SCHOOL OF PUBLIC HEALTH
EPIDEMIOLOGY
UNIVERSITY of WASHINGTON

Session 11:
Gene-Environment Interactions

What is gene-environment interaction?

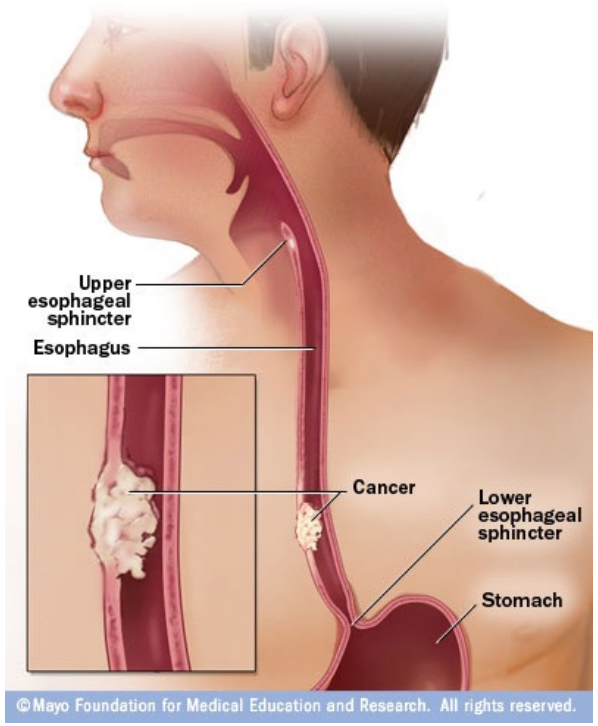
"A different effect of an environmental exposure on disease risk in persons with different genotypes," **or, alternatively,** ***"a different effect of a genotype on disease risk in persons with different environmental exposures."***

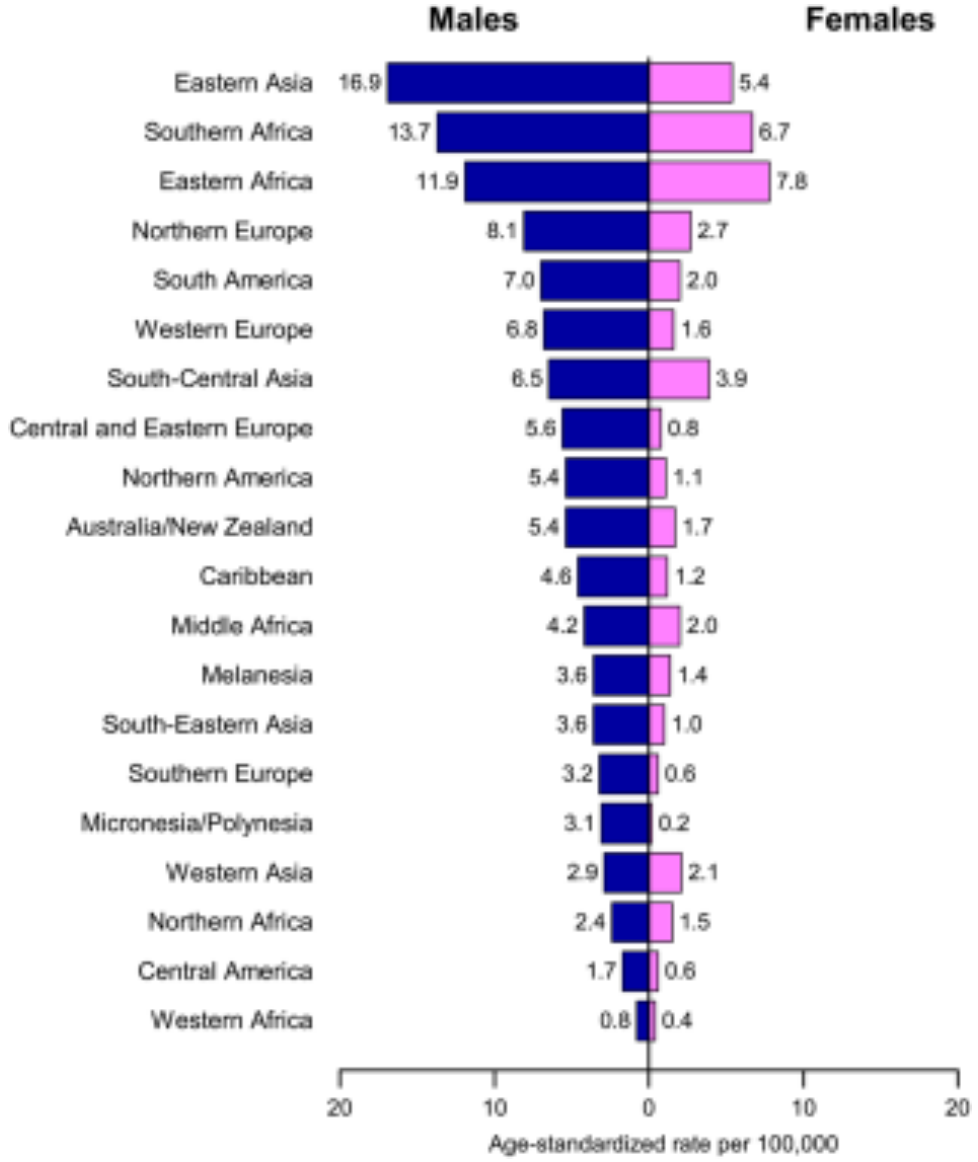
Ottman, Prev Med 1996

Poll: Why study Gene-Environment interactions?

