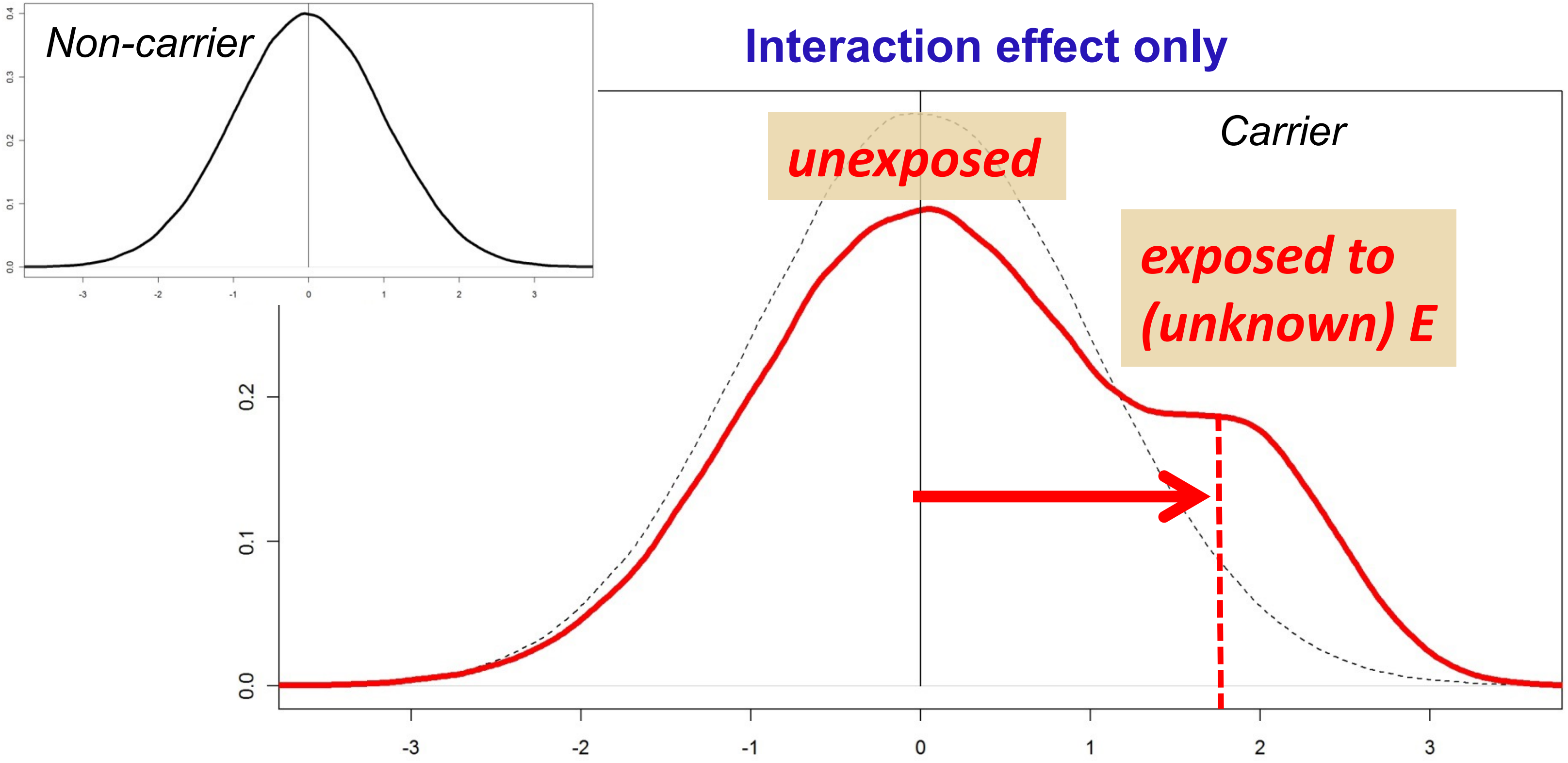
> For quantitative phenotypes, the distribution of phenotypic values by genotypic classes will be different in the presence of main effect only or interaction effect



Main effect only

GE interaction for continuous phenotypes (iii)

> For quantitative phenotypes, the distribution of phenotypic values by genotypic classes will be different in the presence of main effect only or interaction effect



Interaction effect only

unexposed

exposed to (unknown) E

GE interaction studies require large sample sizes

> A common approach is to pool data from multiple studies within large international consortia.

