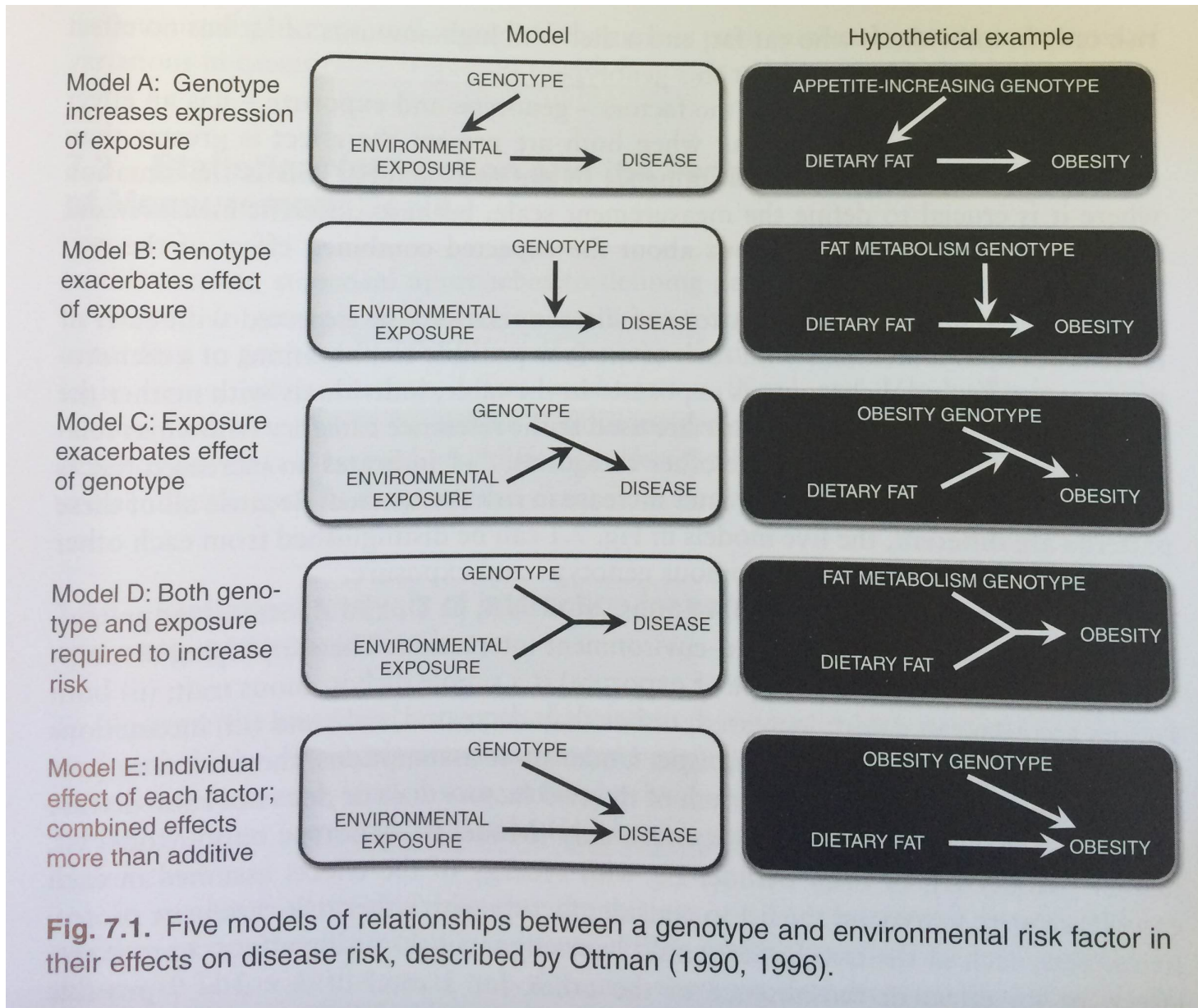


# Gene-Environment Interactions

# 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



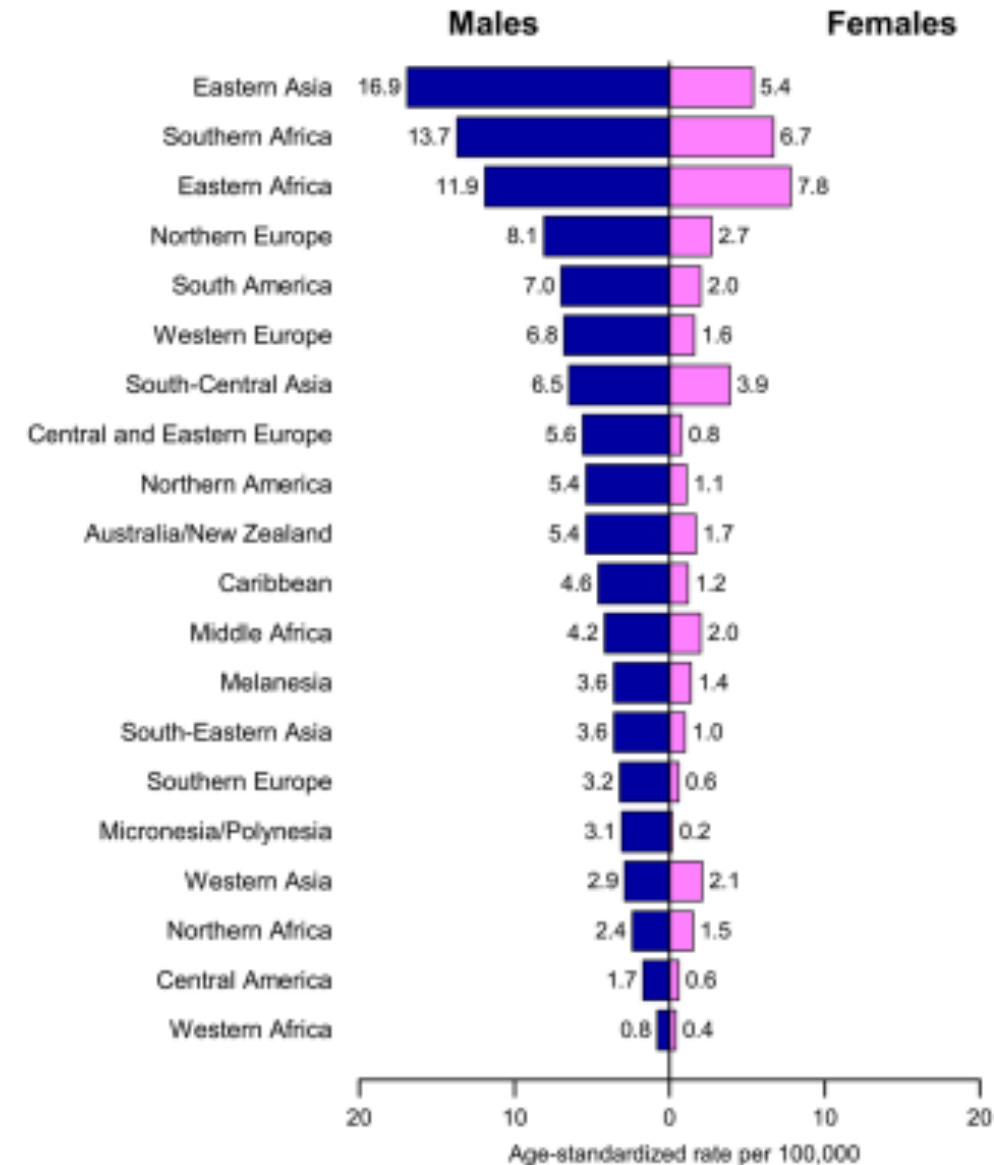
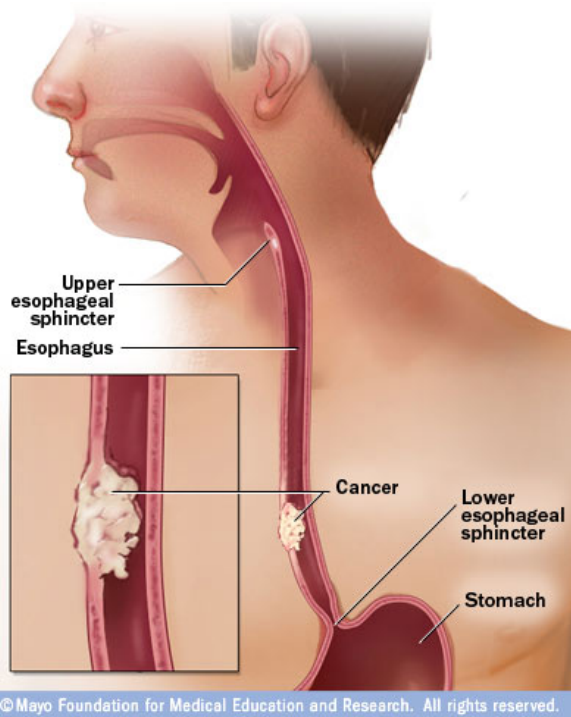
**Fig. 7.1.** Five models of relationships between a genotype and environmental risk factor in their effects on disease risk, described by Ottman (1990, 1996).