Example: Esophageal cancer

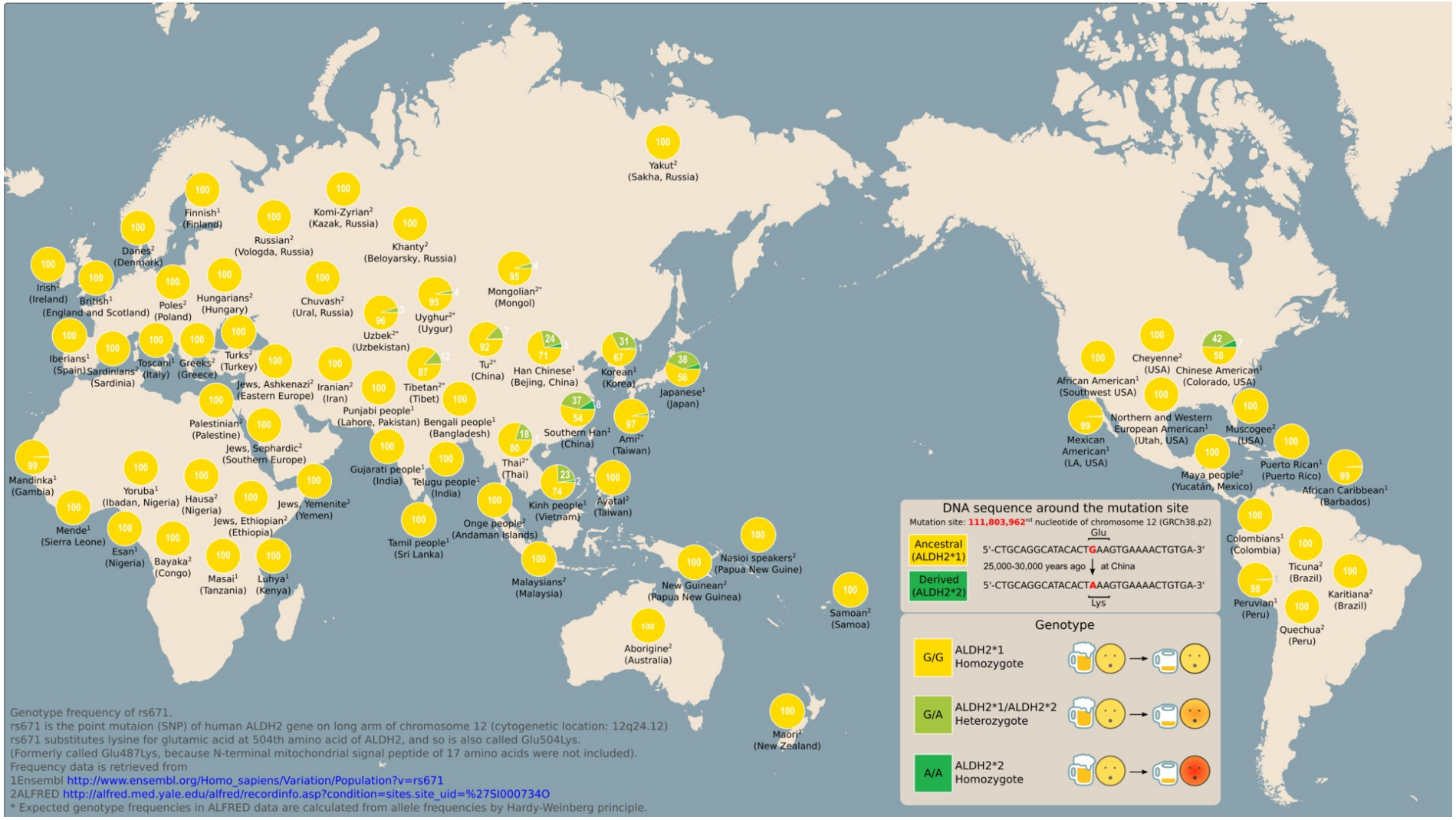
> Risk factors: alcohol intake, tobacco use, Barrett Syndrome, obesity





EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

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



EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

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

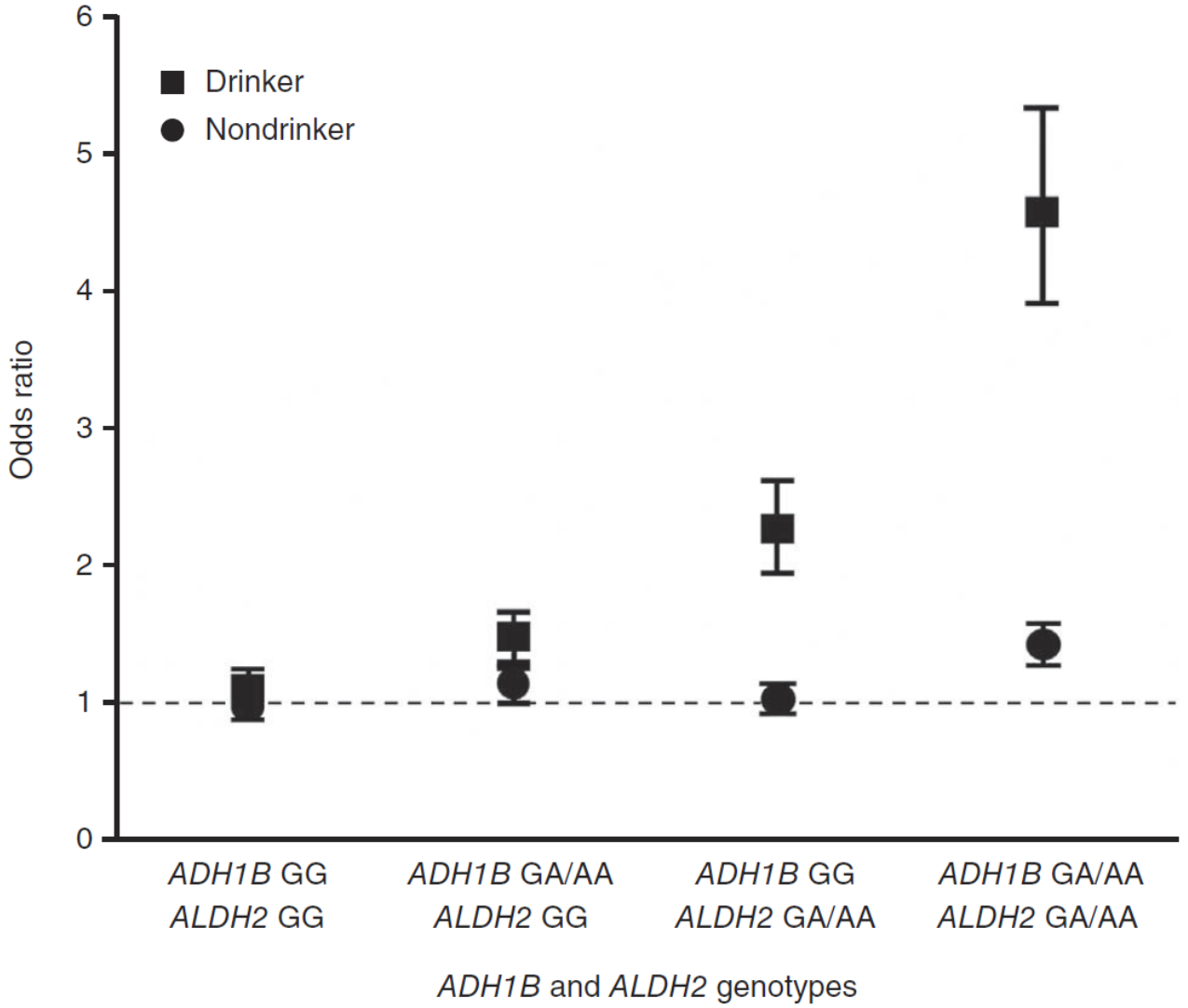


Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

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 main effect analysis

The 4-by-2 table: Interaction on the multiplicative scale

	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 people talk about interaction, they 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: $OR_{GE}=OR_GOR_E$ vs. H_A: $OR_{GE}\neq OR_GOR_E$

In practice, we often test for interaction on the multiplicative scale

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

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

Thomas, 2004

The joint interaction test – a tool for gene discovery

> Is the 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 ONLY

	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.

Assuming G and E are independent in the source population, then if G and E are associated in the cases, this 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!