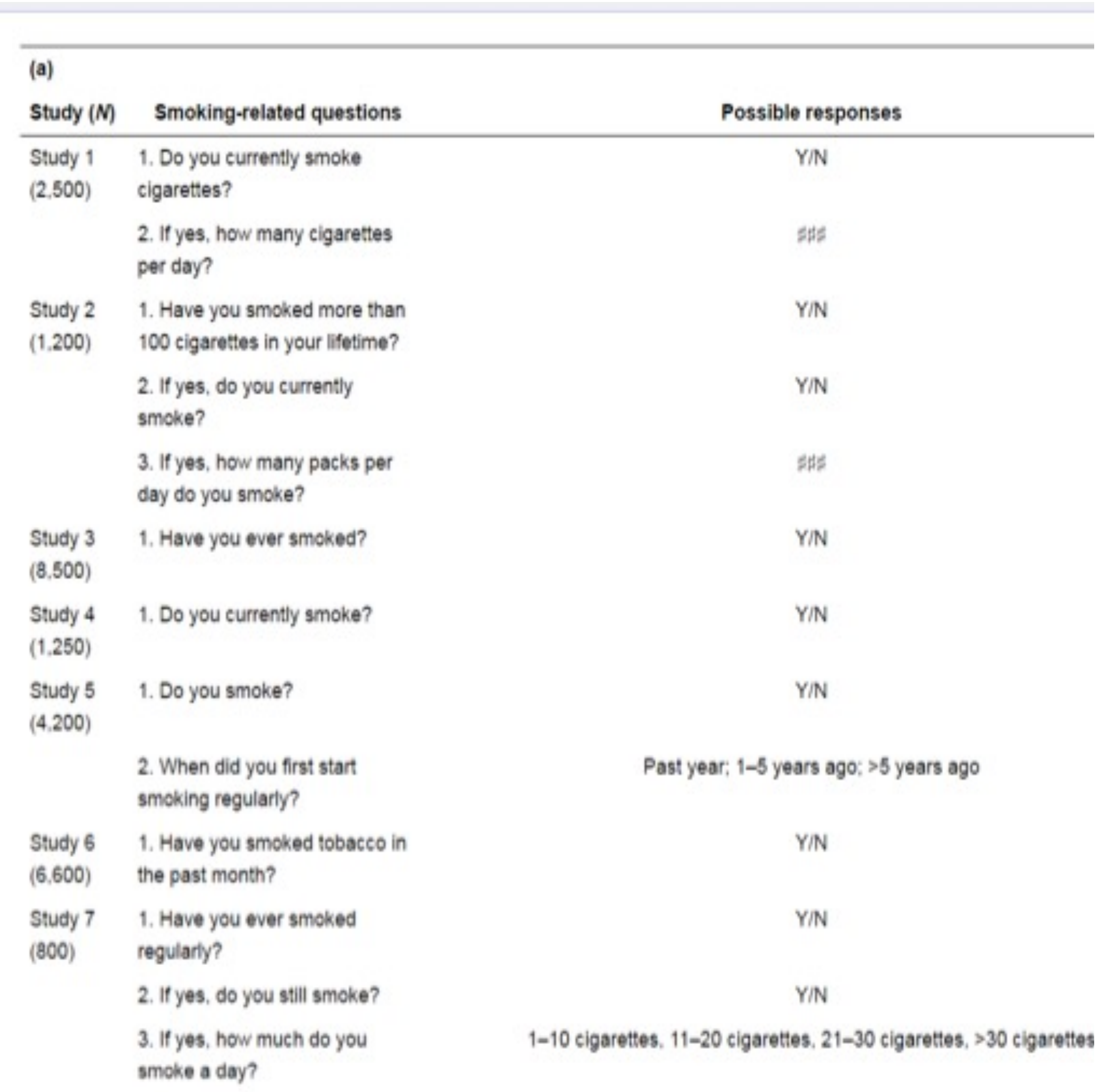
> Although this will result in greatly increases sample size, it introduces challenges for harmonizing data across studies. This is often the most difficult and time-consuming part of multi-study GE interaction research

Breakout Activity

> You are conducting a GE interaction study, where the environmental exposure is smoking. Your have data from multiple studies which means your total sample size is 25,050 subjects!

> You need to harmonize the smoking exposure across studies (see Table). You are trying to build the biggest dataset you can, but you must be able to use the same definition of smoking. What are the samples sizes you could have in your study if you used the following definitions for your "smoking" exposure?

a. Cigarettes per day
b. Ever smoker
c. Current smoker



(a)