# Why study Gene-Environment Interactions?

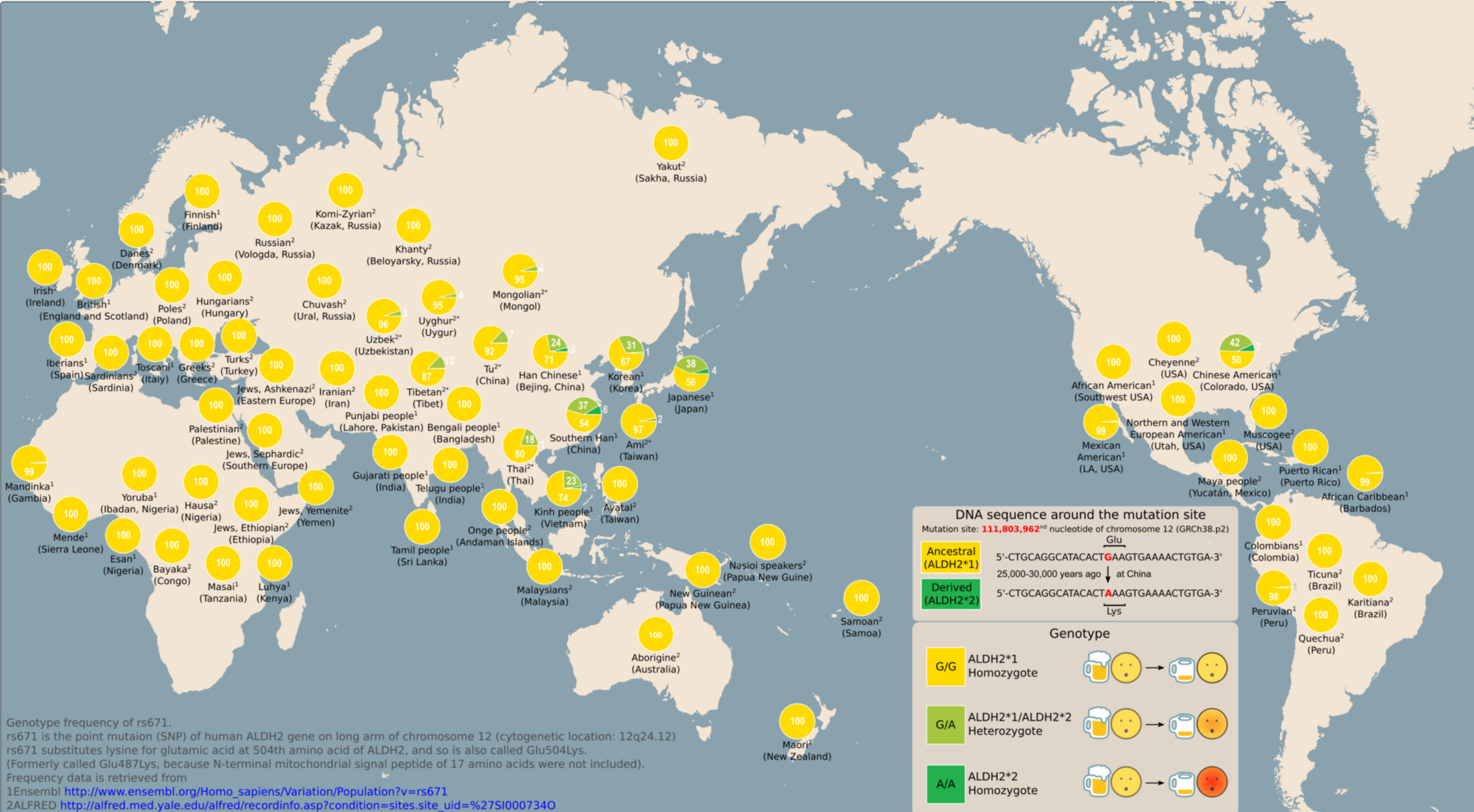
- Gain insights about already known genes
  - Information about effect in different strata might give insights in pathways and biology
- Clinical Importance
  - Disease prediction, pharmacogenetics
- A tool in gene discovery
  - Gene only affective in exposed individuals - Environment only affective in gene carriers
  - Incorporating GxE interactions may boost power in association analysis

# Example: Esophageal cancer

- Risk factors: alcohol intake, tobacco use, being a man, Barrett Syndrome, obesity







Genotype frequency of rs671.  
 rs671 is the point mutation (SNP) of human ALDH2 gene on long arm of chromosome 12 (cytogenetic location: 12q24.12)  
 rs671 substitutes glutamic acid 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](http://www.ensembl.org/Homo_sapiens/Variation/Population?v=rs671)  
 2ALFRED [http://alfred.med.yale.edu/alfred/recordinfo.asp?condition=sites.site\\_uid=%27SI0007340](http://alfred.med.yale.edu/alfred/recordinfo.asp?condition=sites.site_uid=%27SI0007340)  
 \* Expected genotype frequencies in ALFRED data are calculated from allele frequencies by Hardy-Weinberg principle.

**DNA sequence around the mutation site**  
 Mutation site: **111,803,962<sup>nd</sup>** nucleotide of chromosome 12 (GRCh38.p2)

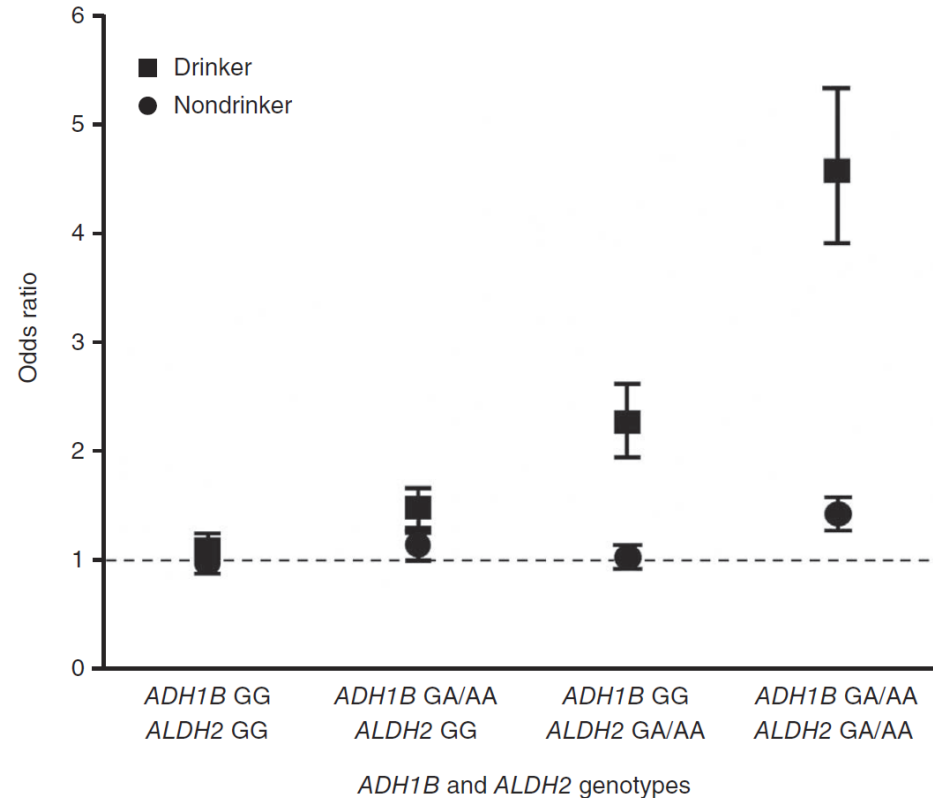
**Ancestral (ALDH2\*1)** 5'-CTGCAGGCATACACT **G**AAGTGAAAACGTGA-3'  
 25,000-30,000 years ago ↓ at China

**Derived (ALDH2\*2)** 5'-CTGCAGGCATACACT **A**AAGTGAAAACGTGA-3'

**Genotype**

<b>G/G</b> ALDH2*1 Homozygote	
<b>G/A</b> ALDH2*1/ALDH2*2 Heterozygote	
<b>A/A</b> ALDH2*2 Homozygote	

# 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).



# 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

# Non-parametric analysis: The 4-by-2 table

	Case	Control
G=0,E=0	$N_{100}$	$N_{000}$
G=1,E=0	$N_{110}$	$N_{010}$
G=0,E=1	$N_{101}$	$N_{001}$
G=1,E=1	$N_{111}$	$N_{011}$

This presentation is “closest to the data” and makes no assumption about genetic model or how the gene and exposure jointly influence risk

For prospective data, yields estimates of relative risks.

For retrospective data, yields estimates of odds ratios.

For rare SNPs or exposures, the GxE-stratified estimates of risks/odds ratios from this table can be very noisy

# 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

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

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

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



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

# Test for Interaction (jointly) – a tool for gene discovery

- Is this gene associated with disease risk in any of the exposure sub-groups?
- Compare “main effect of E only” model to “main effects plus interaction” model in a 2 df test.

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

to

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

# 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.

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 correlated, there is an “interaction.”)

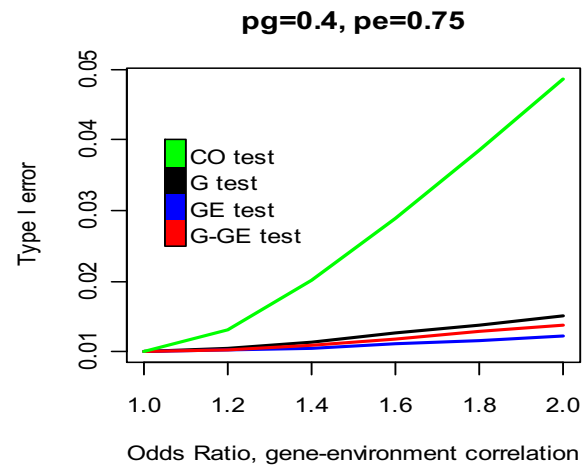
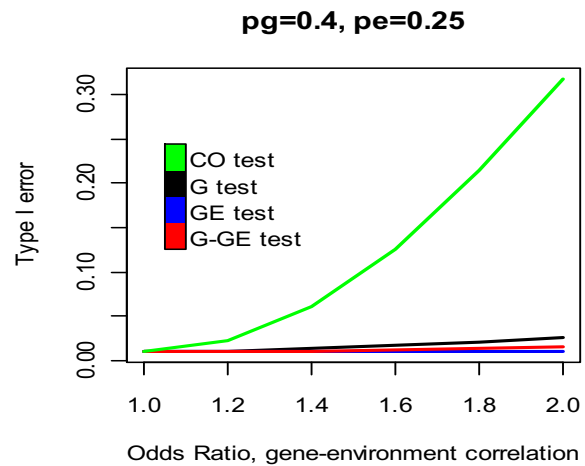
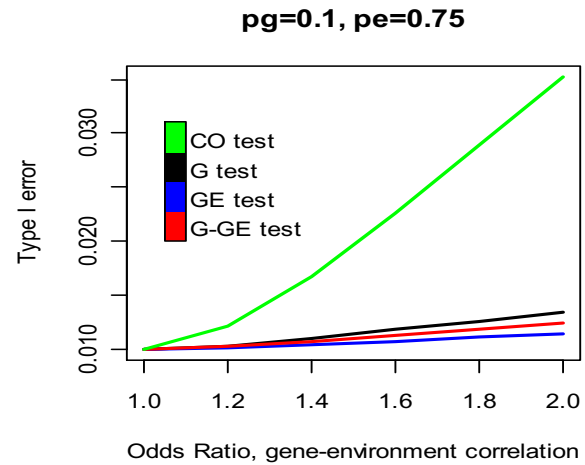
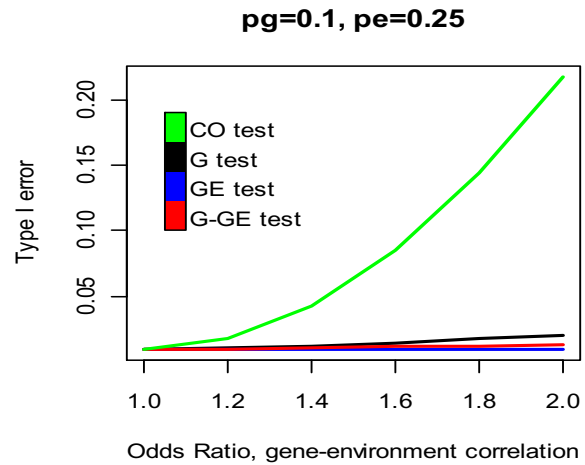
Can be much more powerful than traditional logistic regression analysis!