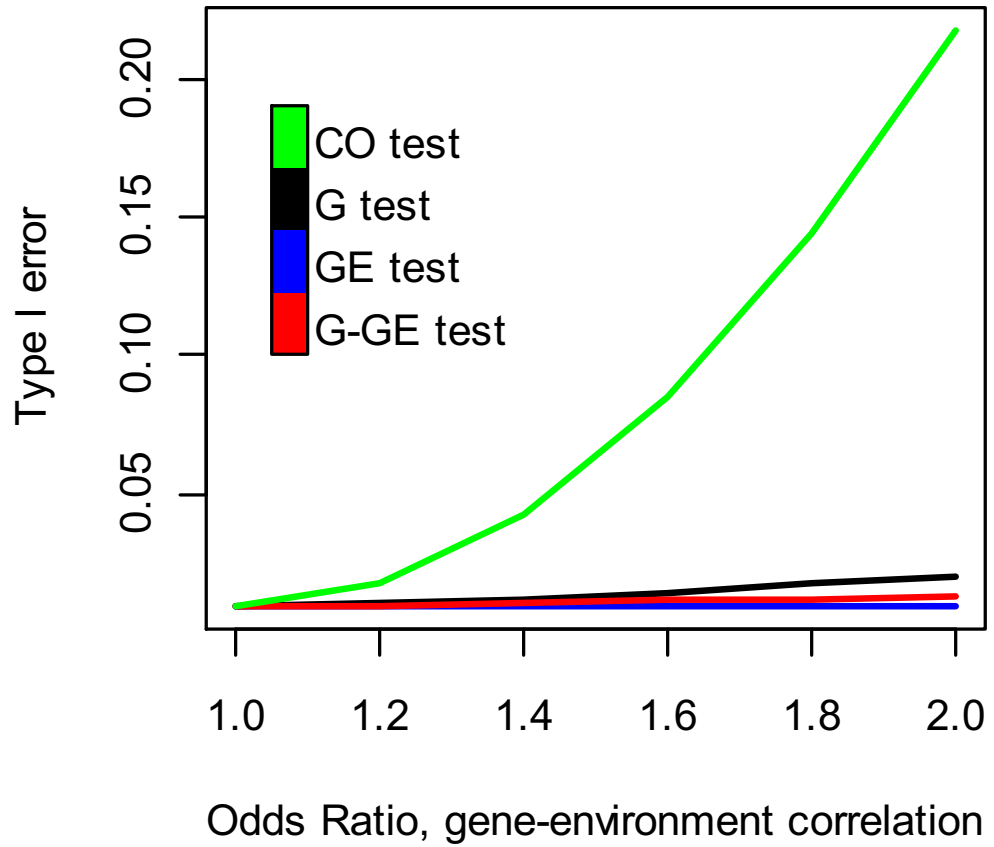
Piegorsch et al, Stat Med 1994

Does this mean I can throw away my controls?

> The increase in power is not due to the restriction to cases per se, rather the additional assumption of G-E independence (which you can test in your controls)

> Data on controls allow for estimation of G and E main effects in addition to the interaction effect and will also allow for calculation of joint G-E-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

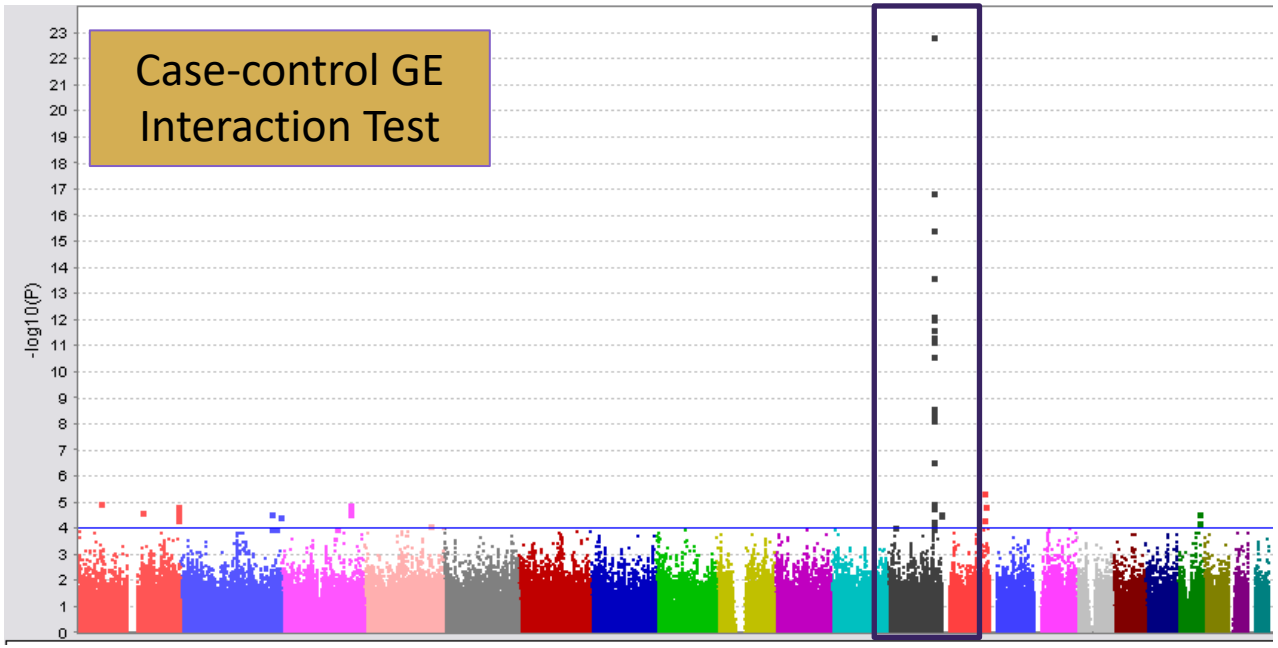
What if there is a negative correlation between G and E?