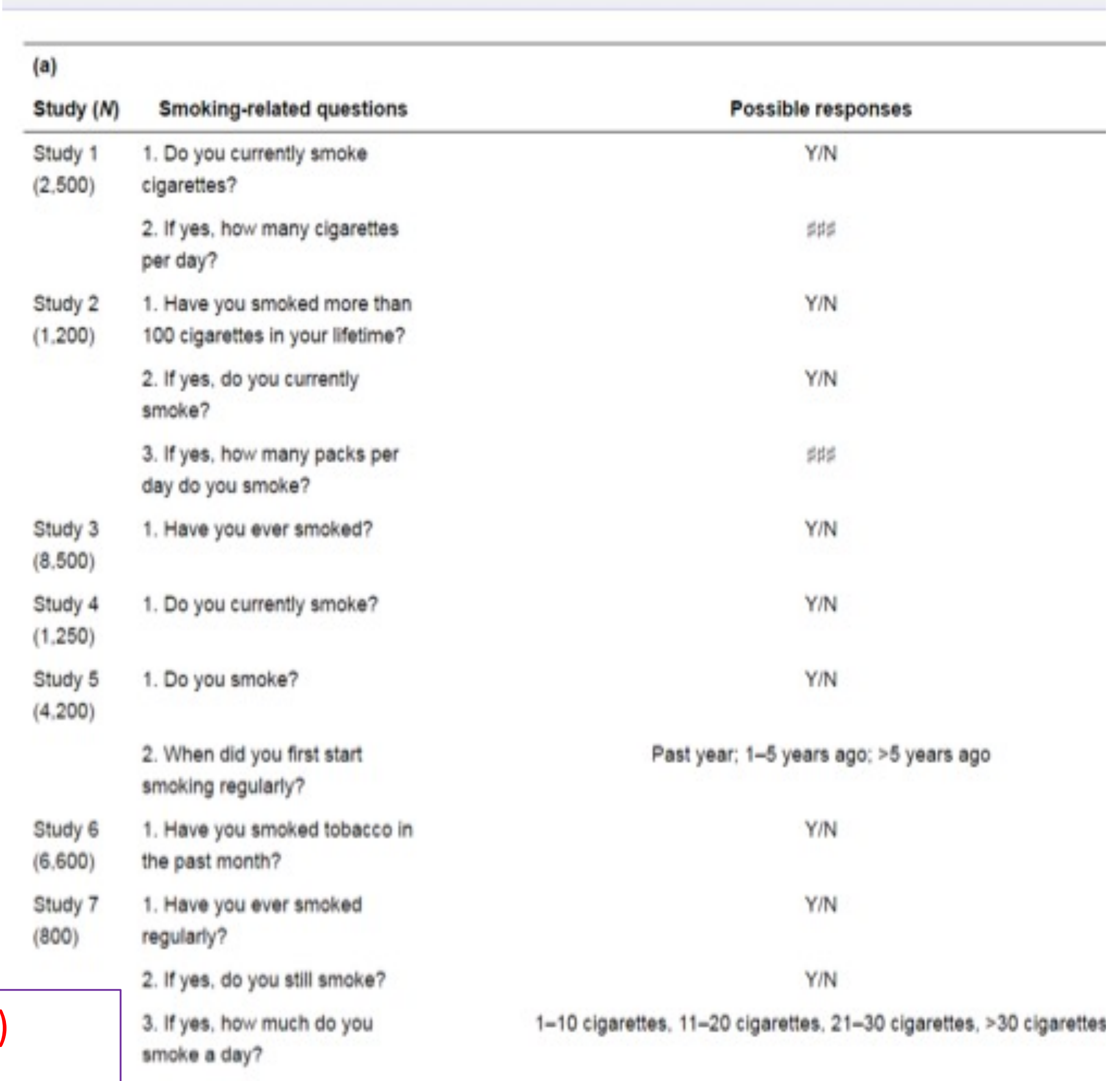
Study (N)	Smoking-related questions	Possible responses
Study 1 (2,500)	1. Do you currently smoke cigarettes?	Y/N
	2. If yes, how many cigarettes per day?	###
Study 2 (1,200)	1. Have you smoked more than 100 cigarettes in your lifetime?	Y/N
	2. If yes, do you currently smoke?	Y/N
	3. If yes, how many packs per day do you smoke?	###
Study 3 (8,500)	1. Have you ever smoked?	Y/N
Study 4 (1,250)	1. Do you currently smoke?	Y/N
Study 5 (4,200)	1. Do you smoke?	Y/N
	2. When did you first start smoking regularly?	Past year; 1–5 years ago; >5 years ago
Study 6 (6,600)	1. Have you smoked tobacco in the past month?	Y/N
Study 7 (800)	1. Have you ever smoked regularly?	Y/N
	2. If yes, do you still smoke?	Y/N
	3. If yes, how much do you smoke a day?	1–10 cigarettes, 11–20 cigarettes, 21–30 cigarettes, >30 cigarettes

Breakout Activity

- > You are conducting a GE interaction study, where the environmental exposure is smoking. Your have data from multiple studies which means your total sample size is 25,050 subjects!