GREAT!!

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

- Well, 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)
- 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.

Sens= 0.6, Spec = 0.9,  
OR(E)= 1.6



Case-only analysis produces inflated results when there is  
a positive correlation between G and E

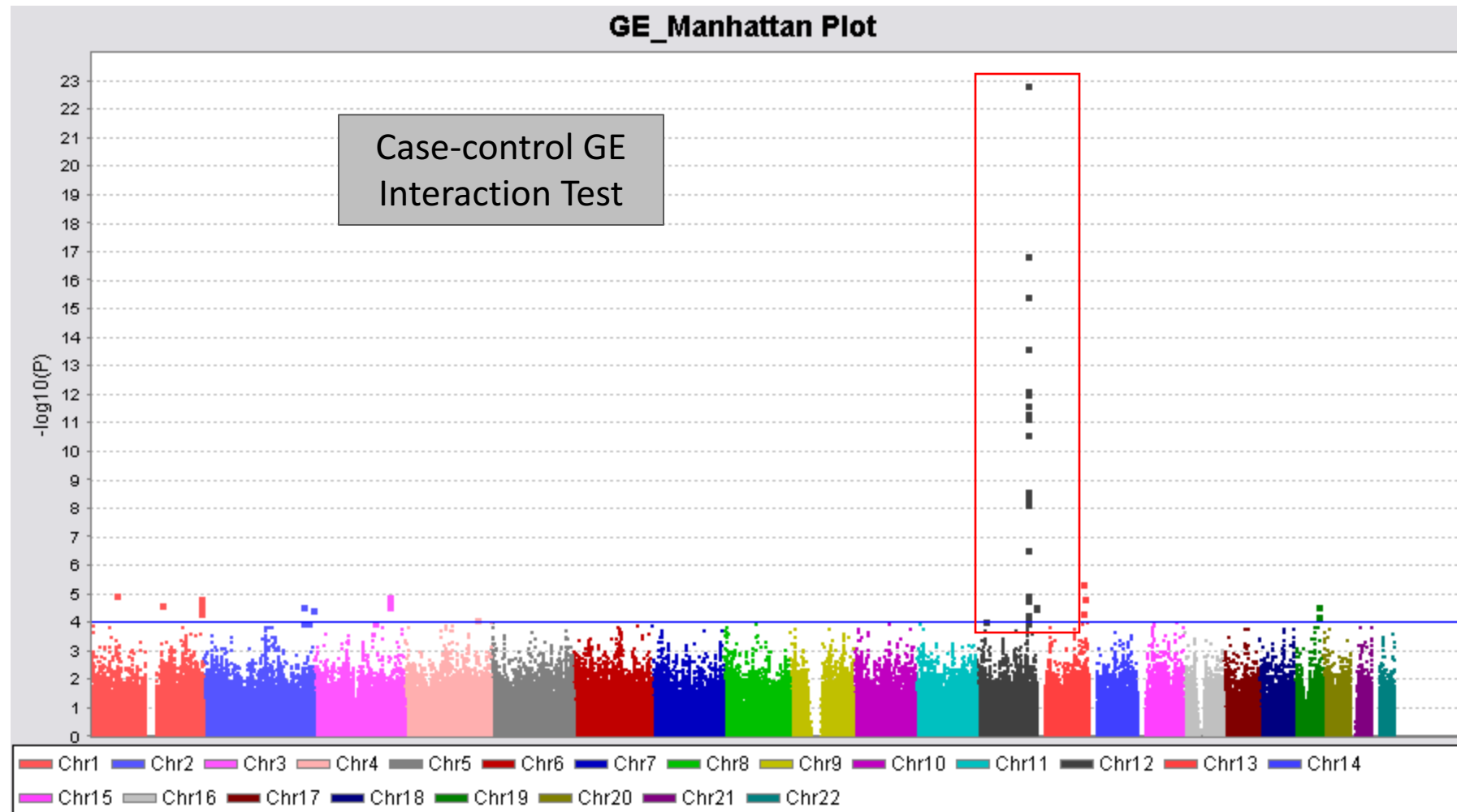
What if there is a negative correlation between G and E?

# Example: ESCC, *ALDH2* and Alcohol Intake

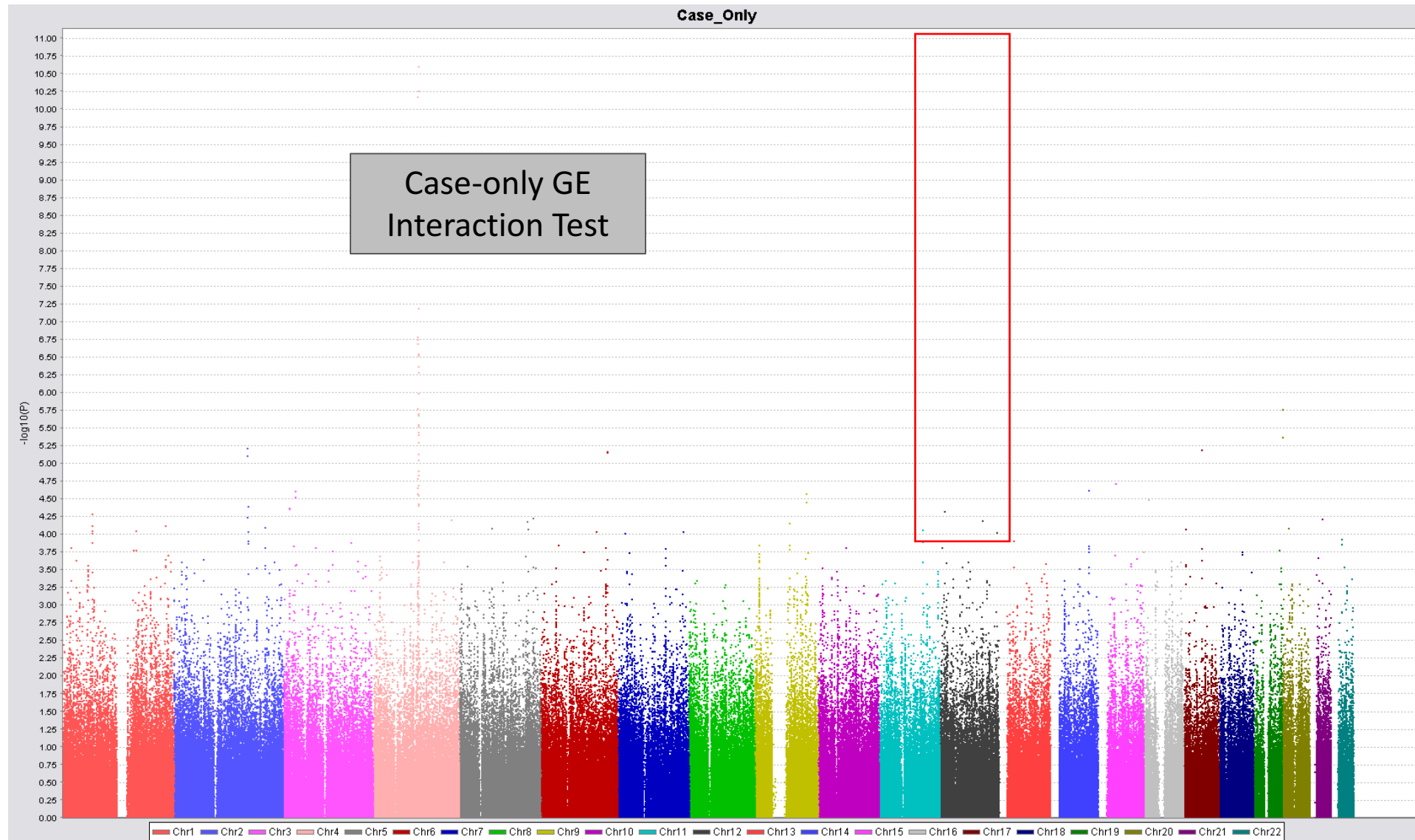
The risk allele is associated with a decreased risk of heavy drinking in the general population, and an increase in the effect of alcohol on ESCC risk

	$OR_{E-G}$	$OR_{GxE}$
rs670 ( <i>ALDH*2</i> )	0.23	2.69

# Example: ESCC, *ALDH2* and Alcohol Intake



# Example: ESCC, *ALDH2* and Alcohol Intake



# Empirical Bayes Estimator

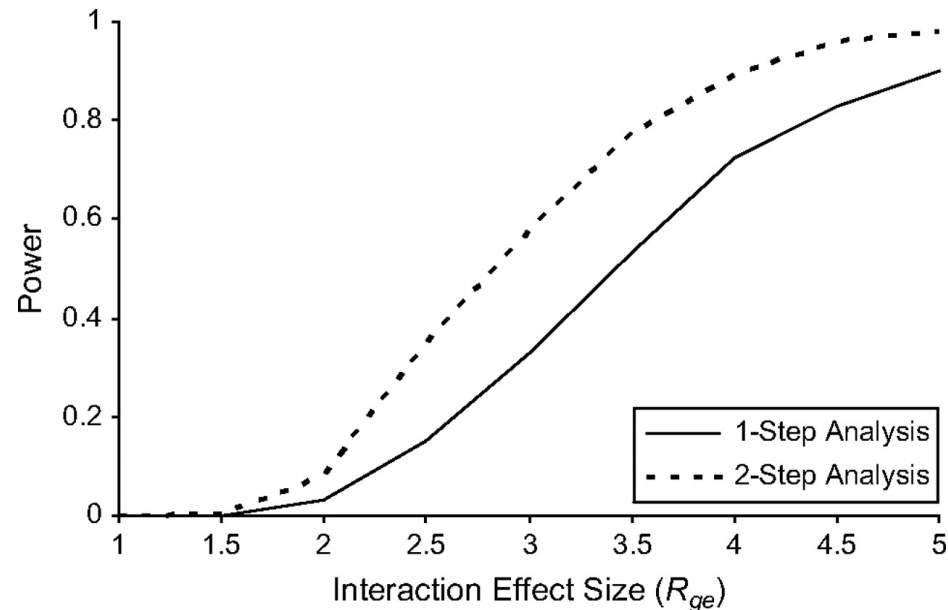
- If G and E are independent– Case-only test. Otherwise GxE interaction tests in a case-control setting (1 df)
- Trade-off between bias and efficiency:

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

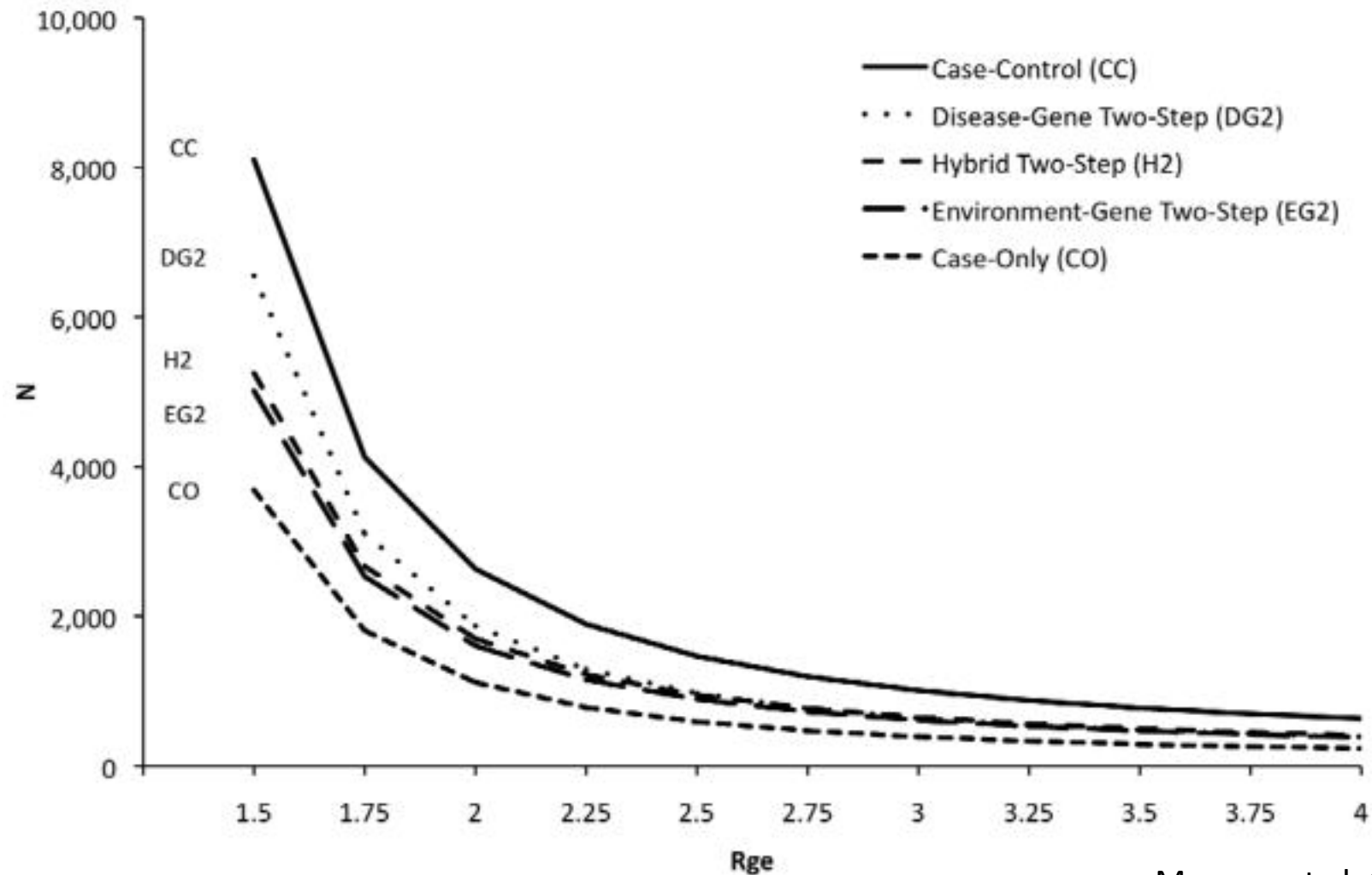
- $\hat{\tau}^2$  is an estimate of the G-E dependence  $\theta_{GE}^2$

# Genome-wide G-E Interaction analysis: 2-step approaches

1. Test for GxE dependence and/or associations between the SNP and your outcome in your entire dataset. Select SNPs with  $p < \alpha_1$
2. Take  $m$  SNPs from stage 1 and perform traditional GxE interaction tests in a case-control setting (1 df). All SNPs with  $p < \alpha/m$  are declared significant



# Sample Size Required to Achieve 80% Power for Tests of Gene-Environment Interaction in a Genome-wide Association Study by Interaction Effect Size for Binary Environmental Exposure



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

	Genome-wide Significant?	
	<i>ALDH2</i> rs11066015 <sup>a</sup>	( $\alpha=5\times 10^{-8}$ ) <i>ADH</i> rs3805322 <sup>b</sup>
Standard case-control test	<b>Yes</b>	no
Case-only test	No	<b>Yes</b>
Empirical Bayes test	<b>Yes</b>	no
Hybrid two-step approach	<b>Yes</b>	no
Cocktail 1	<b>Yes</b>	<b>Yes</b>
Cocktail 2	<b>Yes</b>	<b>Yes</b>

<sup>a</sup> 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  $p_A=6.0\times 10^{-14}<\alpha_A$  and  $p_M=7.3\times 10^{-8}<\alpha_M$ , and the standard test of G×E interaction at the second stage was quite significant ( $p<10^{-16}$ ); for the cocktail methods,  $p^{\text{screen}}=p_M$  for cocktail 1 and  $p^{\text{screen}}=p_A$  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 ( $p<10^{-16}$ ).

<sup>b</sup> Empirical Bayes estimate of  $OR_{G\times E}=1.70$  (1.36,2.20),  $p=5.4\times 10^{-5}$ ; for the screening stage of the hybrid test, both G-E association and marginal G-D tests were significant with  $p_A=1.1\times 10^{-9}<\alpha_A$  and  $p_M=9.3\times 10^{-13}<\alpha_M$ , however, the standard test of G×E interaction at the second stage did not meet the second stage threshold ( $\sim 4.2\times 10^{-4}$ ); for the cocktail methods,  $p^{\text{screen}}=p_M$  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 ( $\sim 4.2\times 10^{-4}$ ).





# GxE 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 a multi-study GxE interaction study

# Harmonizing E

(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

# Harmonizing E

**What are the sample sizes for these derived variables?**

- Cigarettes per day
- Packs per day
- Former smoker
- Ever smoker
- Current smoker

**Table II. Examples of possible (a) smoking-related questions and (b) new variables for cross-study analyses**

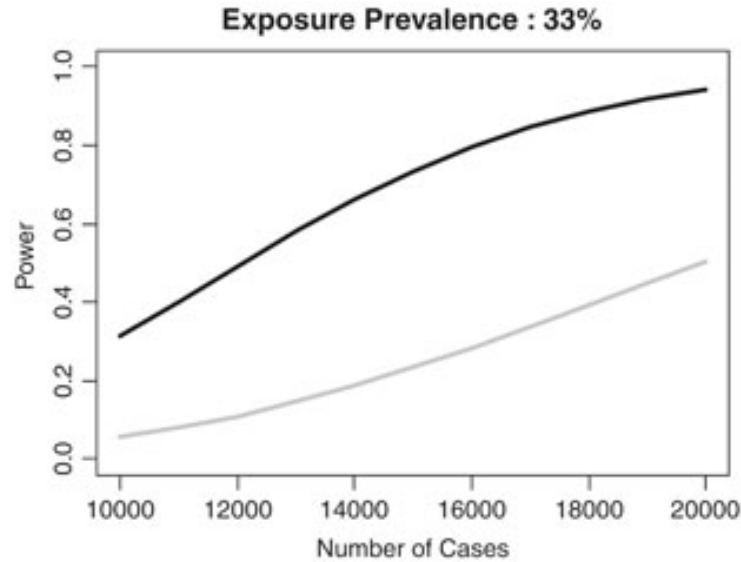
(a)		
Study (N)	Smoking-related questions	Possible responses
Study 1 (2,500)	1. Do you currently smoke cigarettes?	Y/N
	2. If yes, how many cigarettes per day?	###
Study 2 (1,200)	1. Have you smoked more than 100 cigarettes in your lifetime?	Y/N
	2. If yes, do you currently smoke?	Y/N
	3. If yes, how 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

# Harmonizing E

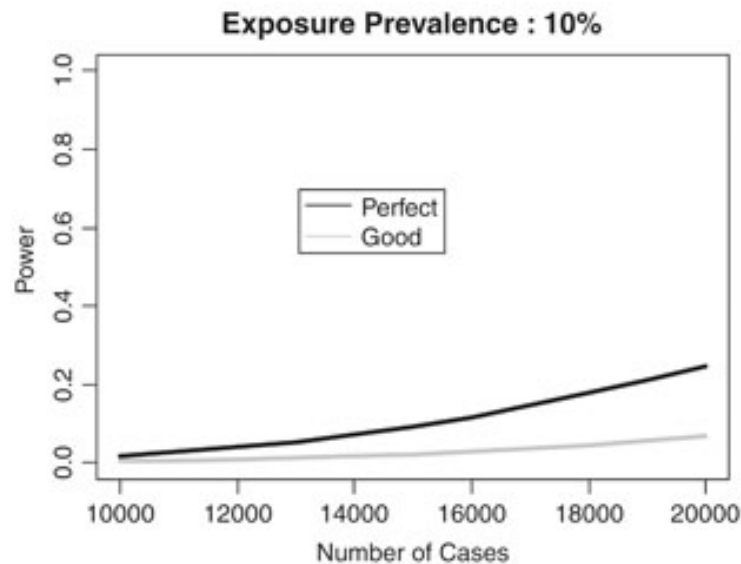
**Table II. Examples of possible (a) smoking-related questions and (b) new variables for cross-study analyses**

(a)			(b)			
Study (N)	Smoking-related questions	Possible responses	New variable	Studies that could contribute data	Total N	Comment
Study 1 (2,500)	1. Do you currently smoke cigarettes?	Y/N	Cigarettes per day	Study 1	2,500	Data from Study 7 might also be included if specific values were assigned to each response category, e.g. 5 for category '1–10 cigarettes', 15 for category '11–20 cigarettes', and so on
	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	Packs per day	Study 1 (if convert cigarettes/day to packs/day) Study 2 Study 7 (if convert categories to packs/day)	4,500	
	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	Former smoker	Study 2 Study 7	2,000	
Study 4 (1,250)	1. Do you currently smoke?	Y/N	Ever smoker	Study 2 Study 3 Study 7	10,500	Requires ability to determine if subjects are former smokers
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	Current smoker	Study 1	16,550	
Study 6 (6,600)	1. Have you smoked tobacco in the past month?	Y/N		Study 2 Study 4 Study 5		
Study 7 (800)	1. Have you ever smoked regularly?	Y/N		Study 6 (if current smoker is defined as having smoked in the past month) Study 7		
	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				

# Even small errors in measurement can greatly decrease power to detect gene-environment interaction

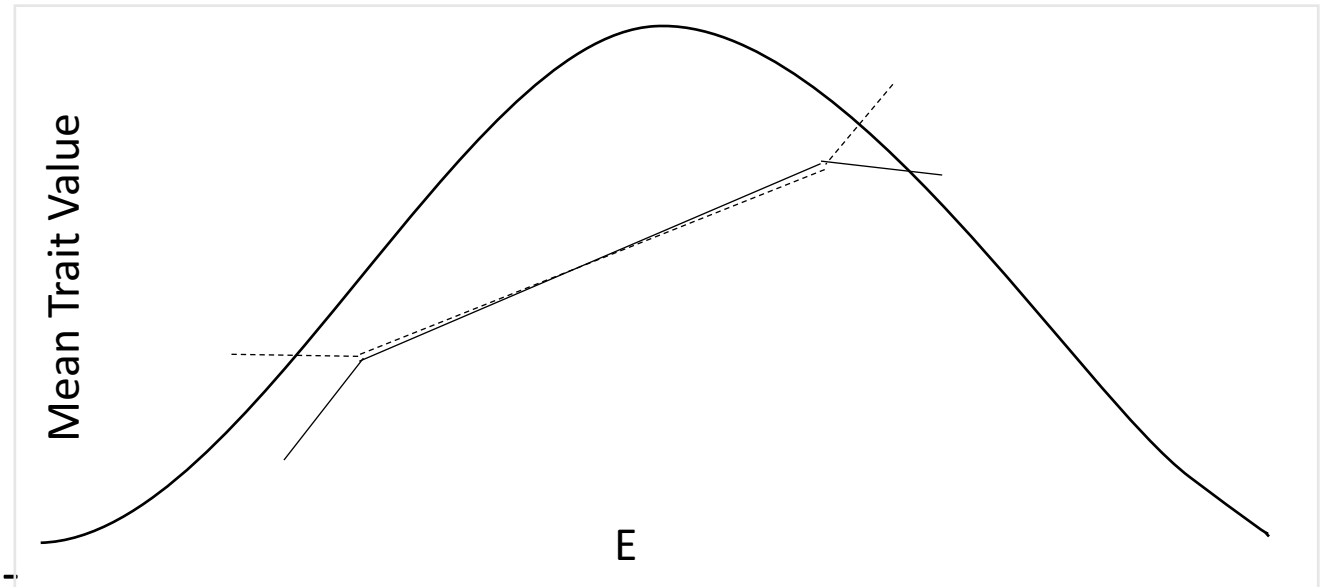


“Good”  
Sensitivity=77%  
Specificity=99%



# How, where, and when you measure exposure have consequences for evaluating gene-environment interactions

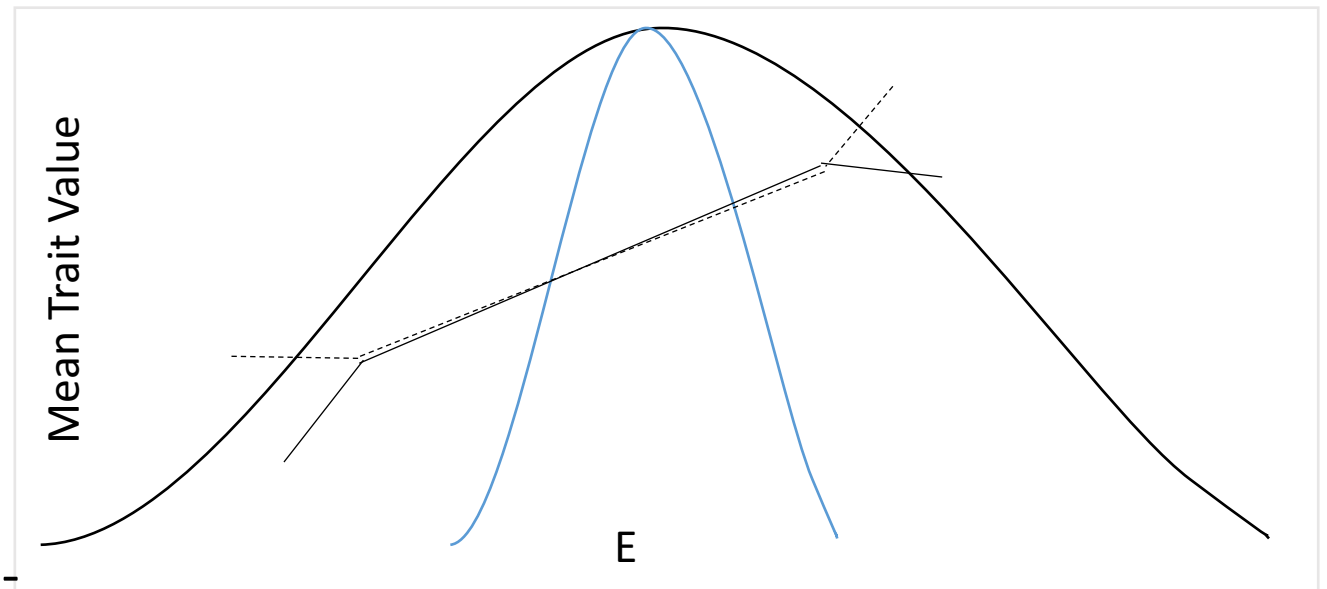
- Issues with replication of GxE interactions
  - E: Different distribution of the exposure across studies
  - G: Different LD patterns, different allele frequencies, (unknown) gene-environment interactions
  - Phenotype: E.g. breast cancer – Estrogen Receptor positive vs. negative



G=0 ———  
G=1 - - - -  
Discovery Sample ———

# How, where, and when you measure exposure have consequences for evaluating gene-environment interactions

- Issues with replication of GxE interactions
  - E: Different distribution of the exposure across studies
  - G: Different LD patterns, different allele frequencies, (unknown) gene-environment interactions
  - Phenotype: E.g. breast cancer – Estrogen Receptor positive vs. negative



G=0 ———  
G=1 - - - -  
Discovery Sample ———  
Replication Sample ———

# *FTO*, Physical Activity and Obesity

---

- Meta-analysis of 218,166 European-ancestry subjects
- Risk of Obesity (BMI  $\geq 30$  vs. BMI  $< 25$  kg/m<sup>2</sup>) for *FTO* 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)
	<i>P-value</i> = 0.0010



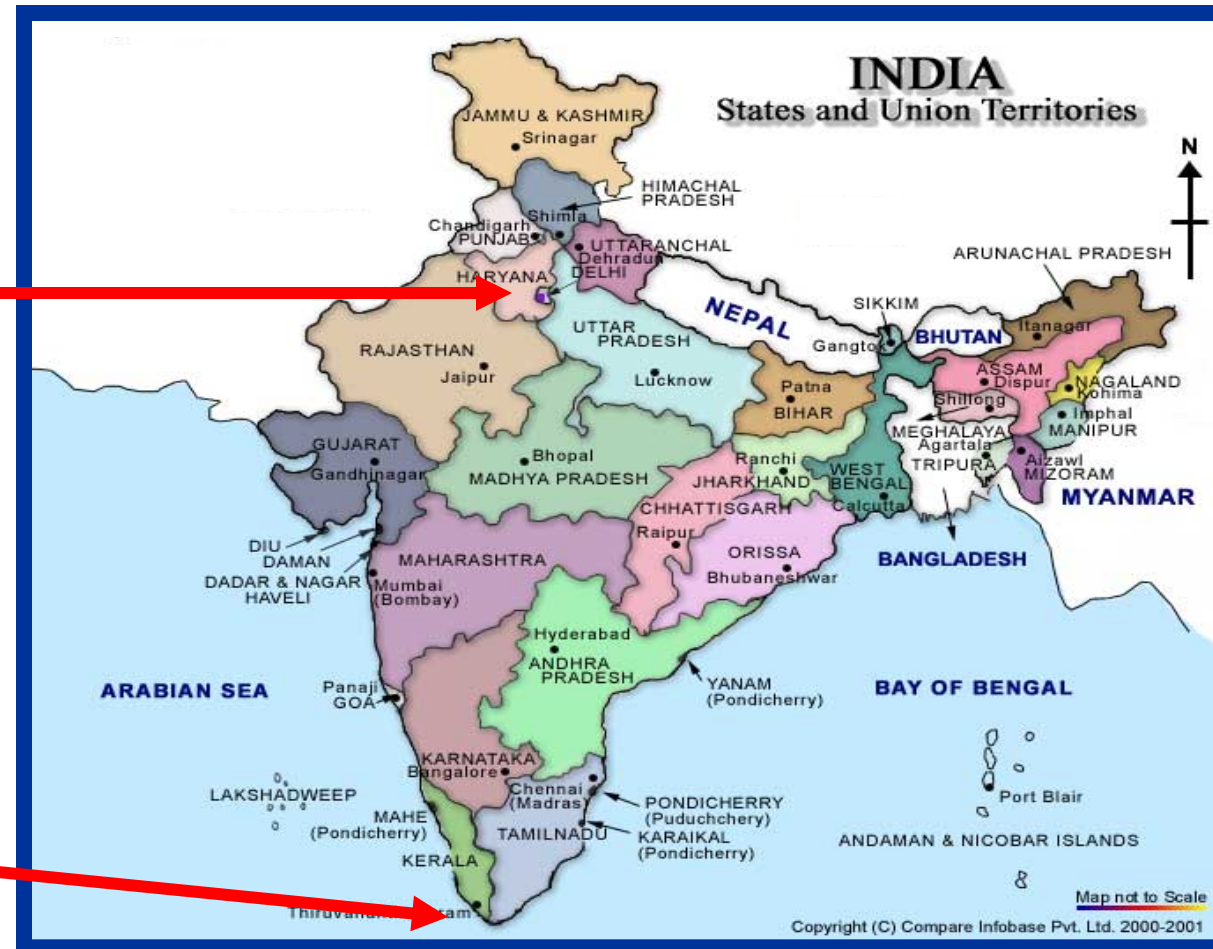
# India health study

Interaction between FTO genotype,  
physical activity and obesity

New Delhi



Trivandrum



## Participant characteristics by region

Characteristic	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

## Association of *FTO* rs3751812 with waist circumference

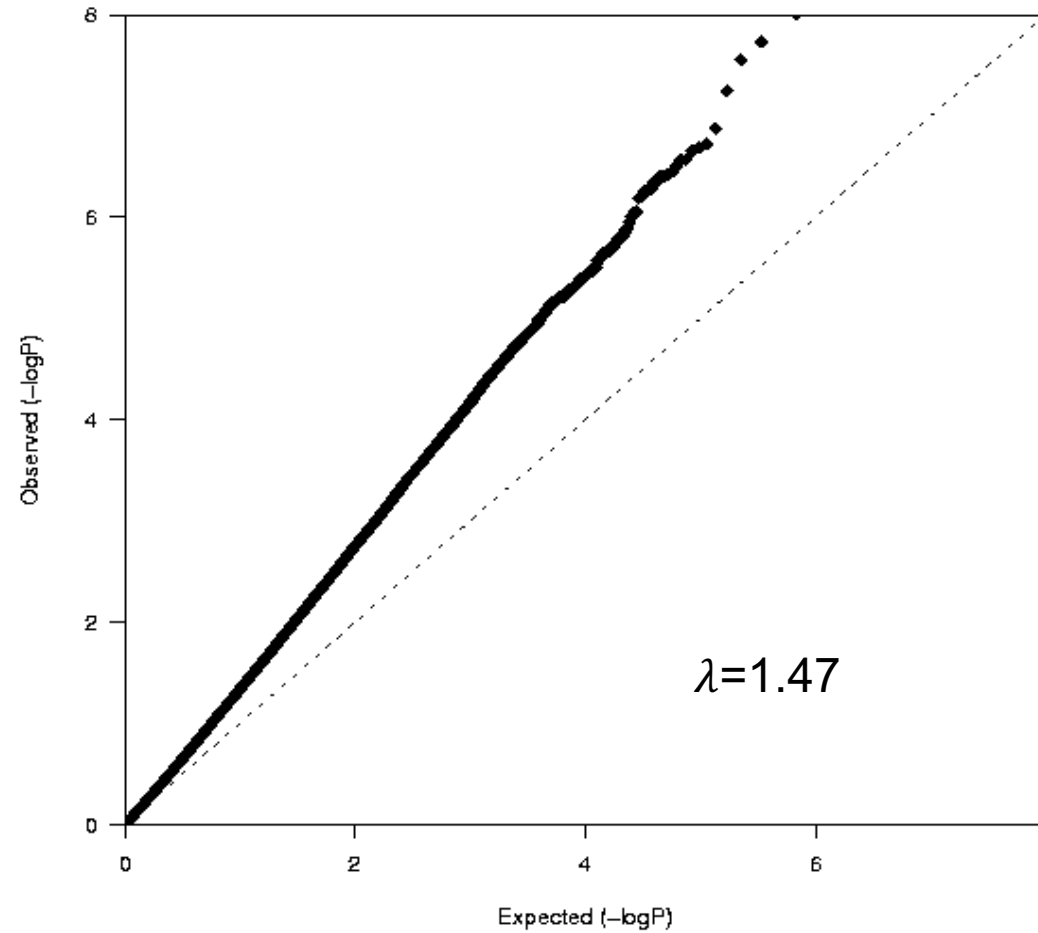
Characteristic	N	Effect size per T allele (95% CI)	P <sub>trend</sub>	Interaction by PA
Overall	1,209	+1.61 cm (0.67, 2.55)	0.0008	<b>0.009</b>
New Delhi				
Overall	578	+2.53 cm (1.08, 3.97)	0.0006	<b>0.59</b>
By PA				
≤ 91 MET-hrs/wk	517	+2.36 cm (0.82, 3.89)	0.003	
92-151 MET-hrs/wk	32	+6.39 cm (1.94, 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				
Overall	574	+0.87 cm (-0.35, 2.08)	0.16	<b>0.004</b>
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, 2012

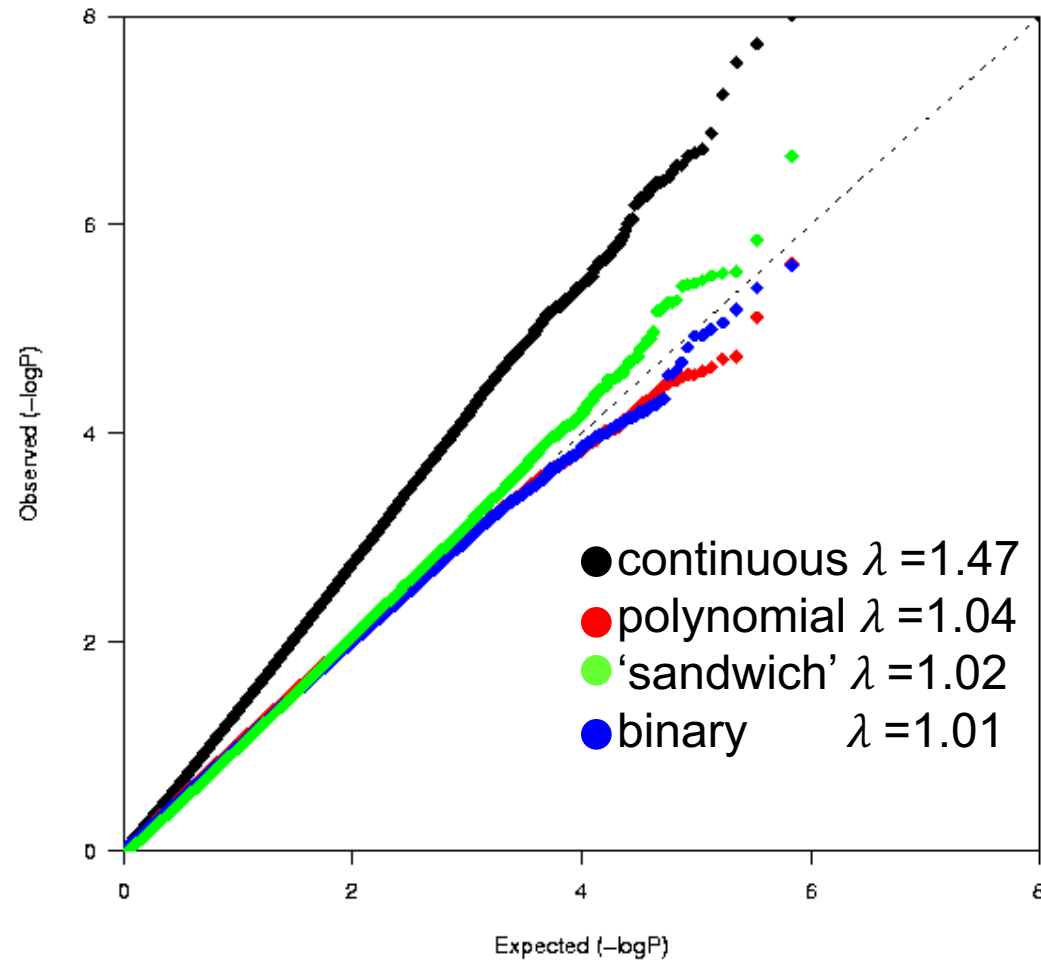
# A note about modeling “E”

Genome-wide GxE Interaction  
study of BMI and Type II  
Diabetes

Standard case-control test for  
GxE Interaction



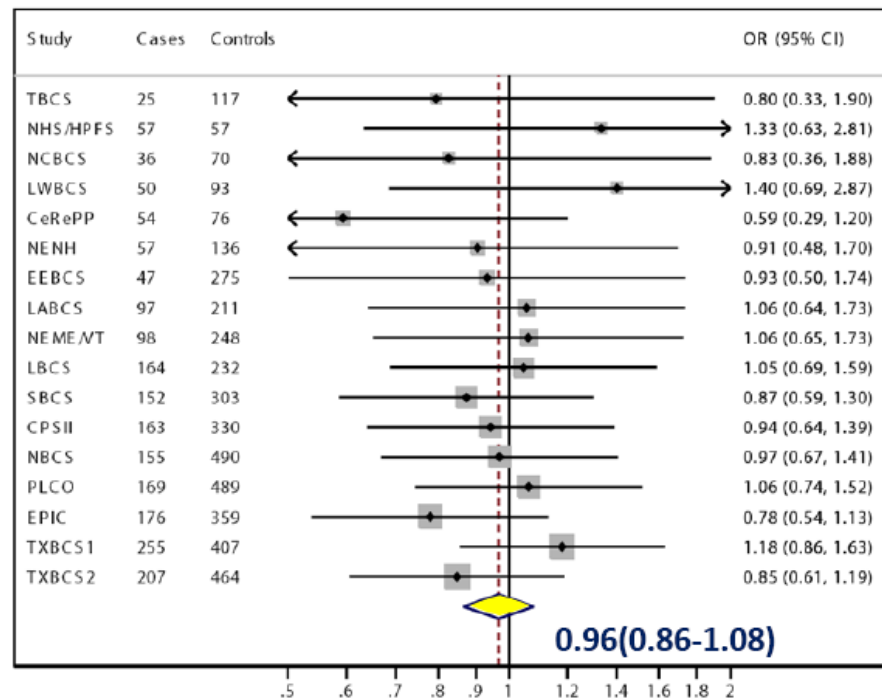
# A note about modeling “E”



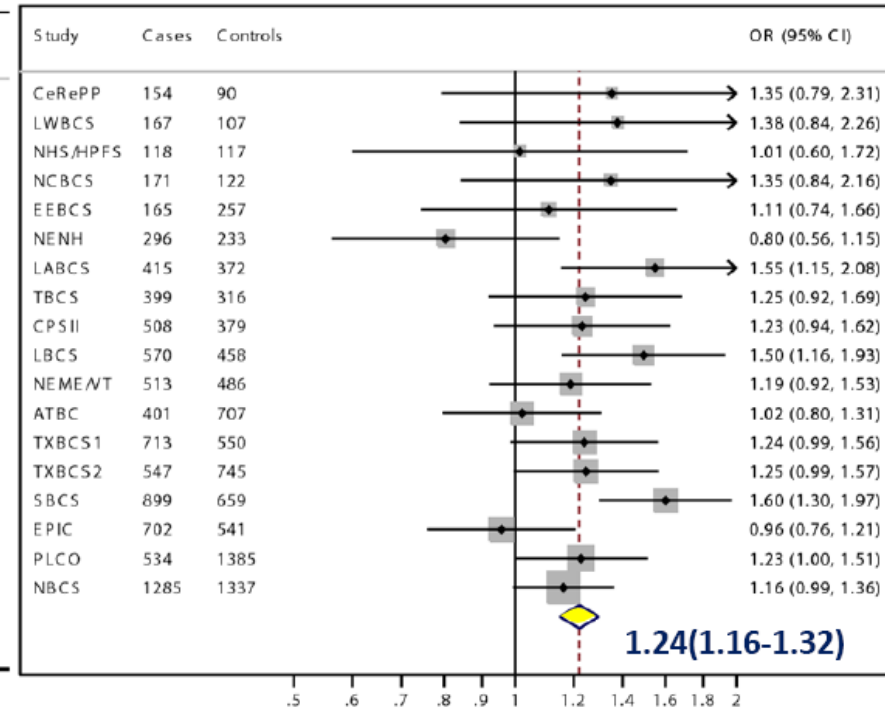
Real data examples

# GE Interaction for Bladder Cancer Risk: NAT2 Slow Acetylation Increases Risk only for Smokers

## Never Smokers

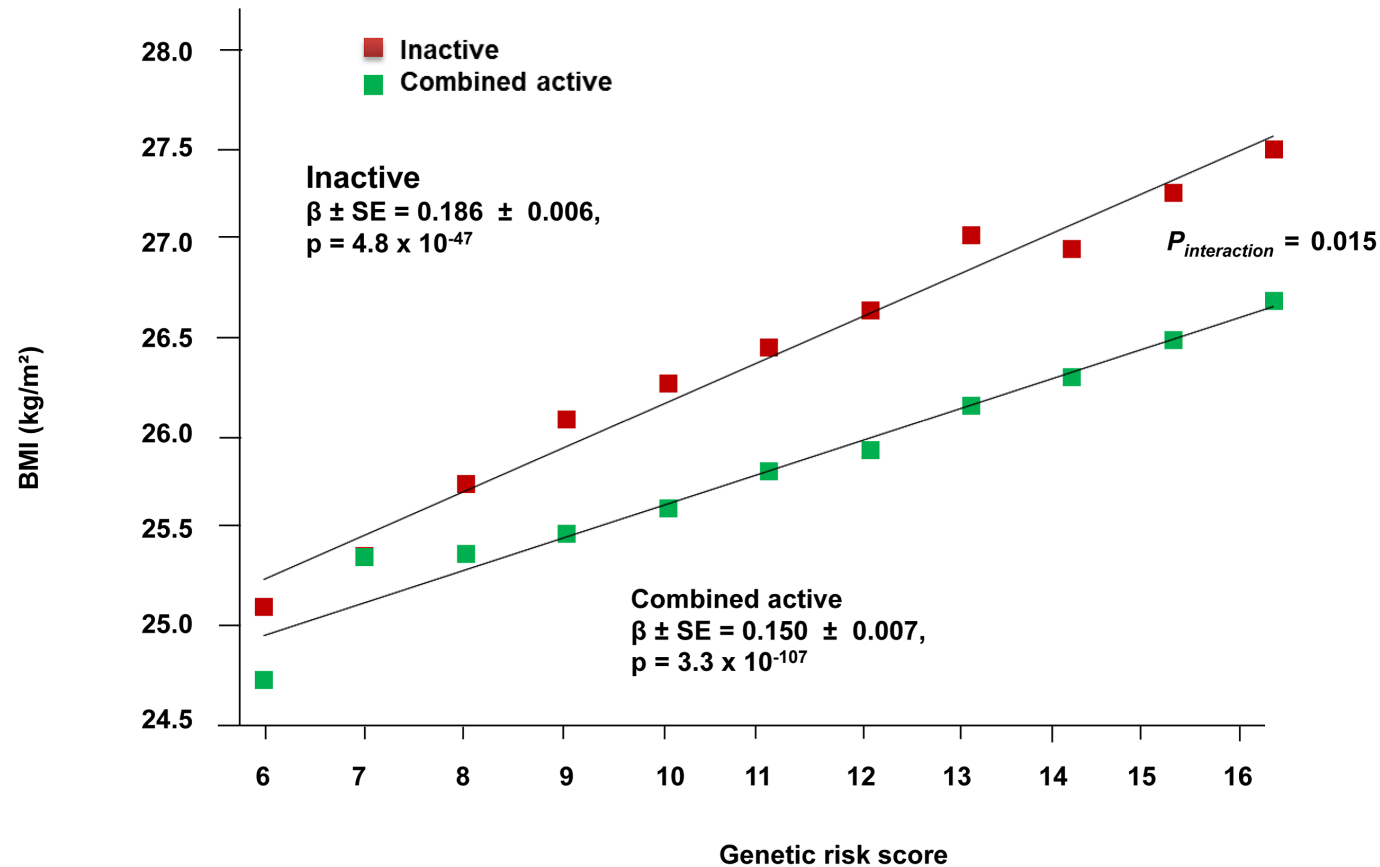


## Ever Smokers



P-interaction =  $2.8 \times 10^{-4}$

# Gene × physical activity interactions in obesity: combined analysis of 111,421 individuals of European ancestry



SNPs	Nearest gene	$\beta_{GE}$	(95% CI)	$P_{interaction}$
rs1121980*	<i>FTO</i>	-0.052	(-0.086, -0.018)	0.003
rs7498665*	<i>SH2B1</i>	-0.003	(-0.039, 0.033)	0.867
rs10913469*	<i>SEC16B</i>	-0.049	(-0.091, -0.006)	0.025
rs10838738*	<i>MTCH2</i>	-0.012	(-0.047, 0.023)	0.502
rs17782313*	<i>MC4R</i>	-0.029	(-0.069, 0.010)	0.147
rs3101336*	<i>NEGR1</i>	0.006	(-0.028, 0.040)	0.728
rs6548238*	<i>TMEM18</i>	0.002	(-0.043, 0.047)	0.936
rs10938397	<i>GNPDA2</i>	-0.001	(-0.036, 0.034)	0.946
rs925946*	<i>BDNF</i>	-0.013	(-0.052, 0.025)	0.491
rs368794*	<i>KCTD15</i>	-0.001	(-0.037, 0.035)	0.969
rs7647305*	<i>ETV5</i>	0.024	(-0.018, 0.066)	0.267
rs7132908*	<i>FAIM2</i>	-0.024	(-0.059, 0.010)	0.164

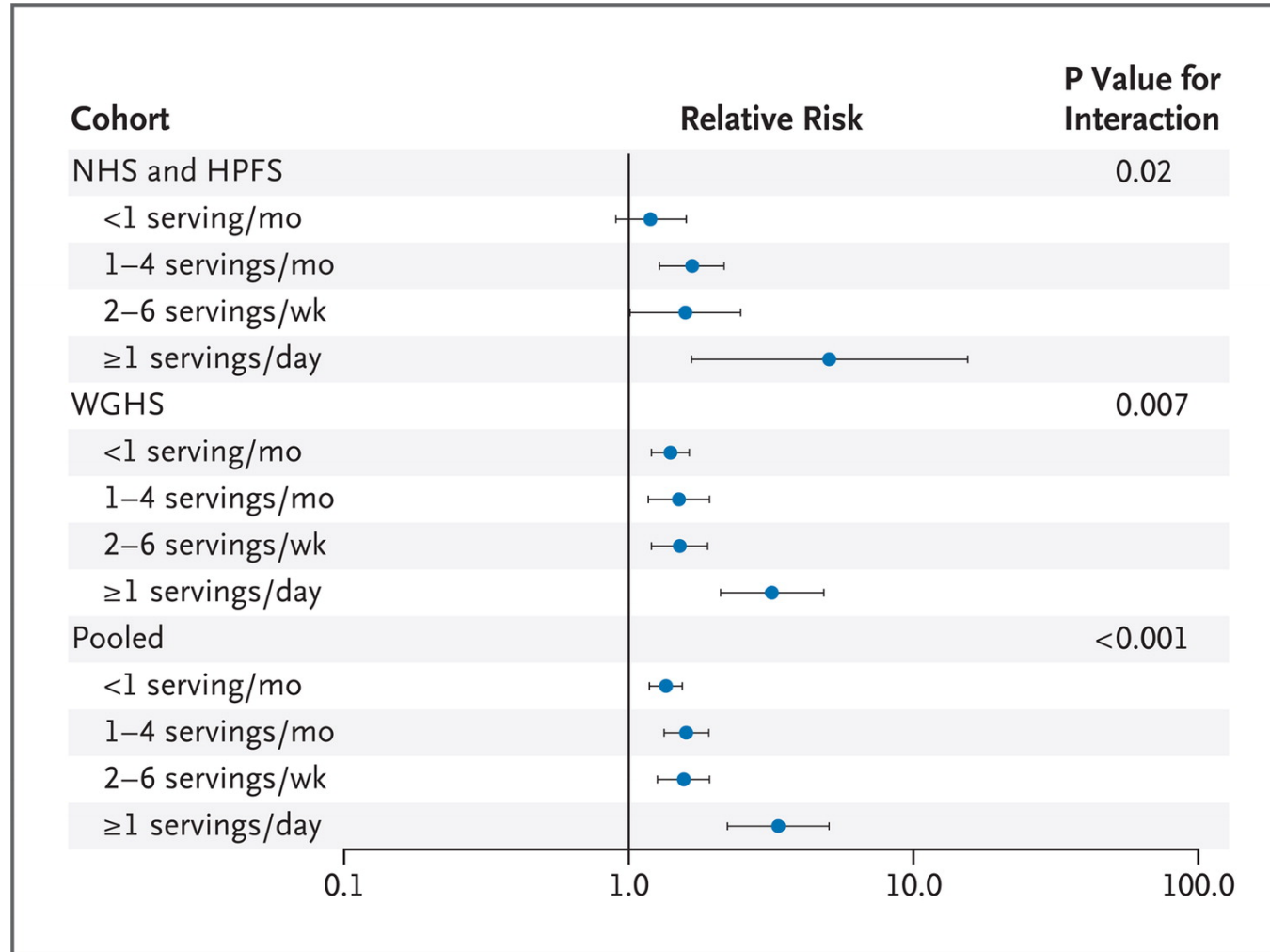
Physical activity was expressed according to the Cambridge Physical Activity Index (CPAI) (4 level scale); further details for the construction of the CPAI can be found in the *Materials and Methods* section and Table S7.

\*Some studies used proxies for these variants, as reported in Table S8.

doi:10.1371/journal.pgen.1003607.t002



# Relative Risk of the Development of Obesity per Increment of 10 Risk Alleles, According to Intake of Sugar-Sweetened Beverages.



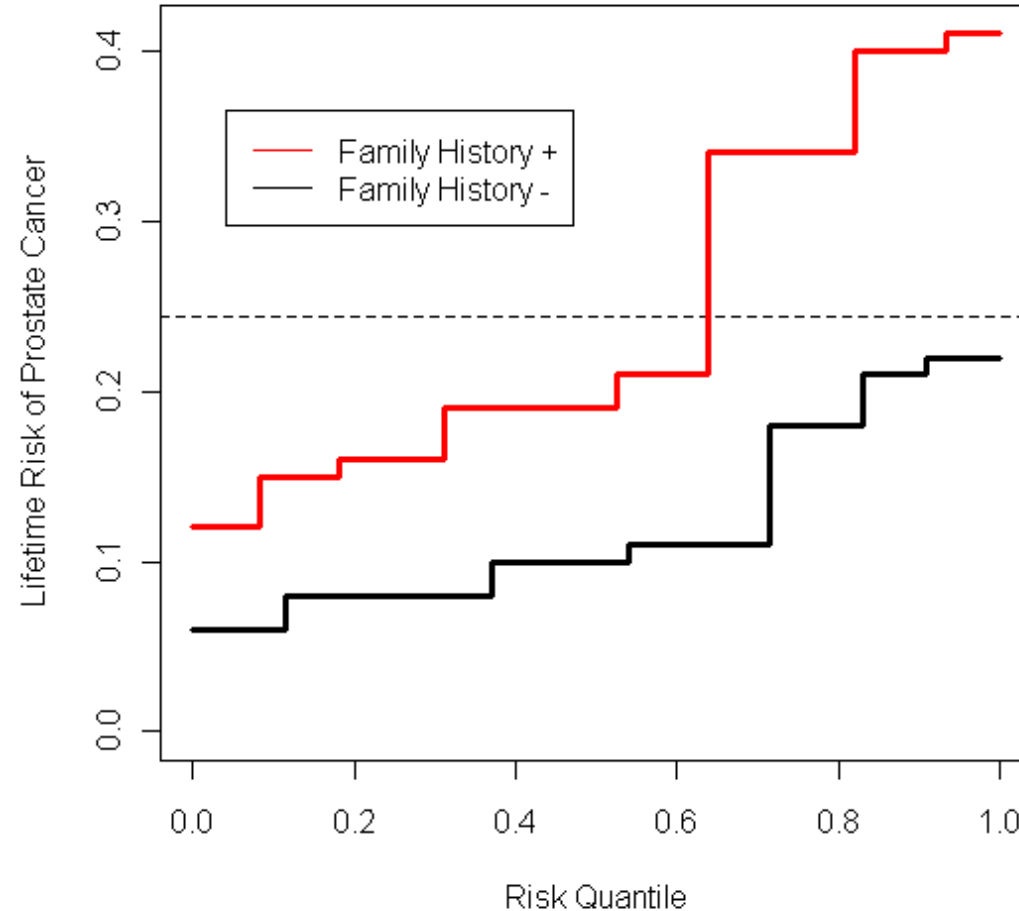
# Joint effects can be very informative regardless of the significance or not of any statistical GxE interactions

## Example: Prostate Cancer

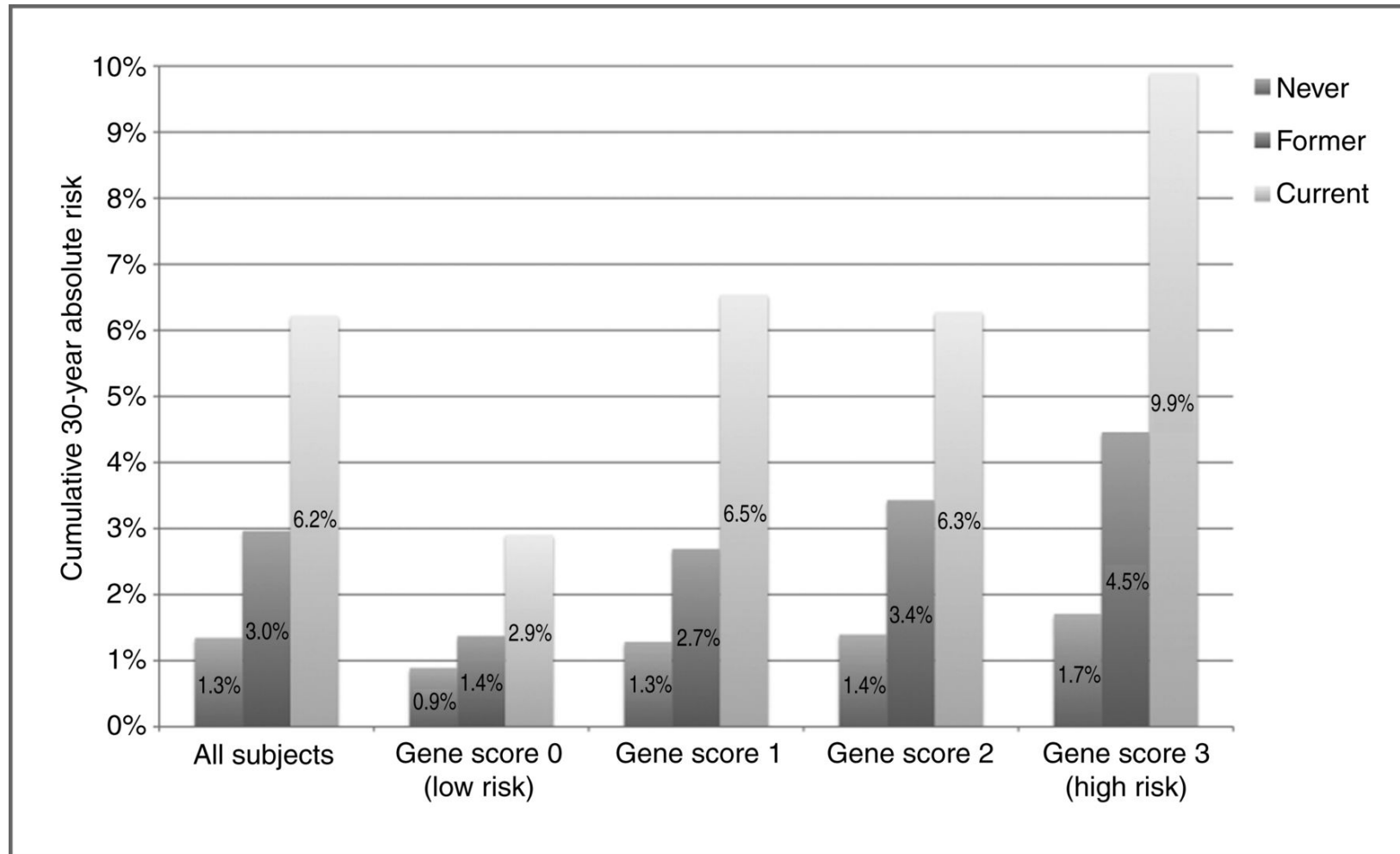
Men who might change their screening strategy based on genetic information:

3.8% of entire population,

63.9% of men with a 1<sup>st</sup> degree relative

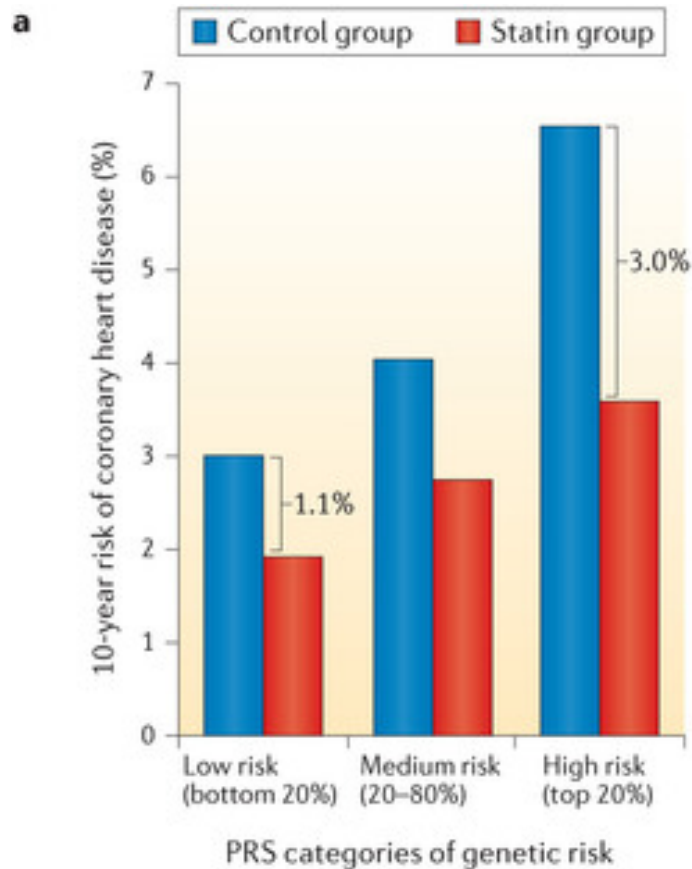


**Cumulative 30-year absolute risk for bladder cancer in a 50-year-old male in the United States, overall and by quartiles of a polygenetic genetic score.**

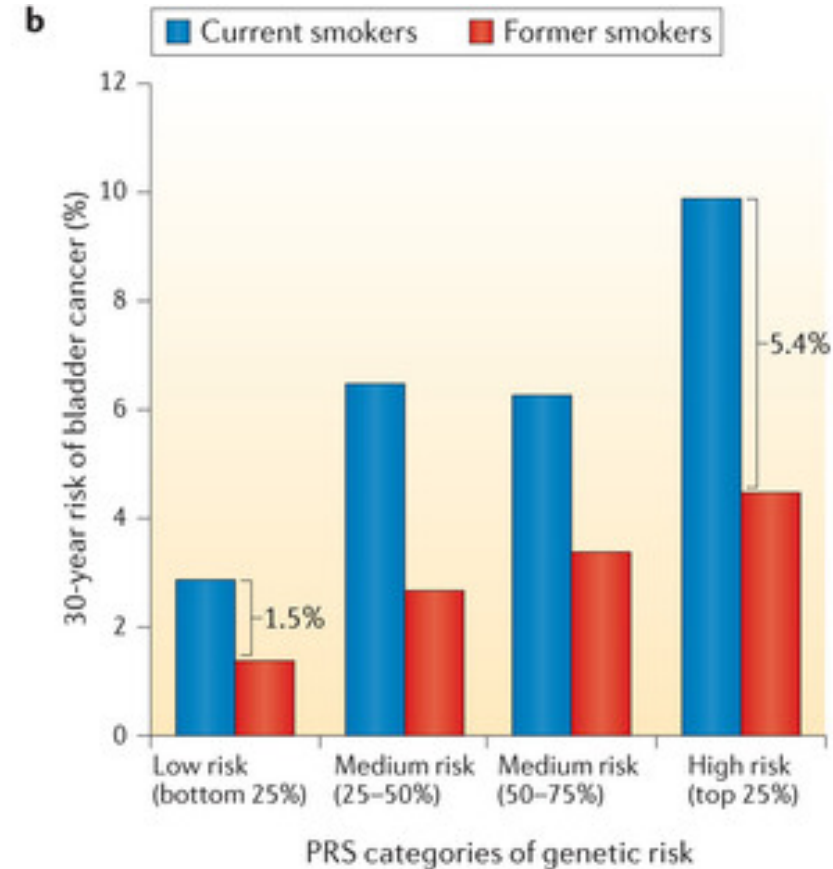


Montserrat Garcia-Closas et al. *Cancer Res* 2013;73:2211-2220

# Intervention in high-risk groups is more efficient



ARR (%)	1.1	1.3	3.0
RRR	0.36	0.32	0.46



ARR (%)	1.5	3.8	2.9	5.4
RRR	0.52	0.60	0.46	0.55

ARR=Absolute risk reduction