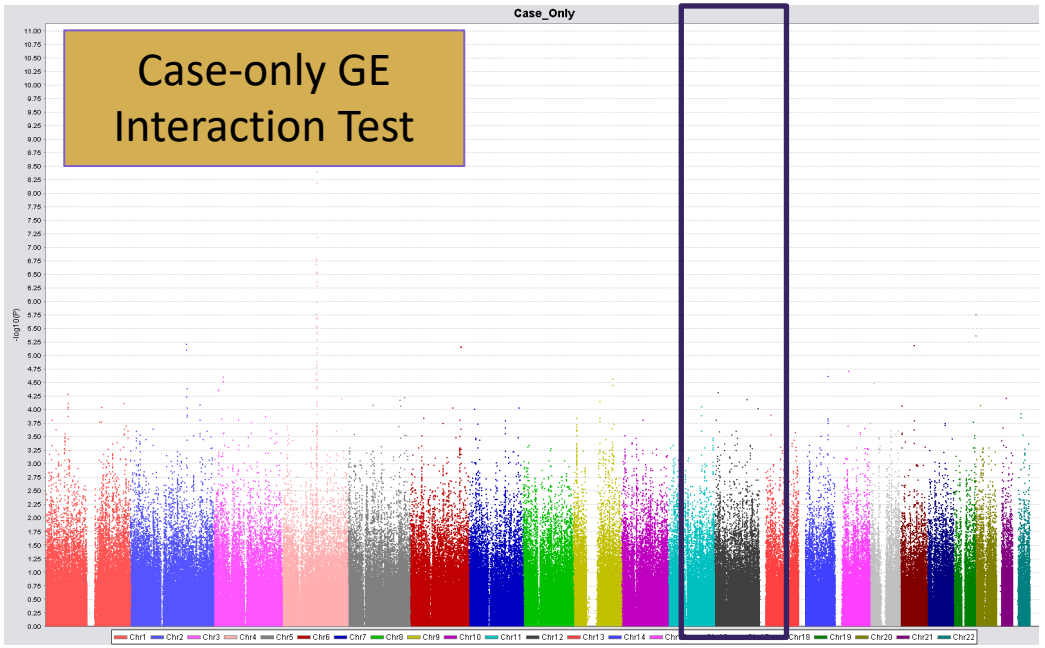
Esophageal cancer, *ALDH2* and Alcohol Intake

	OR_{E-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

Example: ESCC, *ALDH2* and Alcohol Intake





Courtesy of Chen Wu

Empirical Bayes Estimator

- If G and E are independent, do case-only test. If G and E are dependent, do GE interaction test in a case-control setting (1 df)

- Trade-off between bias and efficiency:

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

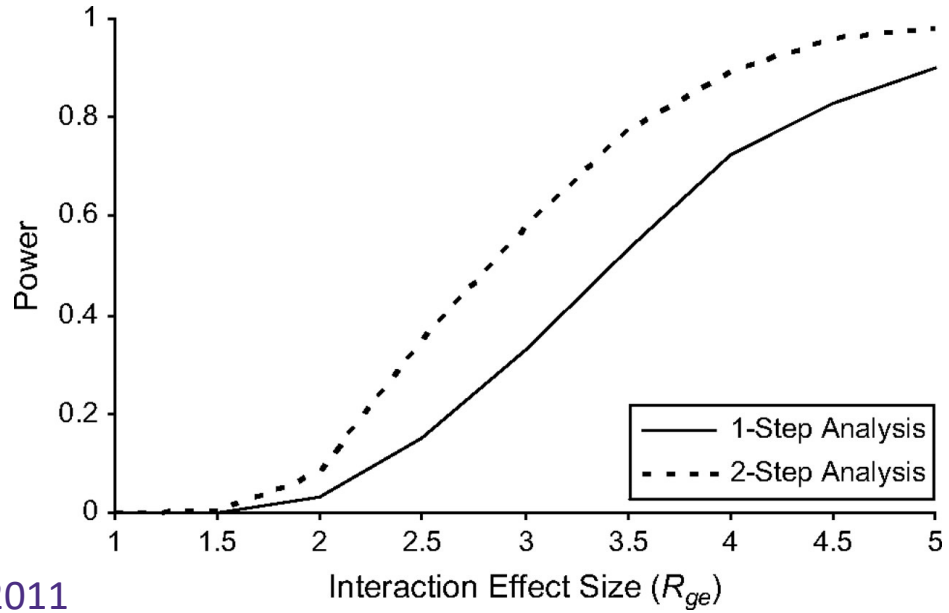
- $\hat{\tau}^2$ is an estimate of the GE dependence

Mukherjee & Chatterjee, Biometrics 2008

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

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

	Genome-wide Significant? ($\alpha=5\times10^{-8}$)	
	rs11066015[a]	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, th standard test of G×E interaction at the second stage did not meet the second stage threshold ($\sim4.2\times10^{-4}$); 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 ($\sim4.2\times10^{-4}$).

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 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 room activity

You are conducting a GE interaction study where the environmental exposure is smoking. Your colleagues have shared their data with you, which means you have 25,050 subjects in your study!