- > You need to harmonize the smoking exposure across studies (see Table). You are trying to build the biggest dataset you can, but you must be able to use the same definition of smoking. What are the samples sizes you could have in your study if you used the following definitions for your “smoking” exposure?

a. Cigarettes per day – 4,500 (Study 1 and 7 and convert 2)
b. Ever smoker – 10,500 (Studies 2, 3 and 7)
c. Current smoker – 16,660 (Studies 1, 2, 4, 5, 6, 7)



(a)

Study (*N*)	Smoking-related questions	Possible responses
Study 1 (2,500)	1. Do you currently smoke cigarettes?	Y/N
	2. If yes, how many cigarettes per day?	###
Study 2 (1,200)	1. Have you smoked more than 100 cigarettes in your lifetime?	Y/N
	2. If yes, do you currently smoke?	Y/N
	3. If yes, how many packs per day do you smoke?	###
Study 3 (8,500)	1. Have you ever smoked?	Y/N
Study 4 (1,250)	1. Do you currently smoke?	Y/N
Study 5 (4,200)	1. Do you smoke?	Y/N
	2. When did you first start smoking regularly?	Past year; 1–5 years ago; >5 years ago
Study 6 (6,600)	1. Have you smoked tobacco in the past month?	Y/N
Study 7 (800)	1. Have you ever smoked regularly?	Y/N
	2. If yes, do you still smoke?	Y/N
	3. If yes, how much do you smoke a day?	1–10 cigarettes, 11–20 cigarettes, 21–30 cigarettes, >30 cigarettes

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Practical issues in GE interaction studies

> Measurement Errors
> Distribution of E/Replication
> Modeling E
> Software

Measurement Error: Continuous Outcome

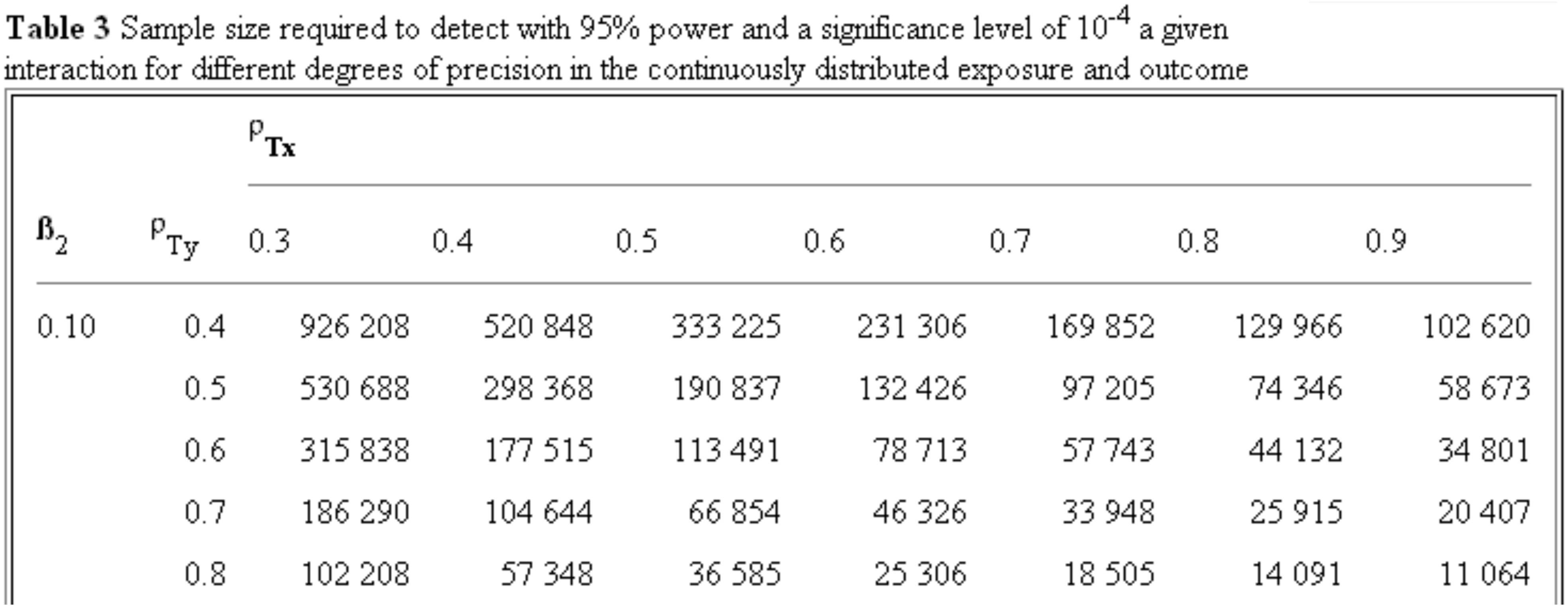
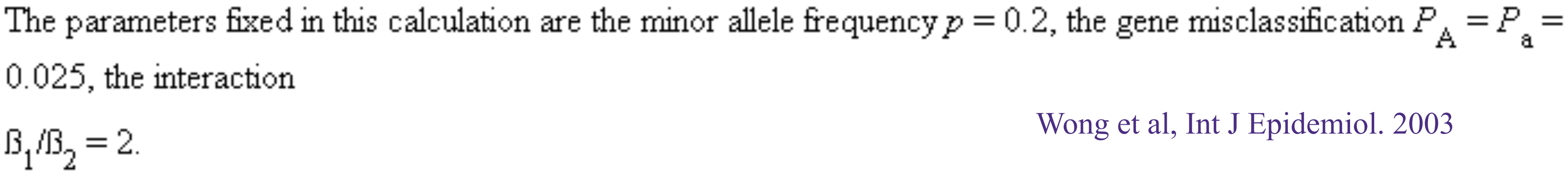


Table 3 Sample size required to detect with 95% power and a significance level of 10^{-4} a given interaction for different degrees of precision in the continuously distributed exposure and outcome

		$\rho_{\mathbf{Tx}}$						
β_2	ρ_{Ty}	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.10	0.4	926 208	520 848	333 225	231 306	169 852	129 966	102 620
	0.5	530 688	298 368	190 837	132 426	97 205	74 346	58 673
	0.6	315 838	177 515	113 491	78 713	57 743	44 132	34 801
	0.7	186 290	104 644	66 854	46 326	33 948	25 915	20 407
	0.8	102 208	57 348	36 585	25 306	18 505	14 091	11 064



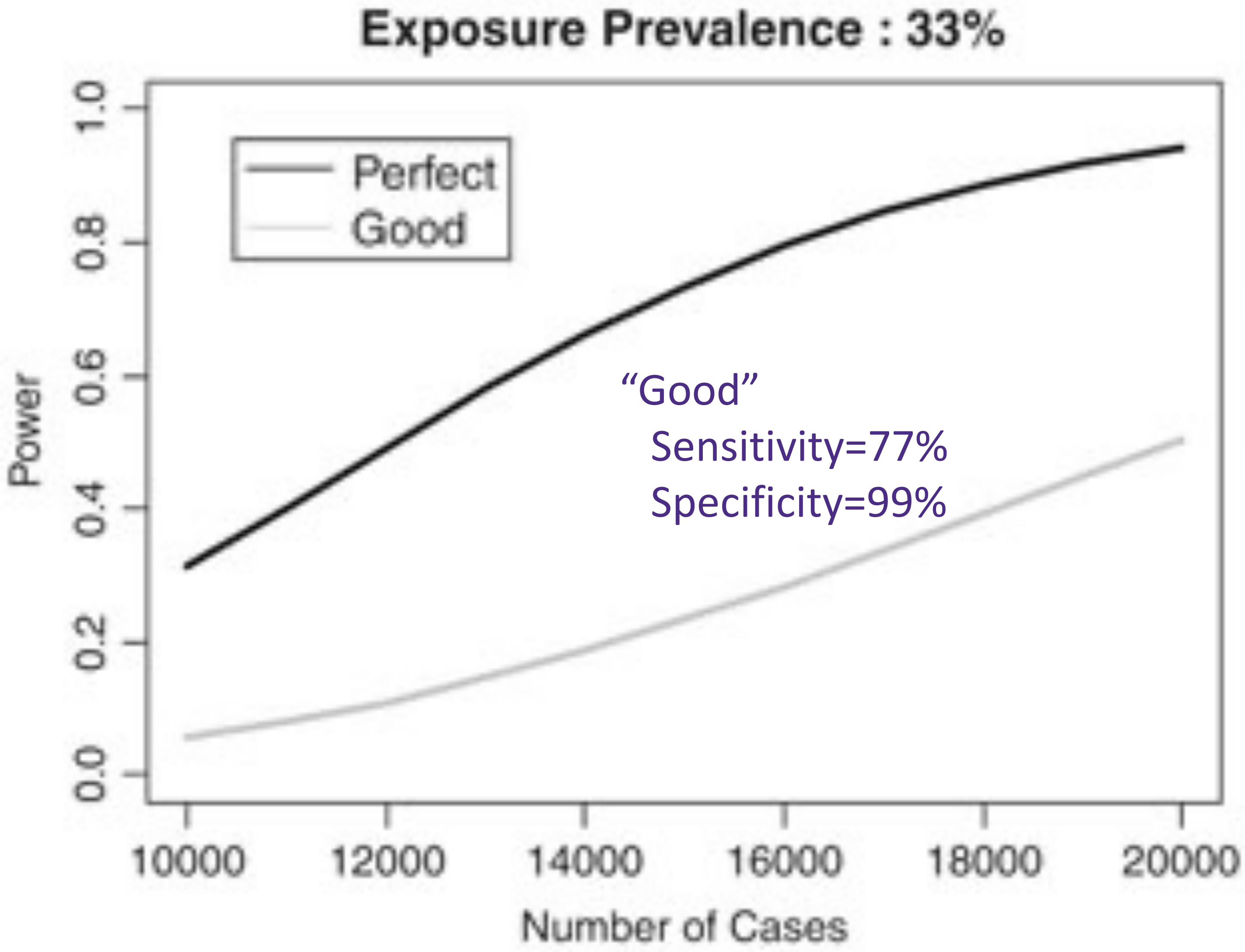
The parameters fixed in this calculation are the minor allele frequency $p = 0.2$, the gene misclassification $P_A = P_a =$ 0.025, the interaction

$\beta_1/\beta_2 = 2$.