You need to harmonize the smoking variable across studies. The studies, their sample size and study-specific questions related to smoking can be found in the 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

Practical issues in GE interaction studies

> Measurement Errors

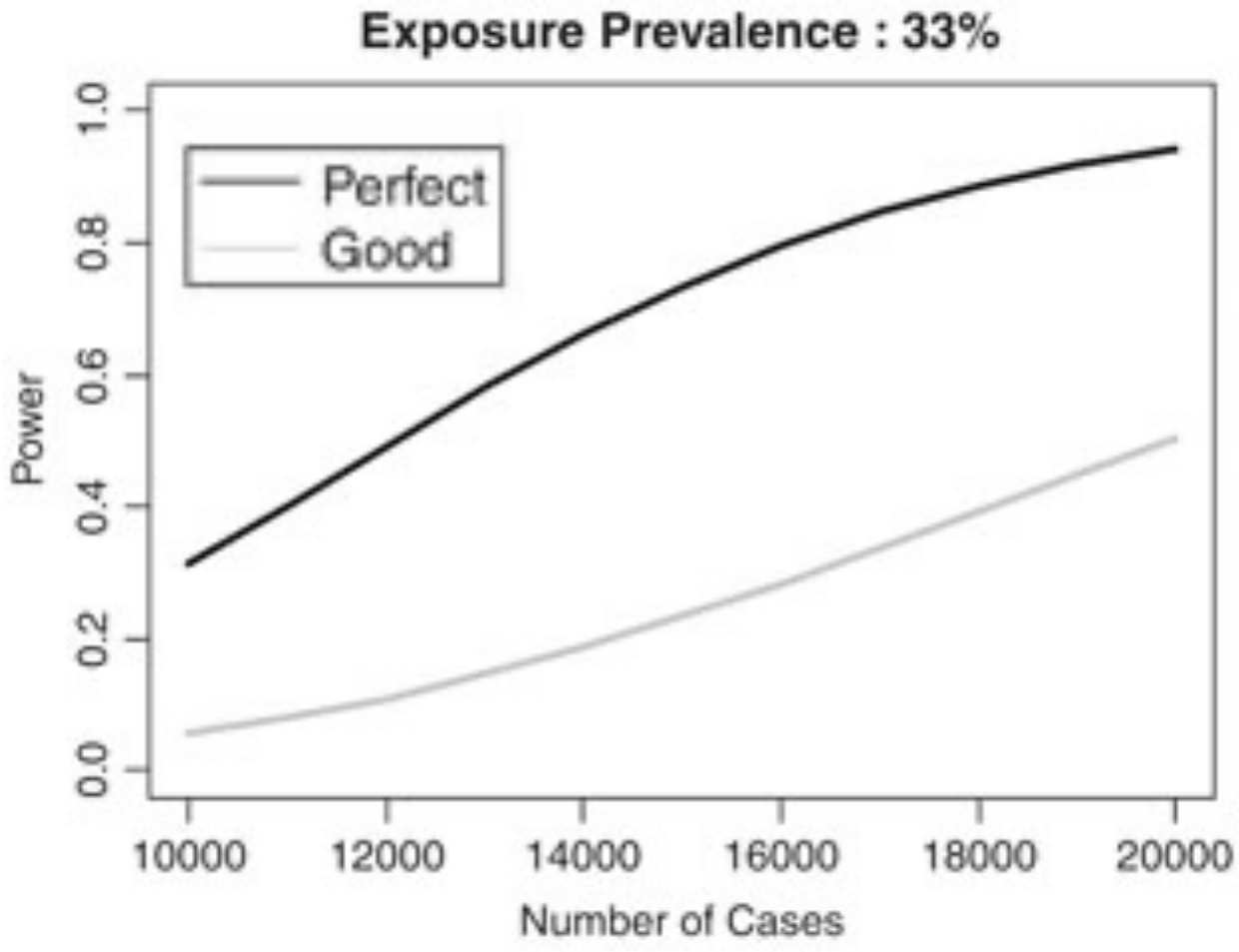
> Distribution of E/Replication

> Modeling E

> Software

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 = 0.0010

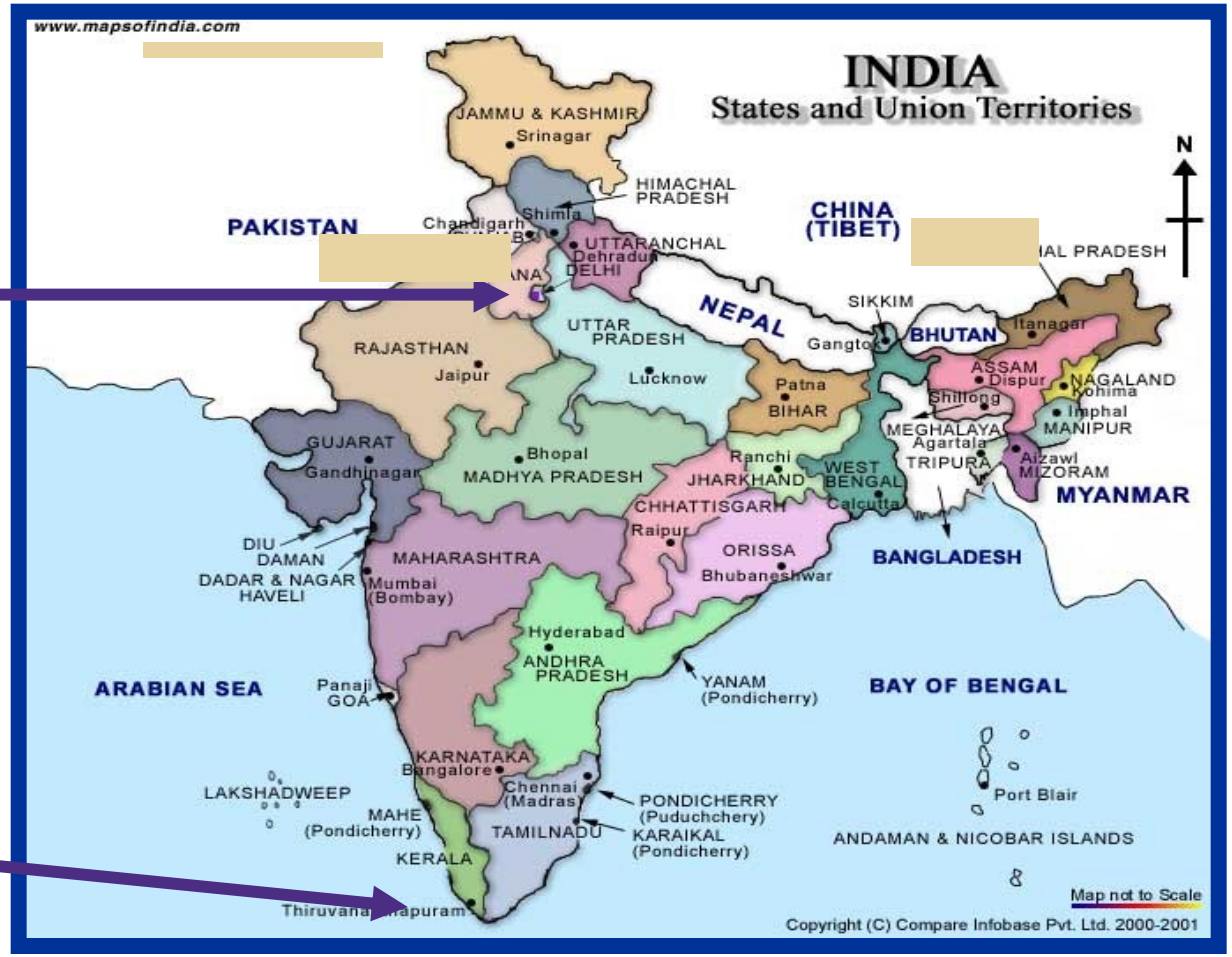
Kilpelainen, 2011

Slide courtesy of Leah Mechanic

The India Health Study

New Delhi

Trivandrum



Moore, Obesity 2012

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 between *FTO* SNP and obesity

	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

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

Association between *FTO* SNP and obesity by physical activity

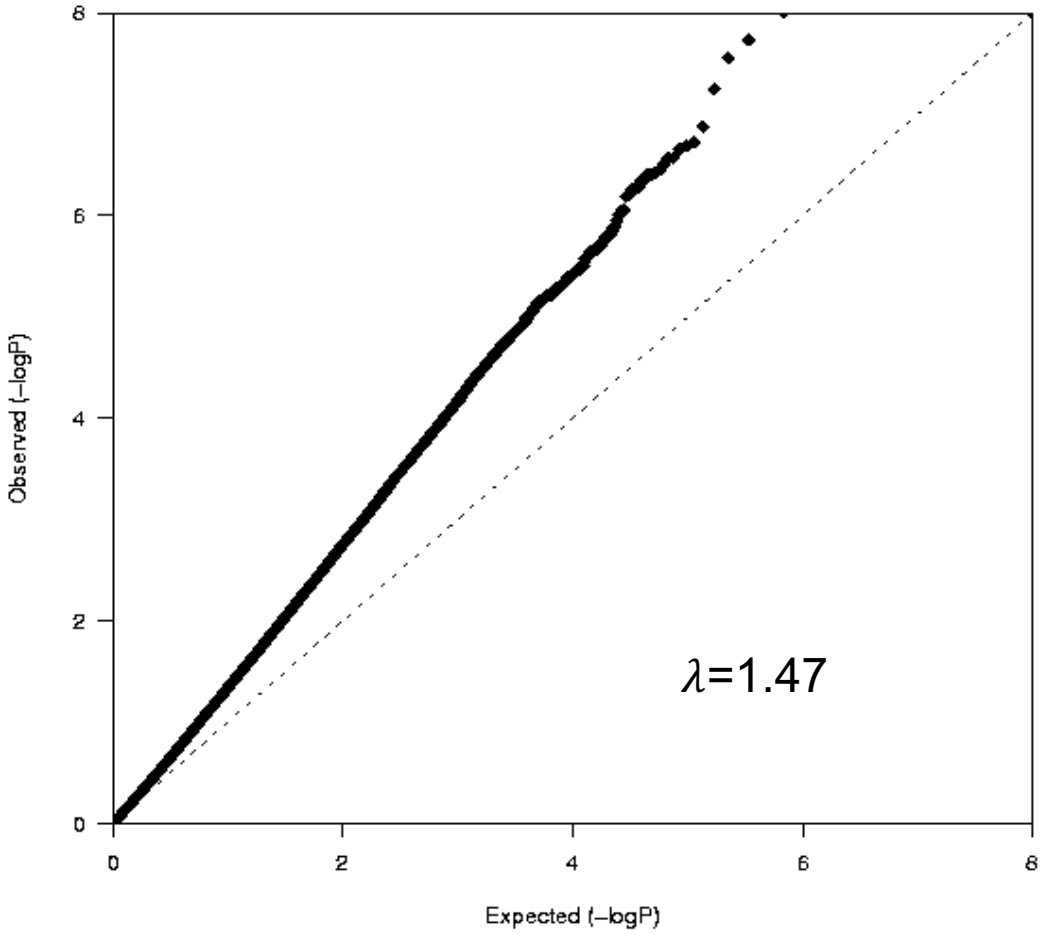
	N	Effect size per T allele (95% CI)	P_{trend}	P_{Int}
Overall	**1,209**	**+1.61 cm (0.67, 2.55)**	**0.0008**	**0.009**
New Delhi	**578**	**+2.53 cm (1.08, 3.97)**	**0.0006**	**0.59**
By PA				
≤ 91 MET-hrs/wk	517	+2.36 cm (0.82, 3.89)	0.003	
92-151 MET-hrs/wk	32	+6.39 cm (1.94, 10.85)	0.005	
152-217 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

EPIDEMIOLOGY
SCHOOL OF PUBLIC HEALTH

A note about modeling “E” (i)

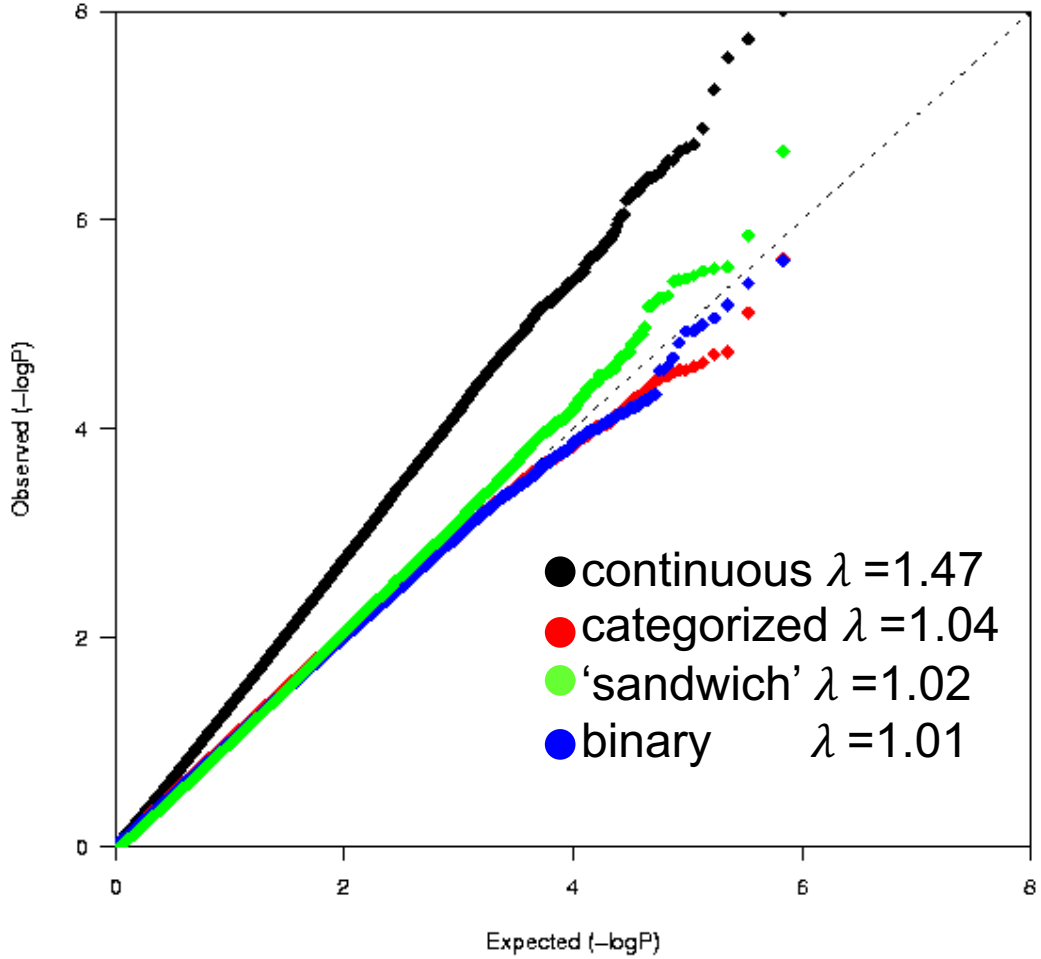
- Genome-wide GE interaction study of BMI and Type II Diabetes
- Standard 1 df case-control test for GE interaction



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