Wong et al, Int J Epidemiol. 2003

Measurement Error: Case-control outcomes

- Even small measurement errors can greatly decrease power to detect GE interactions



“Good”
Sensitivity=77%
Specificity=99%

Bennett SN, et al. Genet Epidemiol. 2011

How, where, and when you measure the exposure has consequences for GE interaction studies

Example: FTO, Physical Activity and Obesity

- Meta-analysis of 218,166 European-ancestry subjects
- Risk of Obesity (BMI ≥ 30 vs. BMI < 25 kg/m^2) for *FTO* SNP rs9939609

	OR (95% CI)
Inactive	1.30 (1.24-1.36)
Active	1.22 (1.19-1.25)
rs9939609 x physical activity interaction	0.92 (0.88-0.97)
	P-value = 0.001

Kilpelainen, 2011

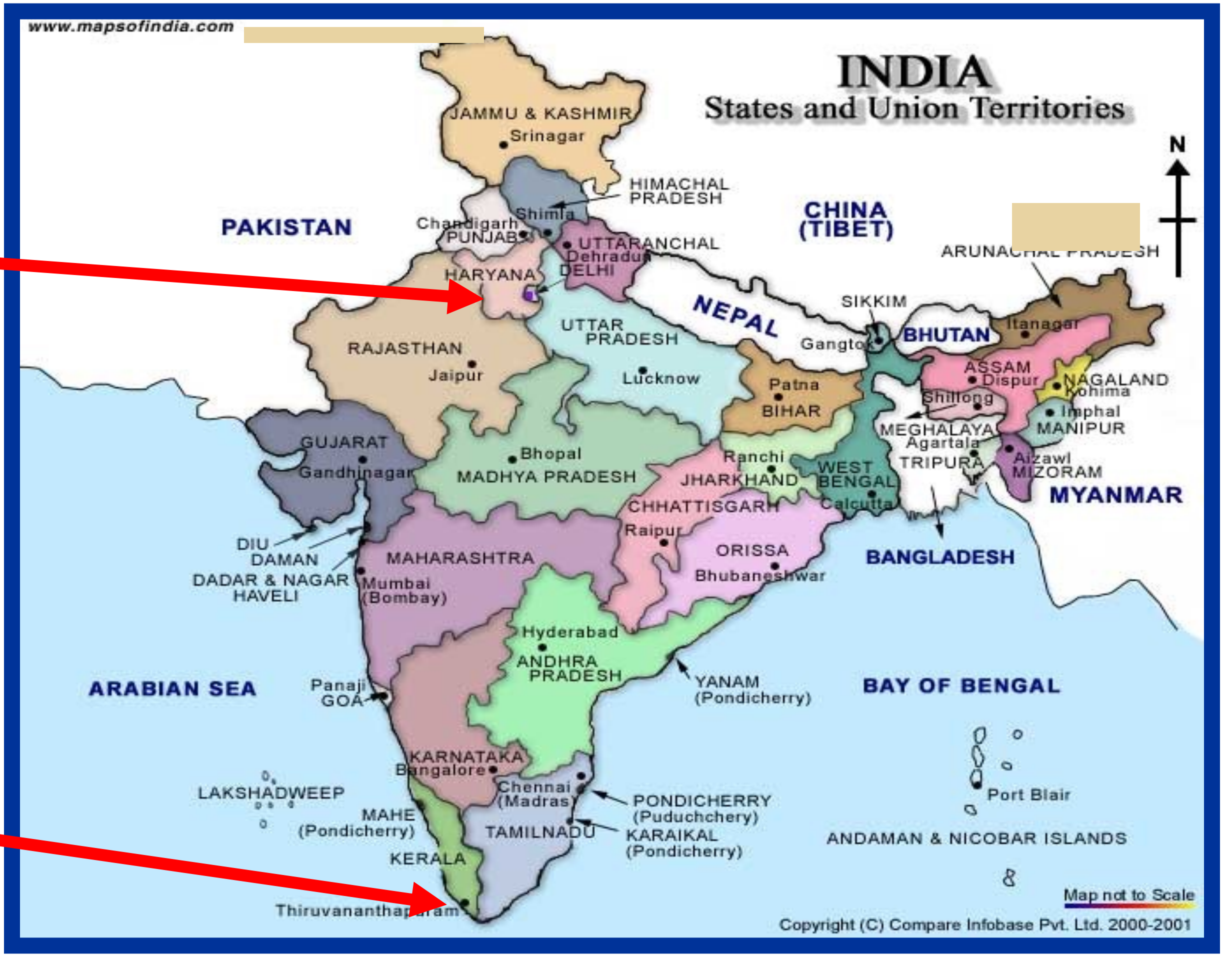
Slide courtesy of Leah Mechanic

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Interaction between *FTO*, physical activity and obesity in the India health study

New Delhi

Trivandrum



Slides courtesy of N Chatterjee

Moore, Obesity 2012

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Characteristics	New Delhi	Trivandrum
Total (n=1,313)	n=619	n=694
Age, years (mean, SD)	47.4 ± 10.0	48.8 ± 9.2
Household monthly income, %		
<5,000 rupees	7.1	71.9
>10,000 rupees	76.7	3.1
Household items, %		
Car	25	7
Refrigerator	87	58
Washing machine	79	14
Total physical activity, MET-hr/wk	42.5 ± 43.8	147.3 ± 85.2
Vigorous physical activity, MET-hr/wk	0.6 ± 6.8	26.2 ± 51.4
Sitting, hr/day	10.4 ± 2.0	5.0 ± 2.3
Centrally obese, %	82.1	60.2

Moore, Obesity 2012

Association of *FTO* SNP rs3751812 with waist circumference

	N	Effect size per T allele (95% CI)	P_{trend}
Overall	1,209	+1.61 cm (0.67, 2.55)	0.0008
New Delhi	578	+2.53 cm (1.08, 3.97)	0.0006
Trivandrum	574	+0.87 cm (-0.35, 2.08)	0.16

Moore, Obesity 2012

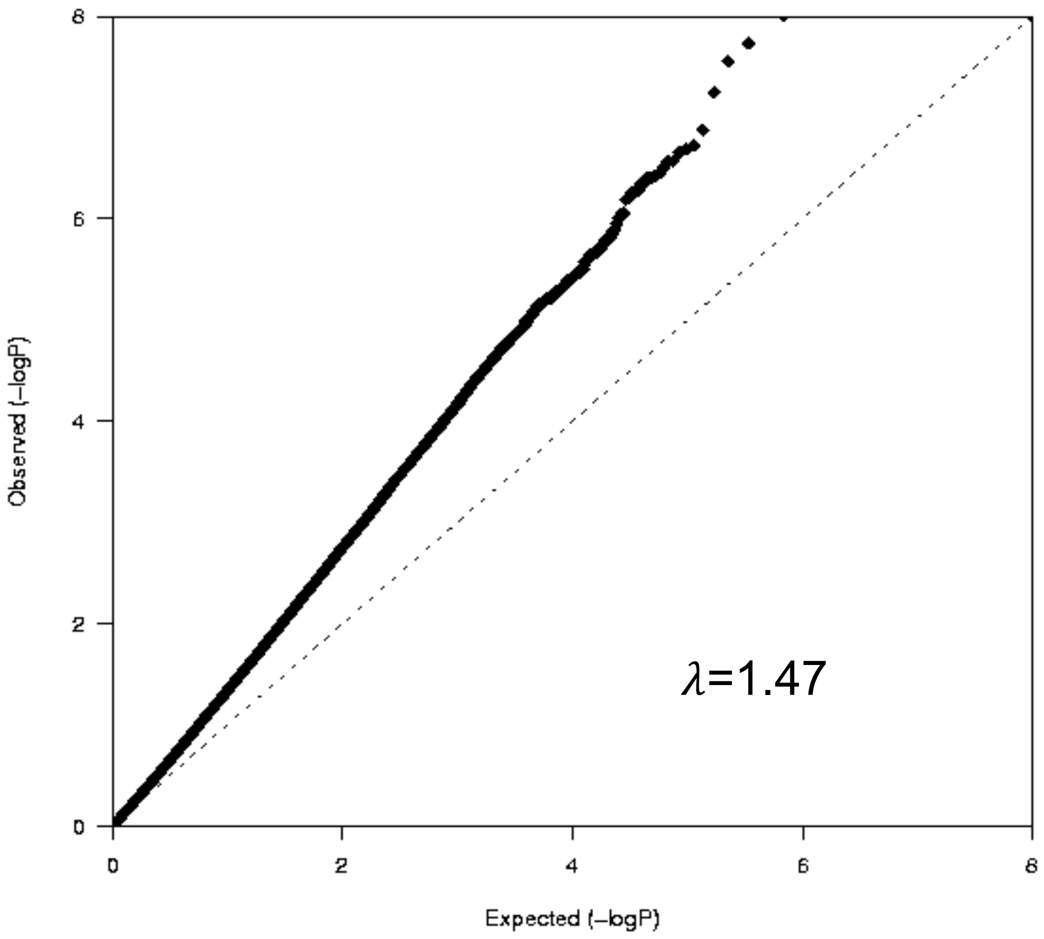
Association between rs3751812 and waist circumference by physical activity

	N	Effect per T allele (95% CI)	P_{trend}	P_{Int}
Overall	**1,209**	**+1.61 cm (0.67, 2.55)**	**0.0008**	**0.009**
New Delhi	**578**	**+2.53 cm (1.08, 3.97)**	**0.0006**	**0.59**
By PA				
≤ 91 MET-hrs/wk	517	+2.36 cm (0.82, 3.89)	0.003	
92-151 MET-hrs/wk	32	+6.39 cm (1.94, 10.85)	0.005	
152-217 MET-hrs/wk	24	-0.95 cm (-7.33, 5.42)	0.77	
218+ MET-hrs/wk	5	N/A	N/A	
Trivandrum	**574**	**+0.87 cm (-0.35, 2.08)**	**0.16**	**0.004**
By PA				
≤ 91 MET-hrs/wk	170	+3.50 cm (0.90, 6.10)	0.008	
92-151 MET-hrs/wk	132	+1.13 cm (-1.08, 3.33)	0.32	
152-217 MET-hrs/wk	141	+1.04 cm (-1.63, 3.70)	0.45	
218+ MET-hrs/wk	131	-2.32 cm (-4.82, 0.18)	0.07	

Moore, Obesity 2012

A note about modeling “E” (i)

- Genome-wide GE interaction study of BMI and Type II Diabetes
- Standard 1 df case-control test for GE interaction

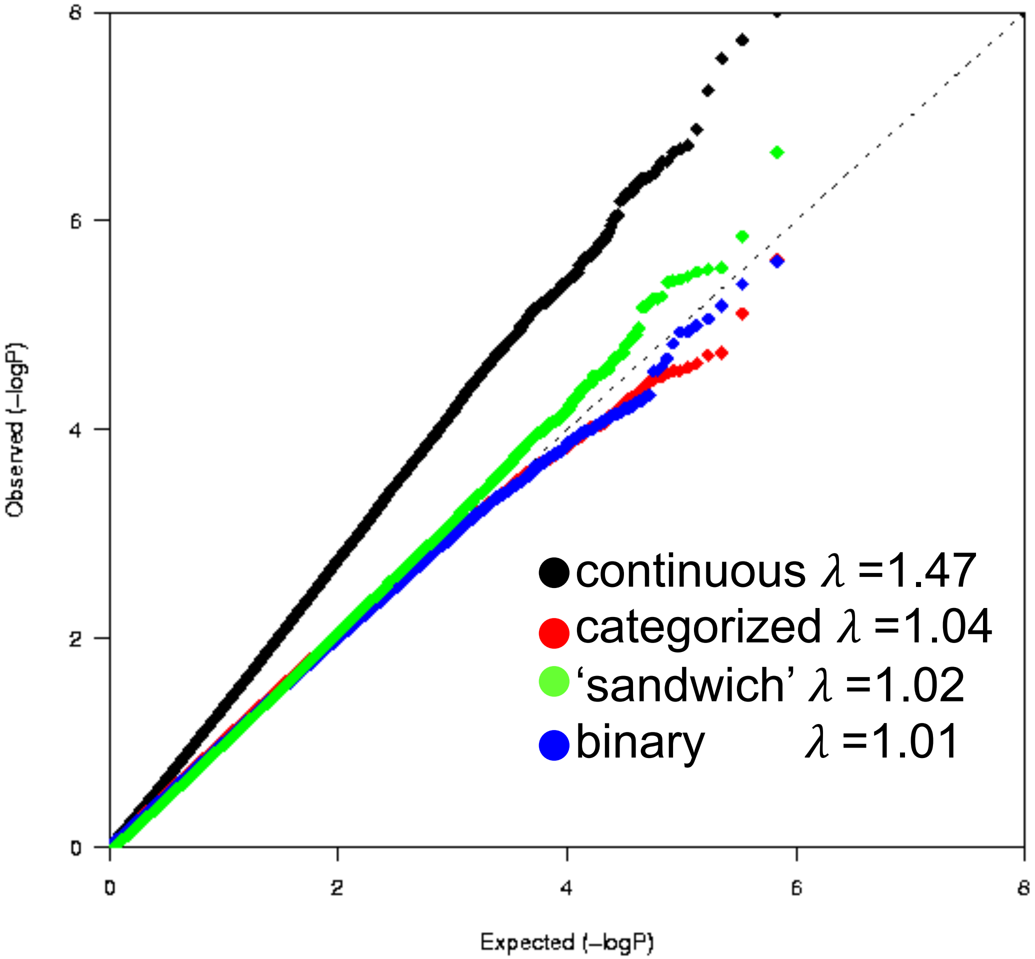


λ=1.47

Slide courtesy Marilyn Cornelis

Tchetgen Tchetgen and Kraft, Epidemiology, 2011

A note about modeling “E” (ii)



Slide courtesy Marilyn Cornelis

Tchetgen Tchetgen and Kraft, Epidemiology, 2011

GE interaction Software

Software	Good for	URL
PLINK	GWAS, data handling, 1df GxE test, joint test	http://pngu.mgh.harvard.edu/~purcell/plink/
GxEscan	R script incorporating multiple genome-wide GxE interaction tests	http://biostats.usc.edu/software
R	Flexible, write your own scripts	http://www.r-project.org/
METAL	Meta-analysis	http://www.sph.umich.edu/csg/abecasis/metal/
CGEN	R package, additive interaction	https://rdrr.io/bioc/CGEN/man/additive.test.html
Quanto (power)	Joint test, GE test, family-based designs, case-control, continuous outcome	http://hydra.usc.edu/gxe/

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH