

PGC Worldwide Lab Call Details

DATE: Friday, March 14th, 2014

PRESENTER: Peter Kraft, Harvard School of Public Health

TITLE: “Gene-Environment Interactions in Genome-Wide Association Studies: Conceptual and Statistical Issues”

START: We will begin promptly on the hour.

1000 EDT - US East Coast

0700 PDT - US West Coast

1400 GMT - UK

1500 CEST - Central Europe

0000 AEST – Australia (Friday, March 14th into Saturday, March 15th, 2014)

DURATION: 1 hour

TELEPHONE:

- US Toll free: 1 866 515.2912

- International direct: +1 617 399.5126

- Toll-free number? See http://www.btconferencing.com/globalaccess/?bid=75_public

- Operators will be on standby to assist with technical issues. “*0” will get you assistance.

- This conference line can handle up to 300 participants.

PASSCODE: 275 694 38 then #

Lines are Muted **NOW**

Lines have been automatically muted by operators as it is possible for just one person to ruin the call for everyone due to background noise, electronic feedback, crying children, wind, typing, etc.

Operators announce callers one at a time during question and answer sessions.

Dial *1 if you would like to ask a question of the presenter. Presenter will respond to calls as time allows.

Dial *0 if you need operator assistance at any time during the duration of the call.

UPCOMING PGC Worldwide Lab

DATE: Friday, April 11th, 2014

PRESENTER: Ole Andreassen, Institute of Clinical Medicine, University of Oslo

TITLE: To Be Announced

START: We will begin promptly on the hour.

1000 EDT - US East Coast

0700 PDT - US West Coast

1500 BST - UK

1600 CEST - Central Europe

0000 AEST – Australia (Saturday, April 12th, 2014)

DURATION: 1 hour

TELEPHONE:

- US Toll free: 1 866 515.2912

- International direct: +1 617 399.5126

- Toll-free number? See http://www.btconferencing.com/globalaccess/?bid=75_public

- Operators will be on standby to assist with technical issues. “*0” will get you assistance.

- This conference line can handle up to 300 participants.

PASSCODE: 275 694 38 then #

Gene-environment interactions in genome-wide association studies: conceptual and statistical issues

Peter Kraft
Professor of Epidemiology and Biostatistics
Harvard School of Public Health

14 March 2014

Caveat #1

45 minutes is barely enough time to do this topic justice.

Good conceptual overviews:

Hutter CM, Mechanic LE, Chatterjee N, Kraft P, Gillanders EM. Gene-environment interactions in cancer epidemiology: a National Cancer Institute Think Tank report. *Genet Epidemiol* 2013;37(7):643-57

Ahmad S, Varga TV, Franks PW. Gene x environment interactions in obesity: the state of the evidence. *Hum Hered* 2013;75(2-4):106-15

Good statistical introductions:

Kraft P, Hunter D. The challenge of assessing complex gene–gene and gene–environment interactions. In: Khoury MJ, Bedrosian S, Gwinn M, Higgins JPT, Ioannidis JPA, Little J, eds. *Human Genome Epidemiology* (2nd ed.) New York: Oxford University Press, 2010.

Chatterjee N, Mukherjee B. Statistical approaches to studies of gene-gene and gene-environment interactions. In: Rebbeck TR, Ambrosone CB, Shields P, eds. *Molecular epidemiology: applications in cancer and other human diseases* New York: Informa Healthcare, 2008.

Caveat #2

All of my examples will be drawn from cancer epidemiology and the epidemiology of obesity.

Outline

- Definition and Notation
- Leveraging G×E Interactions to Discover Risk Markers
- State of the science: cancer and obesity
- Practicalities

Outline

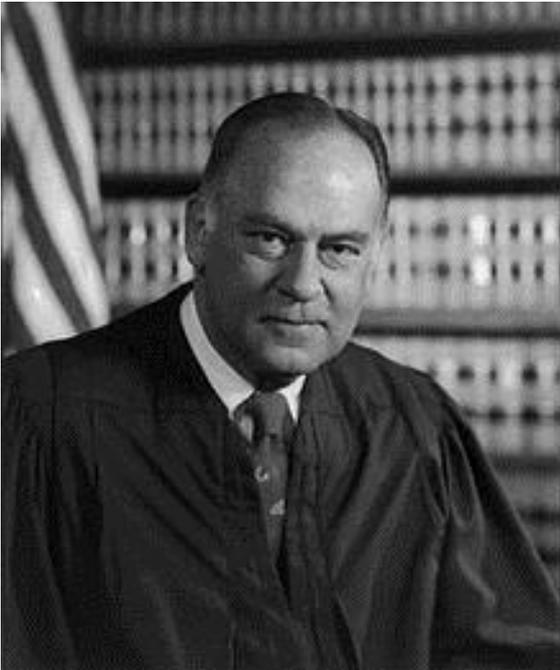
- Definition and Notation
- Leveraging G×E Interactions to Discover Risk Markers
- State of the science: cancer and obesity
- Practicalities

[F]ew epidemiological grant applications now fail to identify the establishment of 'gene–environment interaction' as a primary aim.

Yet much of this discussion is as careless in its use of terms as the early epidemiological literature that first prompted debate about the topic 40 years ago...

Biological interaction, public health interaction, and statistical interaction are distinct concepts.

Biological Interaction



Potter Stewart

I shall not today attempt further to define the kinds of material I understand to be embraced within that shorthand description; and perhaps I could never succeed in intelligibly doing so. But I **know it when I see it...**

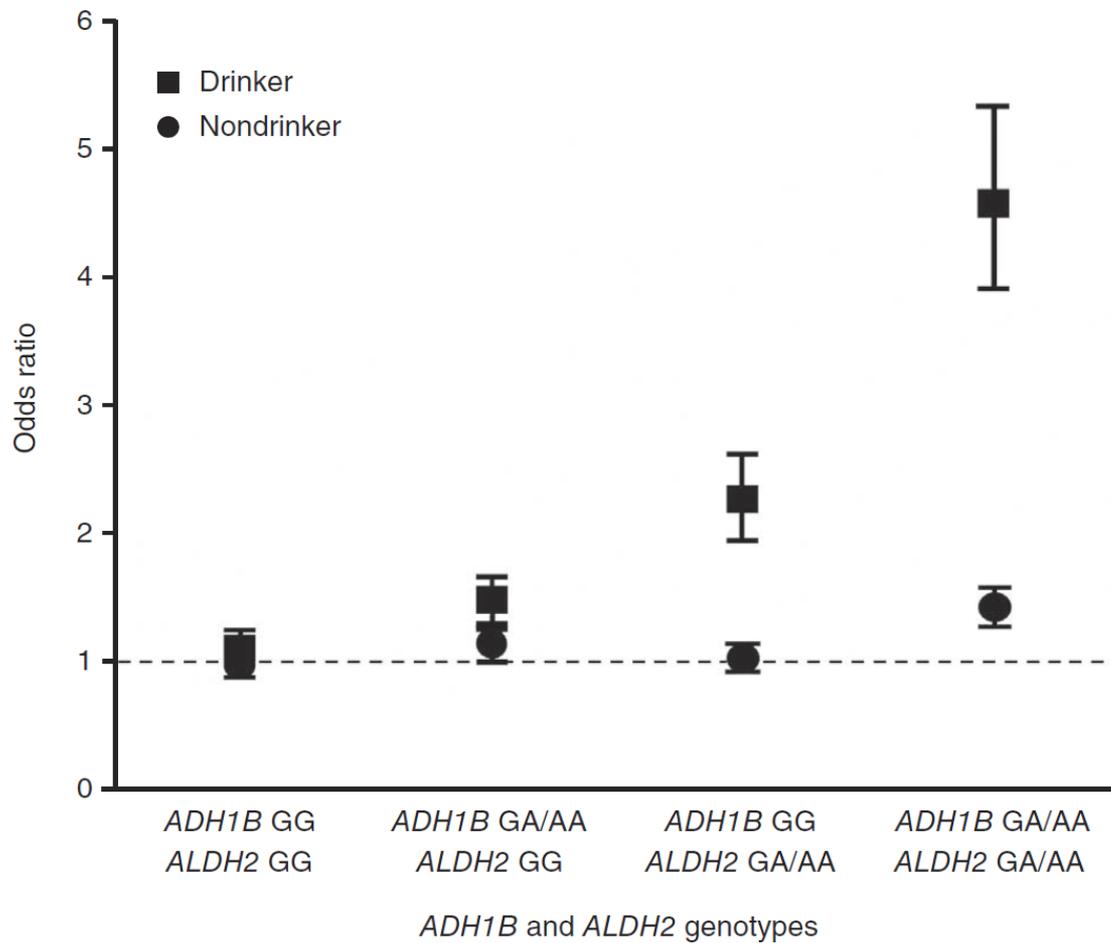
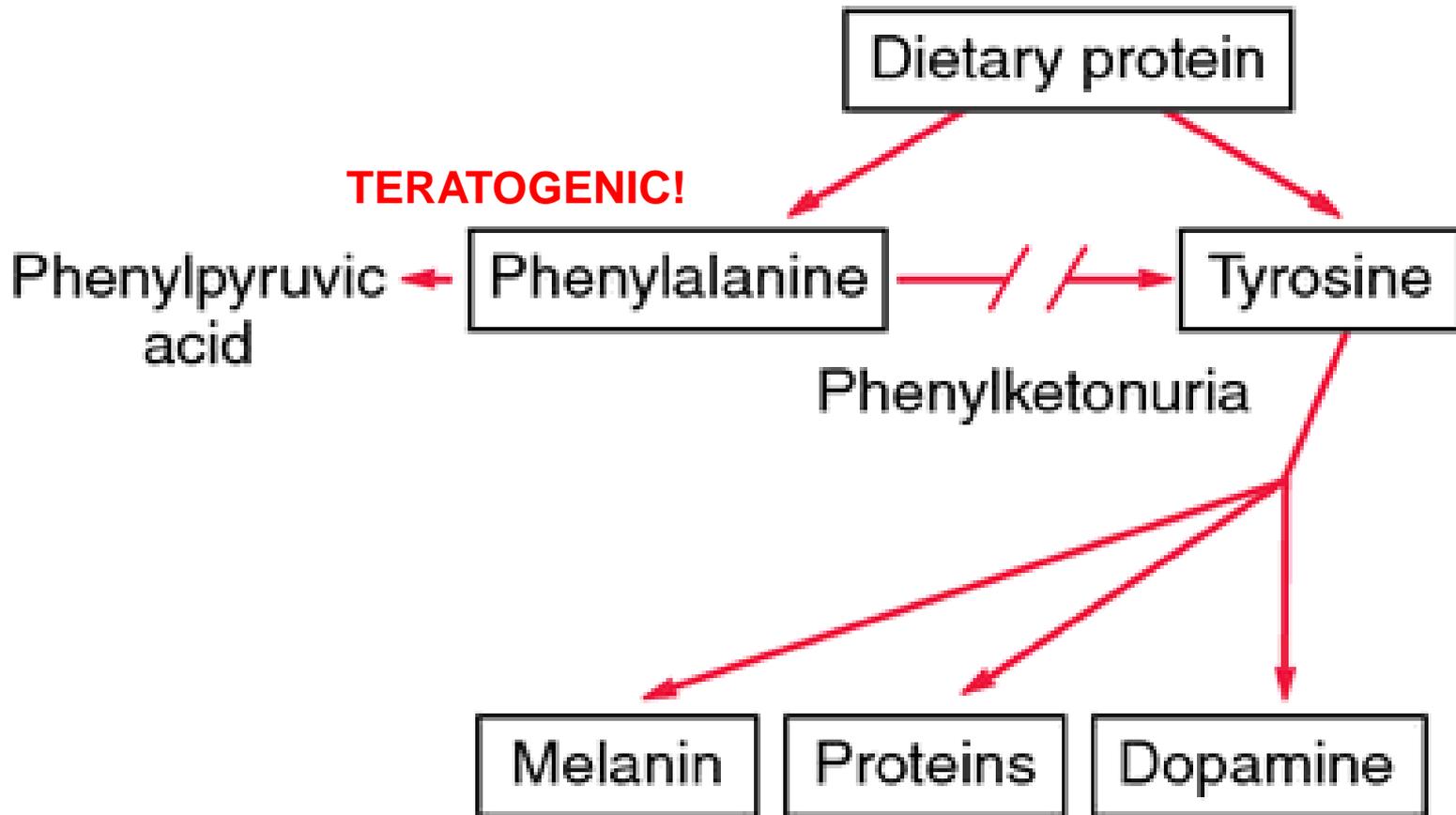


Figure 2 Plots showing the ORs for ESCC in alcohol drinkers and nondrinkers with different *ADH1B* rs1042026 and *ALDH2* rs11066015 genotypes. The vertical bars represent the 95% CIs. The horizontal dashed line indicates the null value (OR = 1.0).

The GxE Poster Child: PKU



The GxE Poster Child: PKU

- Completely penetrant: exposed carriers get the disease if untreated
- Penetrance consistent with biological mechanism: failure of phenylalanine metabolism
- A preventive intervention exists: remove phenylalanine from the diet
- This intervention is too costly to apply to the general population, so targeting carriers makes sense

Sufficient Component Cause/ Counterfactual Interaction

“[W]e may be interested whether, for some individuals, an outcome occurs if both of two exposures are present but not if only one or the other is present.”

Distinct from “biological interaction,” although often referred to as such.

Under strong assumptions, a specific form of statistical interaction—departure from additivity on the absolute risk scale—implies interaction in this sense.

Sufficient Component Cause/ Counterfactual Interaction

“[W]e may be interested whether, for some individuals, an outcome occurs if both of two exposures are present but not if only one or the other is present.”

This is not necessarily the same as intuitive notions of “biological interaction.” Consider the 1986 World Series: it took *both* Bill Buckner’s error and Ray Knight’s earlier single for the Red Sox to lose Game 6. But the two events were not dependent or contemporaneous.

Public Health Interaction

“[P]ublic health interactions correspond to a situation in which the public health costs or benefits from altering one factor must take into account the prevalence of other factors.”

E.g. carriers of a particular allele may benefit disproportionately from a risk-reducing intervention.

If “public health benefit” is measured in terms of reducing incidence, this corresponds to departures from additivity on the absolute risk scale.

Public Health Interaction

“[P]ublic health interactions correspond to a situation in which the public health costs or benefits from altering one factor must take into account the prevalence of other factors.”

Presence of a public health interaction need not imply that a targeted intervention strategy is ideal: if the intervention is inexpensive and risk-free, a population-based strategy may be better.

Statistical Interaction

- “By interaction or effect [measure] modification we mean a variation in some measure of the effect of an exposure on disease risks across the levels of [...] a modifier. [...] The definition of interaction depends on the measure of association used.”
- In other words, a statistical interaction between two factors refers to departure from an additive effects model on a particular scale

Simple example

$$G \begin{cases} 1 & \text{if carrier} \\ 0 & \text{if non-carrier} \end{cases} \quad E \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if unexposed} \end{cases}$$

Risk of disease

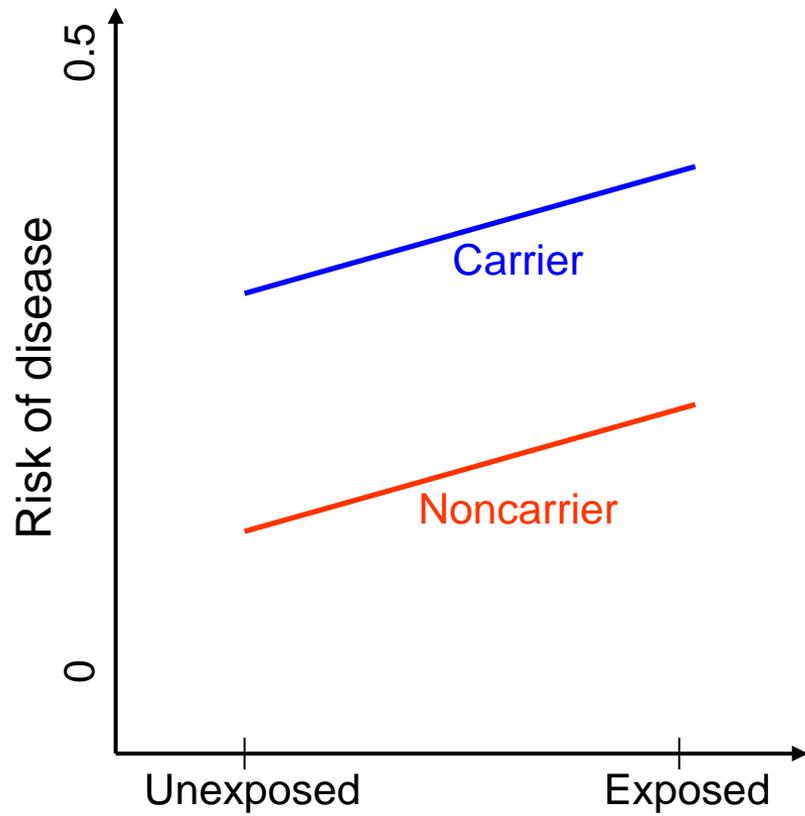
$$p_{GE} = b_0 + b_g G + b_e E + b_{ge} GE$$

Log odds of disease

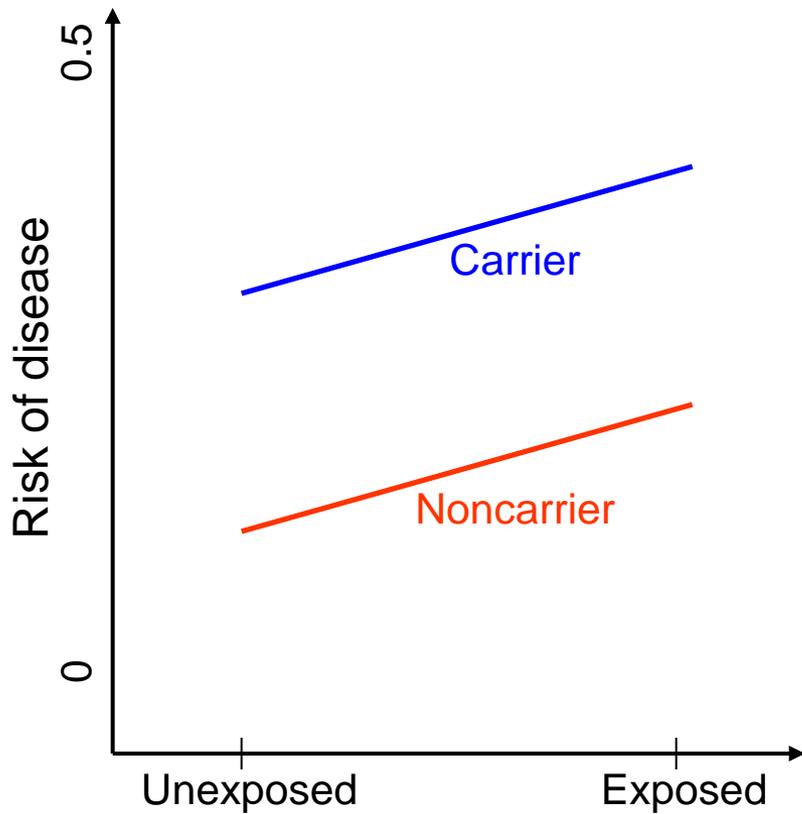
$$\log \frac{p_{GE}}{1-p_{GE}} = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE$$

Test for “additive interaction:” H_0 is $b_{ge}=0$

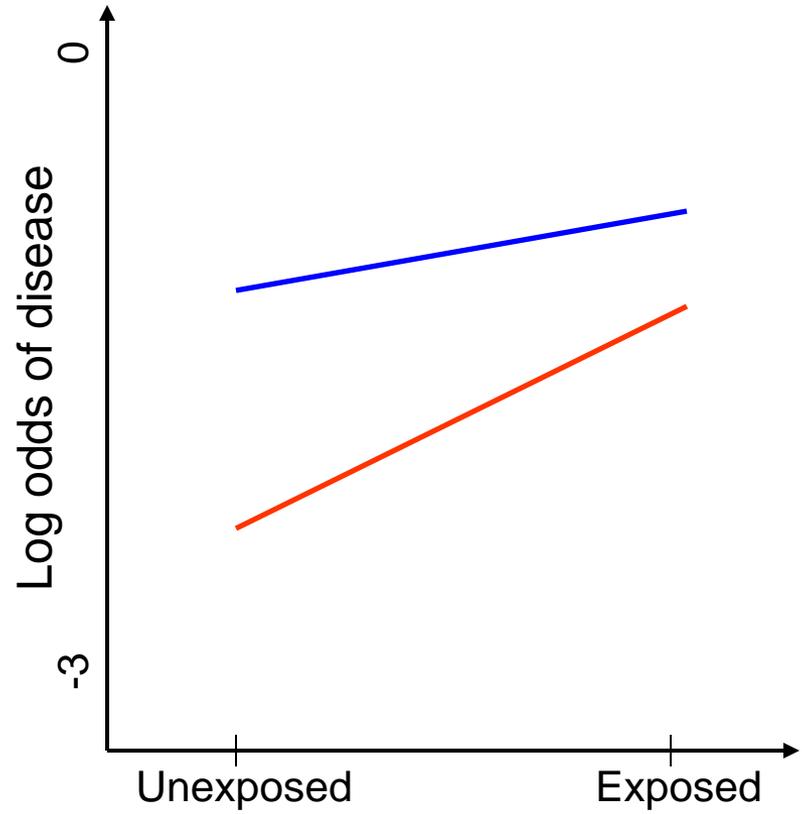
Test for “(multiplicative) interaction:” H_0 is $\beta_{ge}=0$



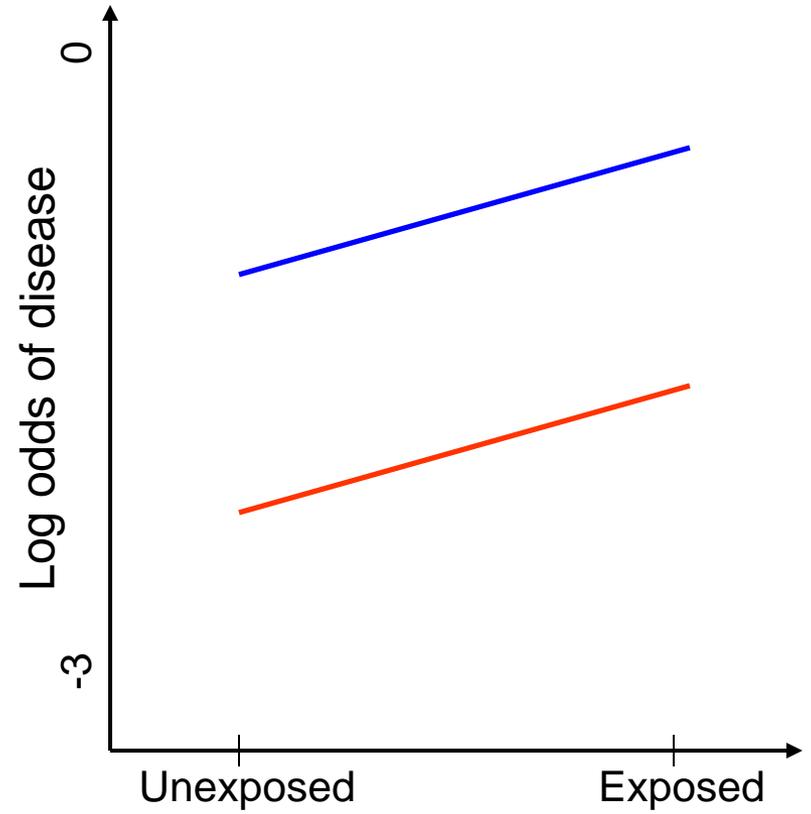
$$b_{ge}=0$$



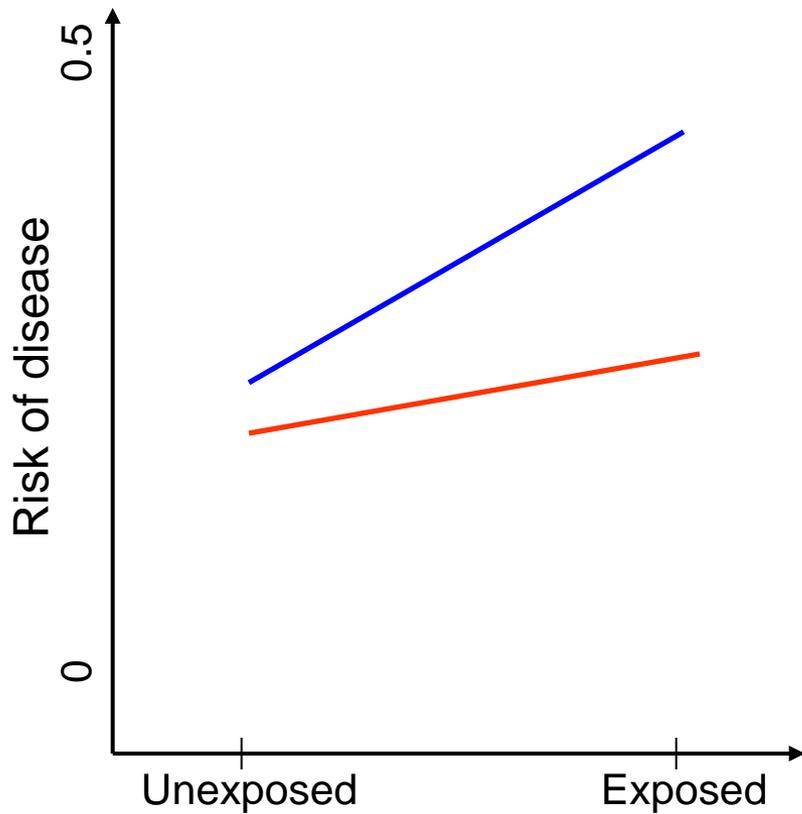
$$b_{ge}=0$$



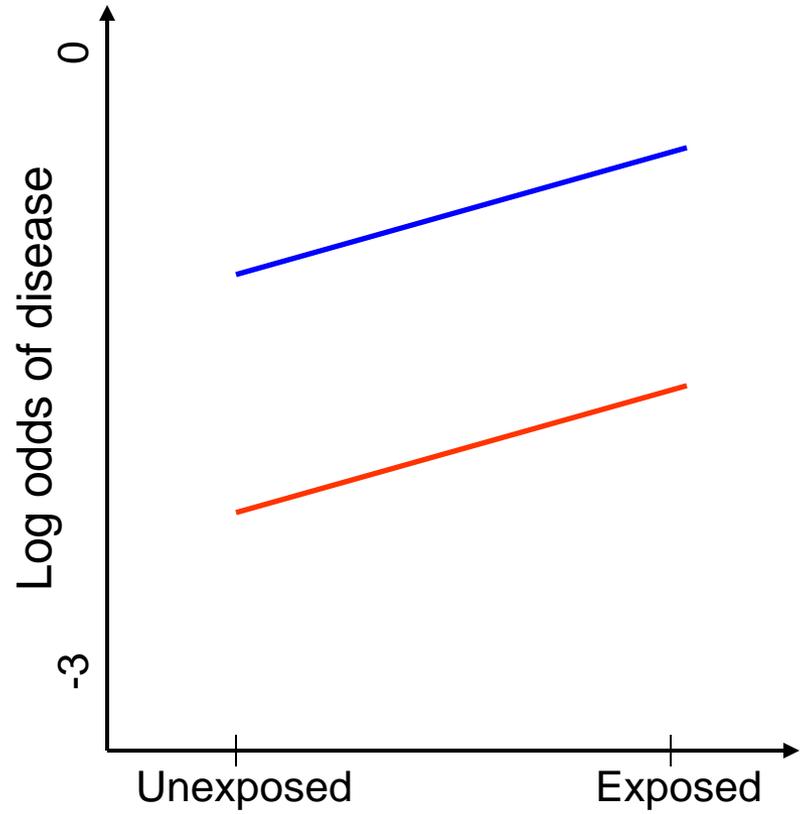
$$\beta_{ge} \neq 0$$



$$\beta_{ge}=0$$



$$b_{ge} \neq 0$$



$$\beta_{ge} = 0$$

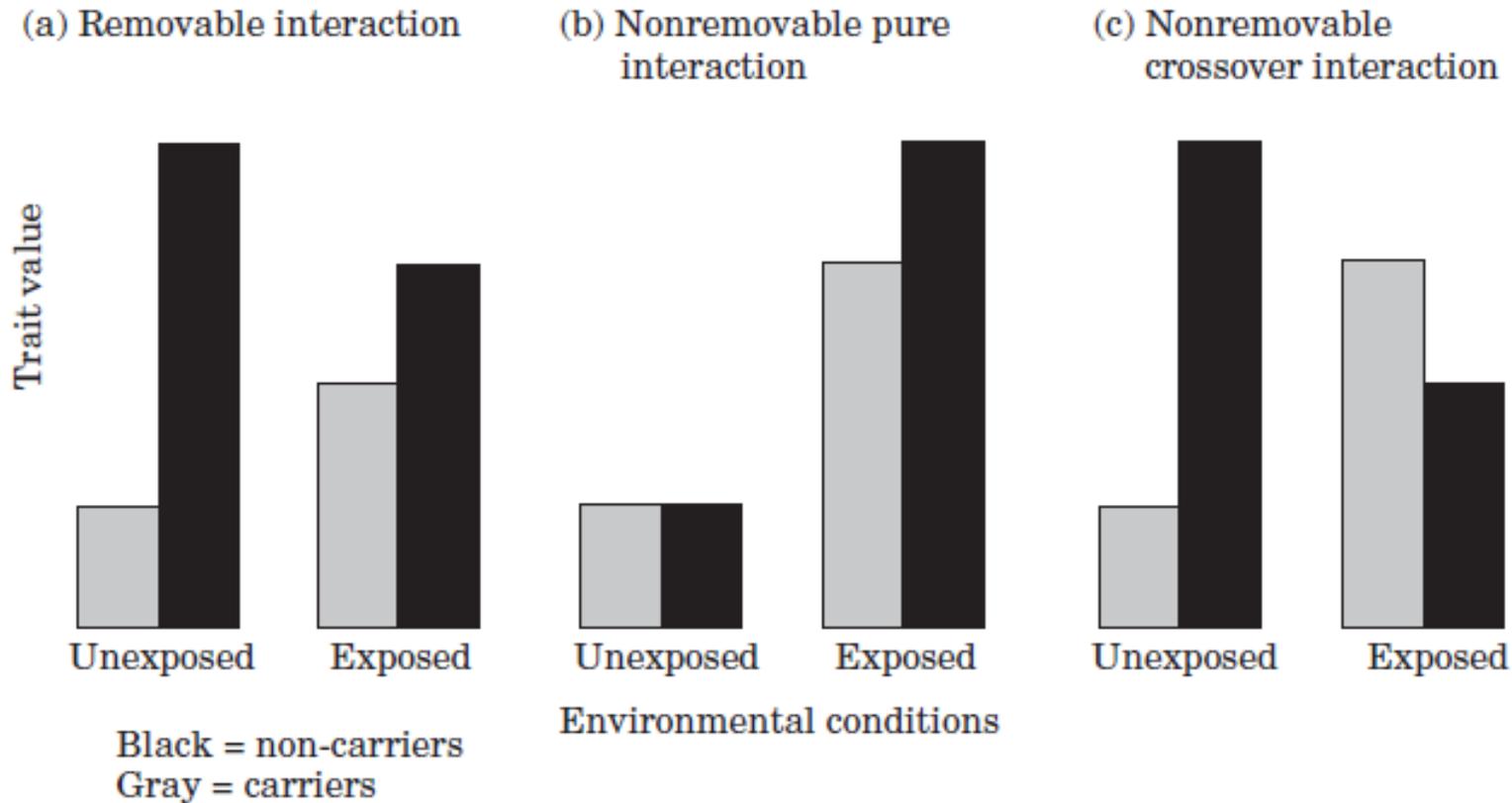
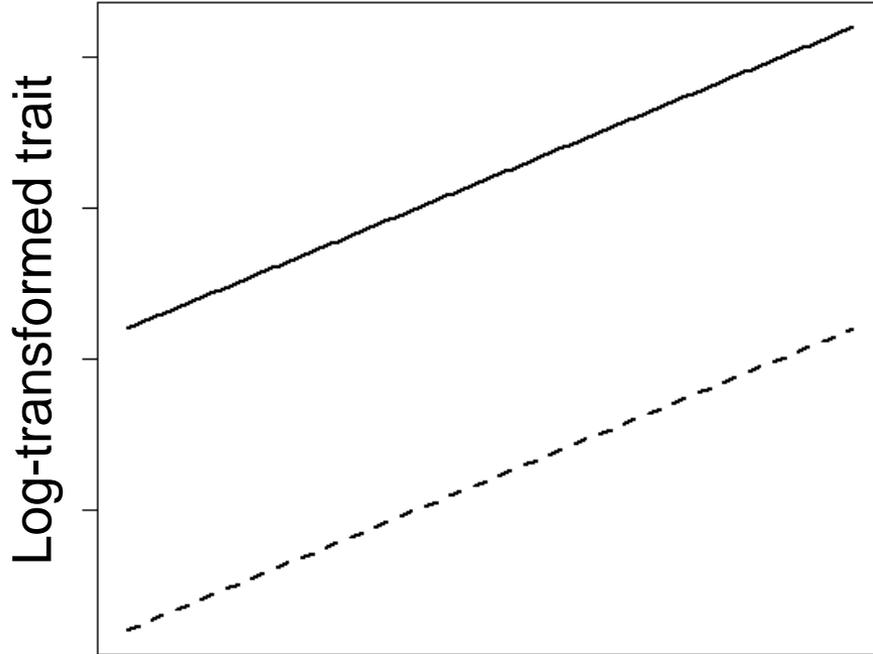


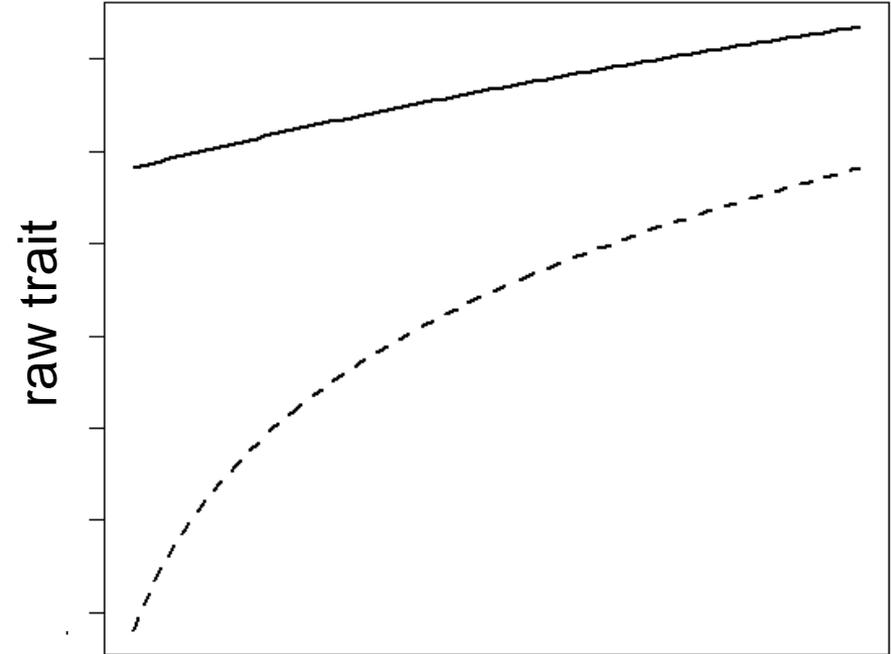
Figure 9.2 Three qualitative patterns of gene–environment interaction. The y-axis represents a trait value (e.g., mean height, disease prevalence, expected survival); the x-axis represents two environmental conditions; the black bars denote noncarriers and the gray bars represent carriers. (a) gives an example of a removable interaction. (b) and (c) are non-removable interactions (“pure” and “crossover” interactions, respectively).

Continuous Y



E

$Y = aG + bE + \text{error}$
fits well

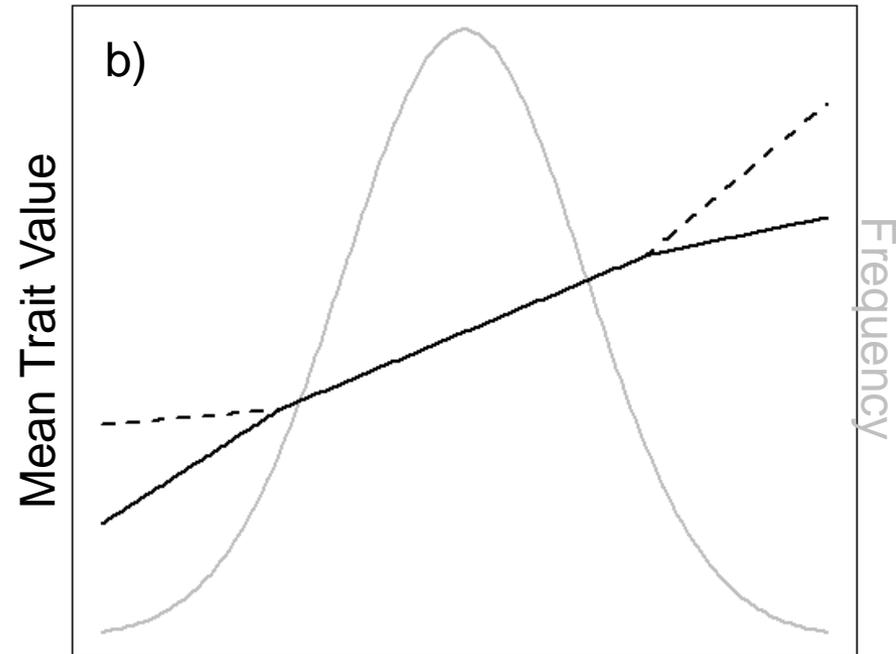
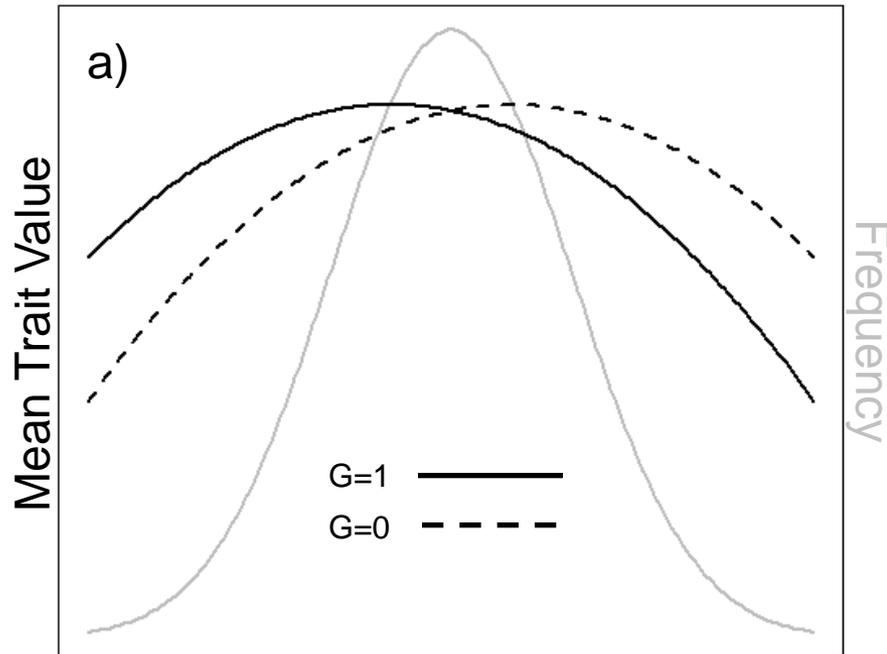


E

$Y = aG + bE + \text{error}$
does not fit well

Gene-environment interactions can be easily created or eliminated by changing the scale of Y. There is no universally appropriate scale.

Two “reaction norms” (i.e. gene-environment interaction patterns)
after Lewontin (1974) Am J Hum Genet



Genetic effect, environmental effect, and gene-environment interaction all depend on what part of the E distribution you've sampled: this has implications for discovery and replication

Nota Bene!

Response to exposure is not the same as gene-environment interaction as typically defined.

In this setting, the phenotype is missing in unexposed individuals.

For example: change in mammographic density in response to tamoxifen, nicotine and alcohol addiction, post-traumatic stress disorder, etc.

This has implications for selection of controls: if exposure is uncommon, but proportion of responders among exposed is high, unexposed “controls” may be inefficient.

Gene-dose interactions among exposed are more akin to gene-environment interactions as typically defined.

Recap

- Biological, public health, and statistical interactions are distinct concepts
- Changes in scale can create or remove statistical interactions
- The appropriate choice of scale depend on the problem at hand

Outline

- Definition and Notation
- **Leveraging G×E Interactions to Discover Risk Markers**
- State of the science: cancer and obesity
- Practicalities

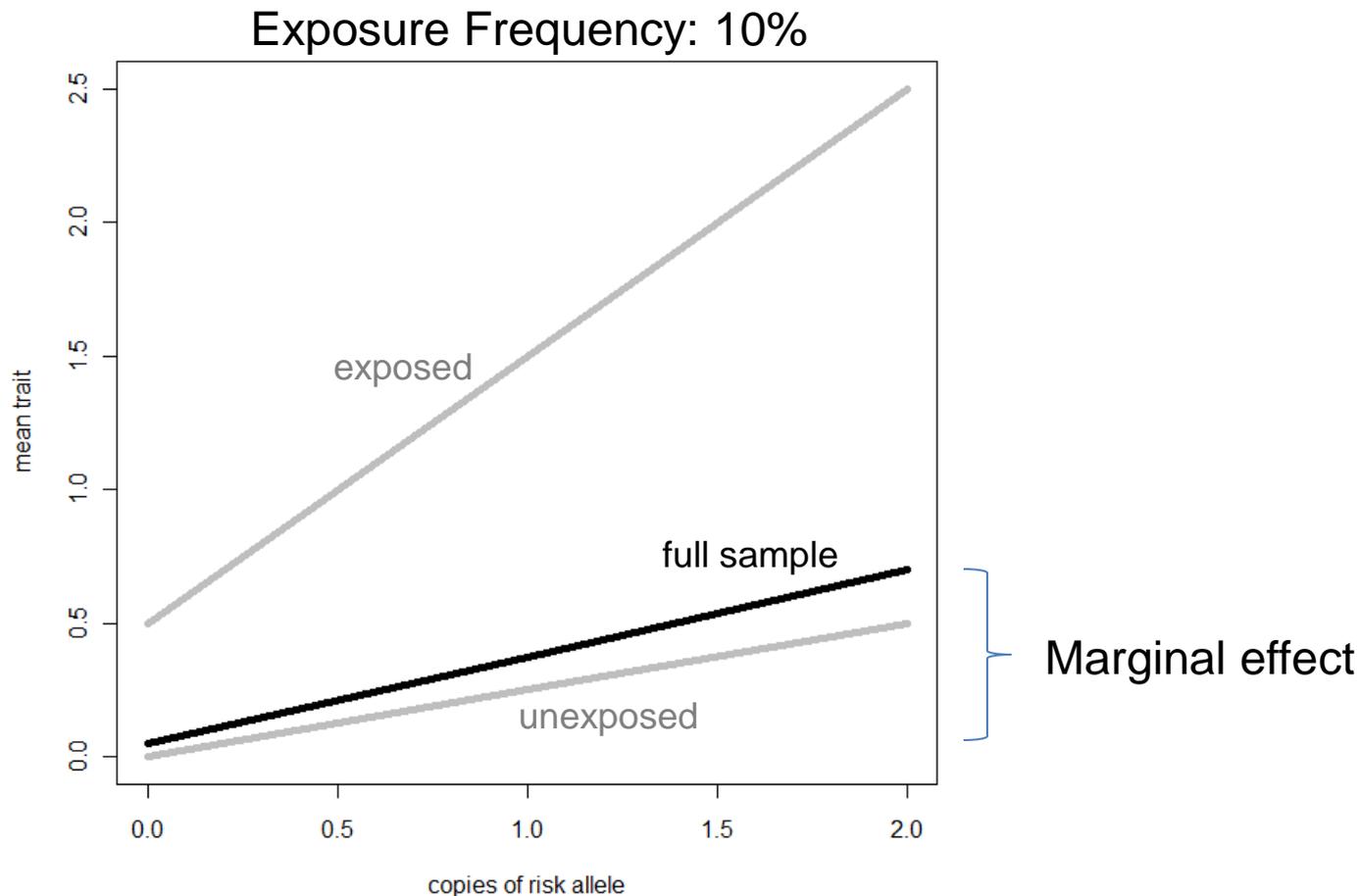
The appropriate choice of test depends on the problem at hand.

The problem at hand in genome-wide association studies is often to find markers that are associated with phenotype in any exposure stratum.

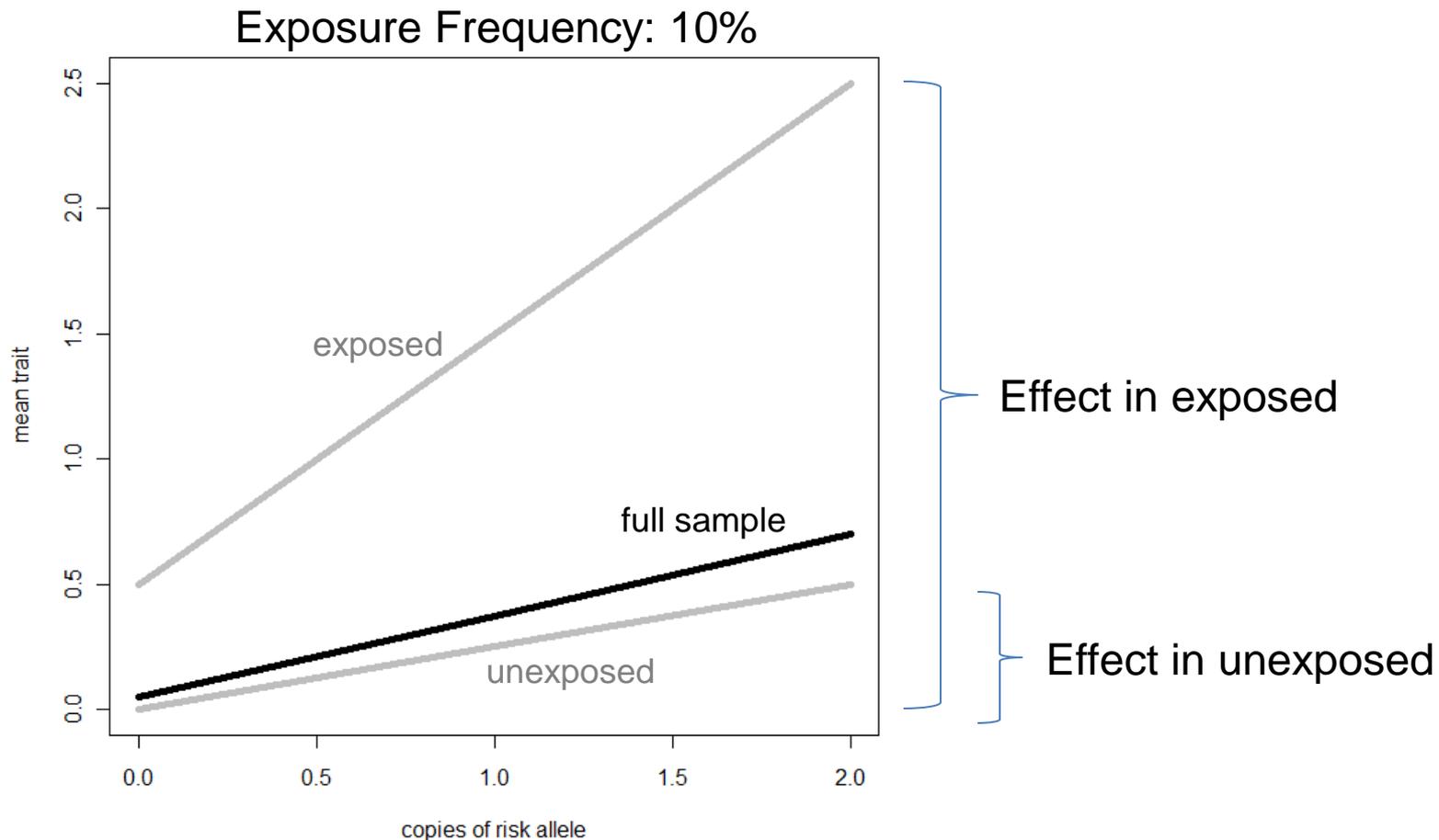
Classical tests for statistical interaction are not testing the appropriate null hypothesis.

But we can leverage the presence of statistical interaction to increase power relative to the marginal test of gene-environment interaction.

- If the genetic effect is restricted to the exposed subgroup, then the marginal test (which averages over exposure) may lose power



- If the genetic effect is restricted to the exposed subgroup, then the marginal test (which averages over exposure) may lose power



Simple example

$$G \begin{cases} 1 & \text{if carrier} \\ 0 & \text{if non-carrier} \end{cases} \quad E \begin{cases} 1 & \text{if exposed} \\ 0 & \text{if unexposed} \end{cases}$$

Risk of disease

$$p_{GE} = b_0 + b_g G + b_e E + b_{ge} GE$$

Log odds of disease

$$\log \frac{p_{GE}}{1-p_{GE}} = \beta_0 + \beta_g G + \beta_e E + \beta_{ge} GE$$

Test for “additive interaction:” H_0 is $b_{ge}=0$

Test for “(multiplicative) interaction:” H_0 is $\beta_{ge}=0$

These tests throw away information on effect of G among unexposed

Testing for association allowing for interaction

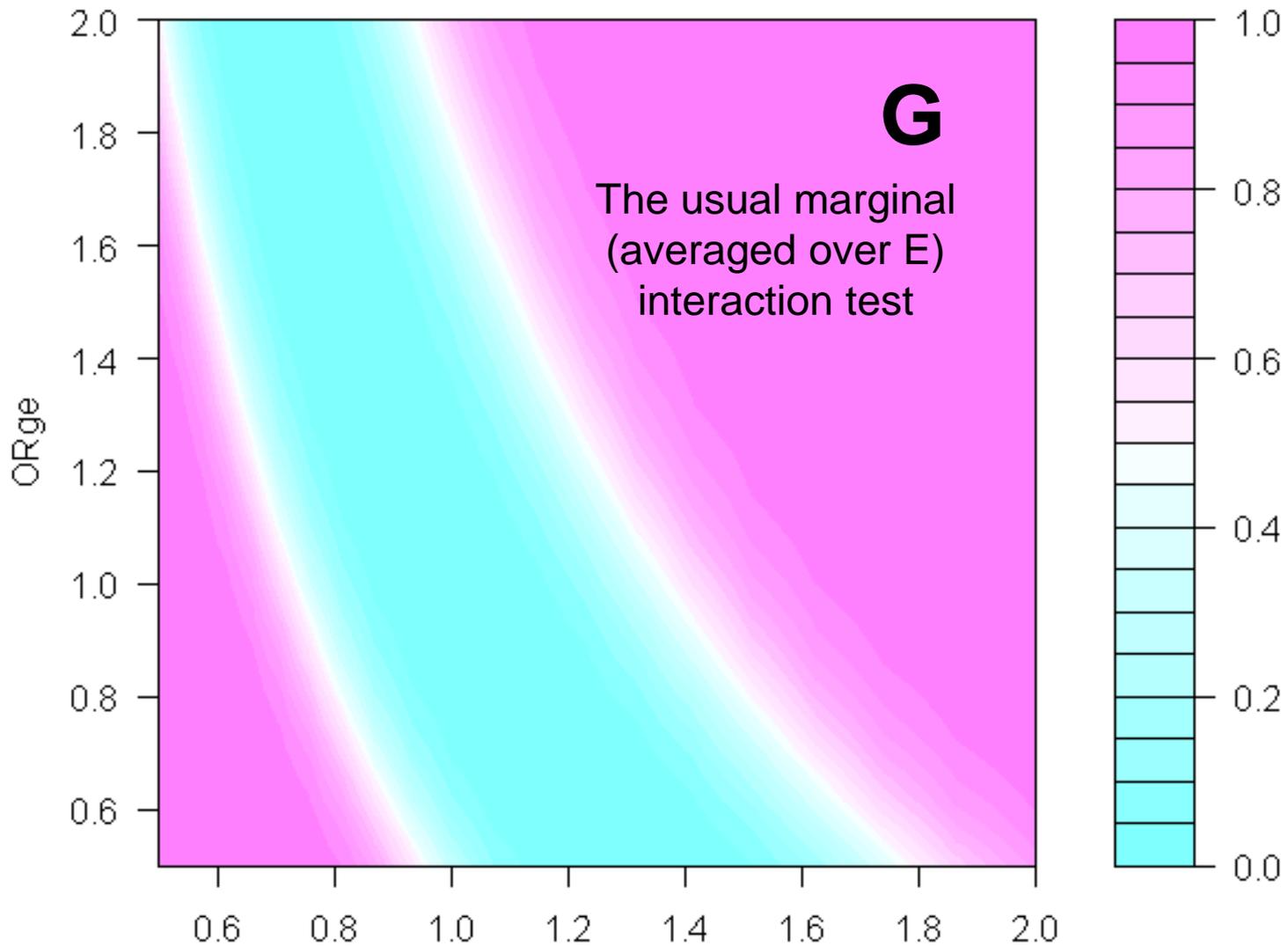
- Is this marker associated with risk of disease in any exposure subgroup?

Compare two models

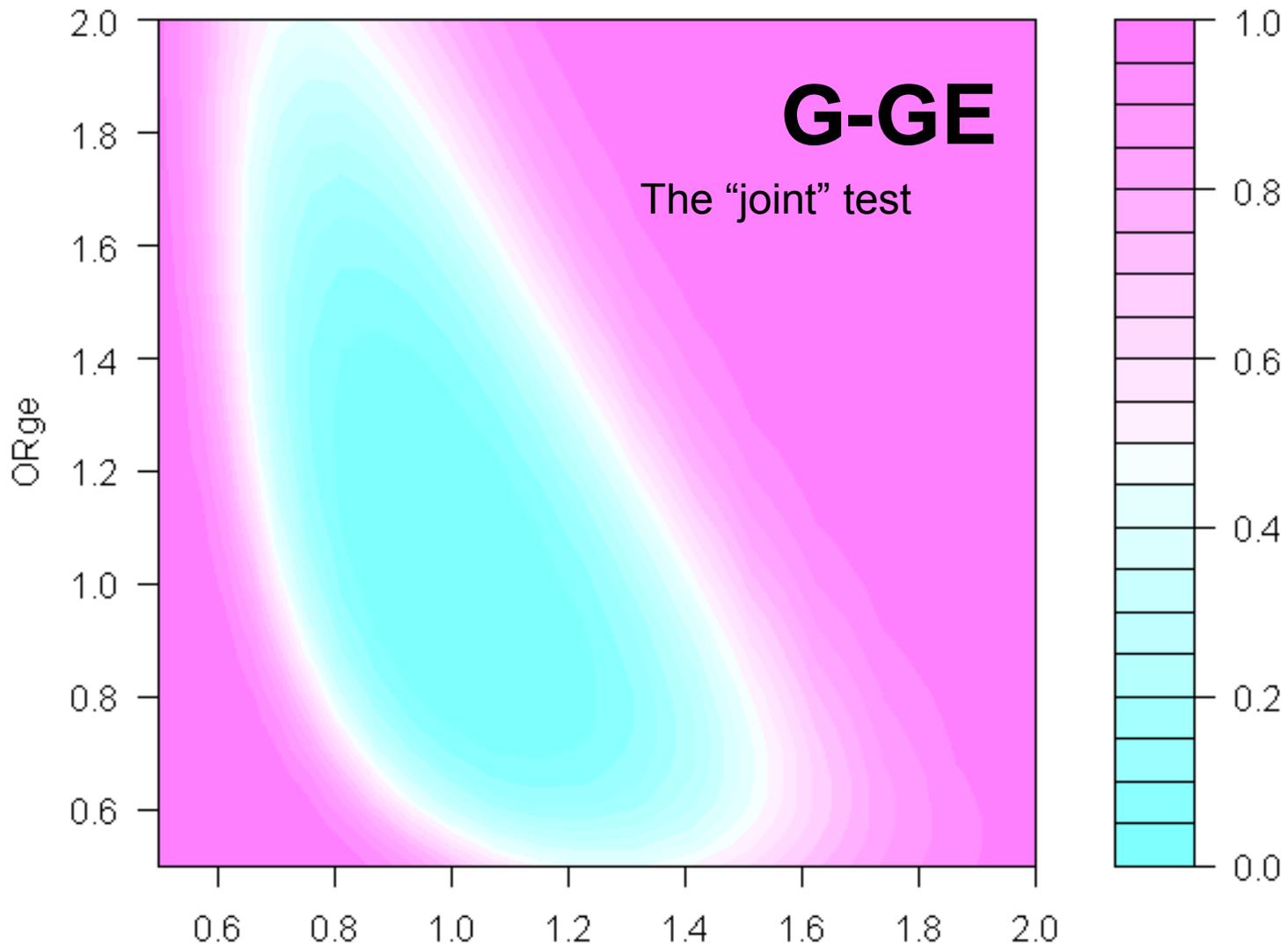
Null
$$\log \frac{p_{GE}}{1-p_{GE}} = \beta_0 + \beta_e E$$

Alternative
$$\log \frac{p_{GE}}{1-p_{GE}} = \beta_0 + \beta_e E + \beta_g G + \beta_{ge} GE$$

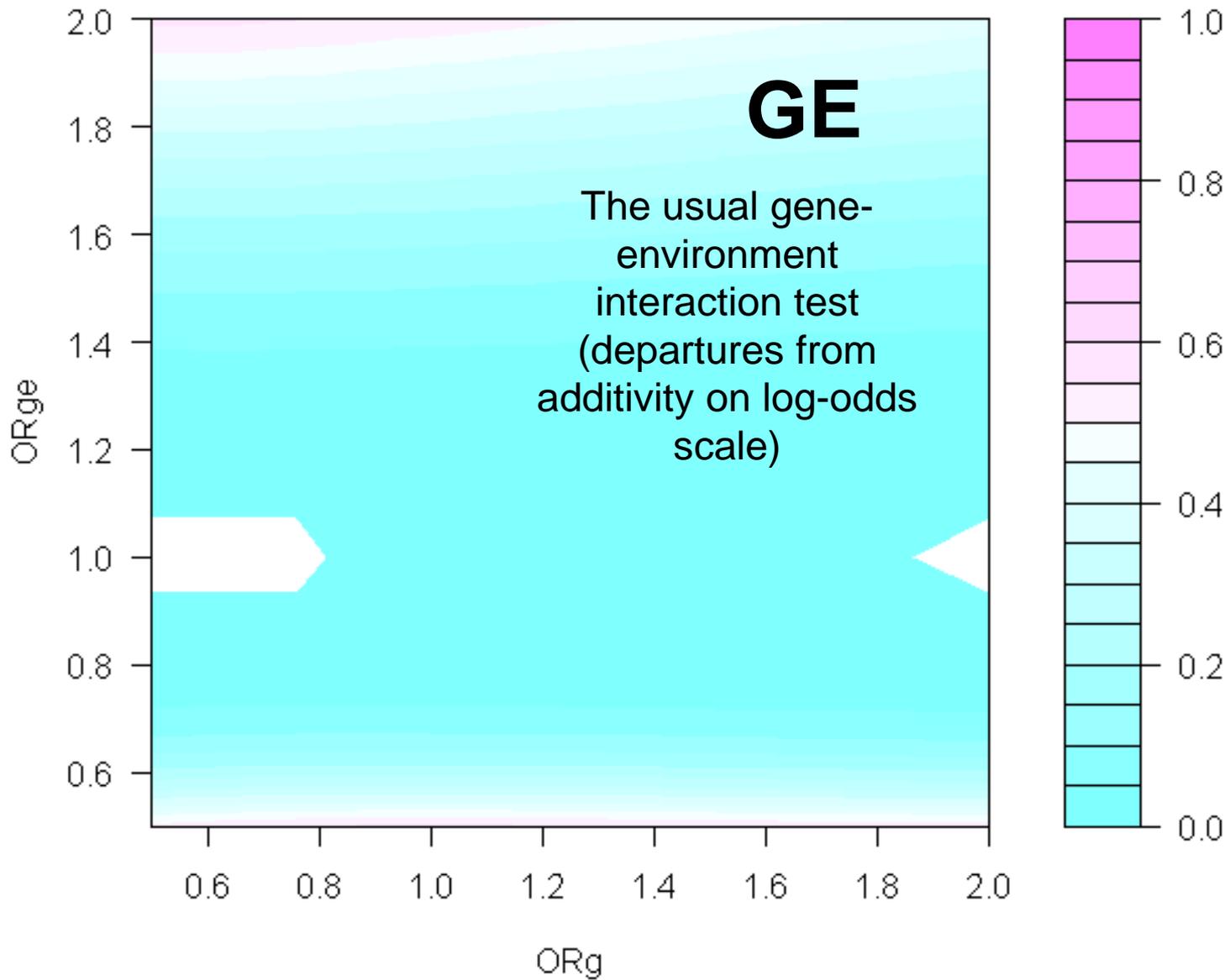
- Can also ask: Is this exposure associated with risk of disease among individuals with any genotype?



$N=900$ $p_g=0.35$ $p_e=0.30$ $OR_e=2$
 OR_g



$N=900$ $p_g=0.35$ $p_e=0.30$ $OR_e=2$



$N=900$ $p_g=0.35$ $p_e=0.30$ $OR_e=2$

Novel genetic associations discovered using the joint test

Parkinson's and coffee intake

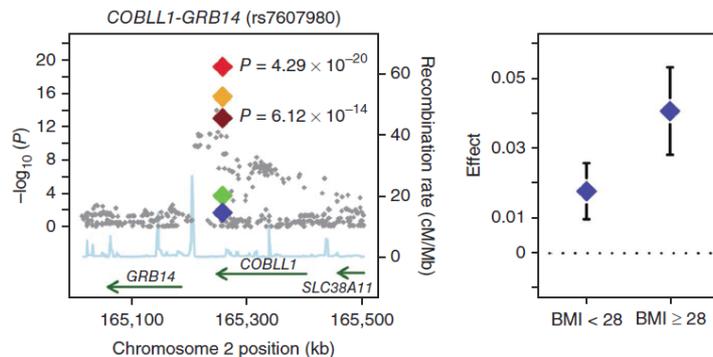
Hamza TH, et al. (2011). PLoS Genet 7(8): e1002237

2,400 cases, 2,500 controls

Fasting glucose and BMI

Manning AK, et al. (2012) Nat Genet

83,000 subjects



Type 2 Diabetes and Gender

Morris AP, et al. (2012) Nat Genet

35,000 cases and 114,000 controls

Type 2 Diabetes and BMI

Perry JRB, et al. (2012) PLoS Genet

16,000 cases and 75,000 controls

Esophageal cancer and alcohol

Wu C, et al. (2012) Nat Genet

10,000 cases and 10,000 controls

Lung function and smoking

Hancock, et al. (2012) Nat Genet

50,000 subjects

Leveraging G×E Interactions to Discover Risk Markers

- Joint test (binary or continuous outcomes)
- “Case-only” test (binary outcomes)
- Hedge methods (binary outcomes)
- Genotype-dependent variance methods (continuous outcomes)

We can also squeeze more information out of our data by assuming the tested genetic marker and the environmental exposure are independently distributed in the general population.

2x2x2 Representation of Unmatched Case-Control Study Examined by Standard Test for GxE Interaction

Environment	Gene			
	G=1		G=0	
	D=1	D=0	D=1	D=0
E=1	<i>a</i>	<i>b</i>	<i>e</i>	<i>f</i>
E=0	<i>c</i>	<i>d</i>	<i>g</i>	<i>h</i>
OR (D-E)	<i>ad / cb</i>		<i>eh / gf</i>	
OR (G×E)	<i>adfg / bceh</i>			

$$\text{OR}(G \times E) = \text{OR}(G-E|D=1) / \text{OR}(G-E|D=0).$$

Assuming $\text{OR}(G-E|D=0)=1$ greatly reduces the variability in $\text{OR}(G \times E)$.

The case-only estimate of $\text{OR}(G \times E)$ is ag/ce .

The gain in power comes from the assumption of G-E independence, not the fact that only cases are used.

It is possible to build this assumption into the analysis of case-control data. These approaches retain the efficiency of the case-only test, but also allow for estimation of main effects of G and E, and estimates/tests of interaction effects other than departure from a multiplicative odds model.

The price for the increased power for the case-only test is increased Type I error rate if $OR(G-E | D=0) \neq 1$, i.e. if G and E are associated in controls.

How could this happen?

1. Population stratification
2. “E” is an intermediate on the $G \rightarrow D$ pathway

How likely is this?

1. Population stratification could affect many markers, but can also be controlled at design and analysis stage
2. A small number of markers out of the many many markers tested in a GWAS will affect E, and those may be known.

Leveraging G×E Interactions to Discover Risk Markers

- Joint test (binary or continuous outcomes)
- “Case-only” test (binary outcomes)
- Hedge methods (binary outcomes)
- Genotype-dependent variance methods (continuous outcomes)

Hedge Methods

Can we have our cake and eat it too?

- Empirical Bayes methods
 - Averages the standard logistic regression and case-only estimates of the interaction effect, weighted by evidence for/against G-E independence
- Two-step approaches
 - Screening step followed by testing step
 - Screening step may leverage G-E independence
 - Testing step robust to departures from G-E independence
 - Screening and test step chosen to be independent

Leveraging G×E Interactions to Discover Risk Markers

- Joint test (binary or continuous outcomes)
- “Case-only” test (binary outcomes)
- Hedge methods (binary outcomes)
- Genotype-dependent variance methods (continuous outcomes)

These tests are based on shifts in the mean trait values across $G \times E$ categories. What if we look at general differences in distribution across genotype?

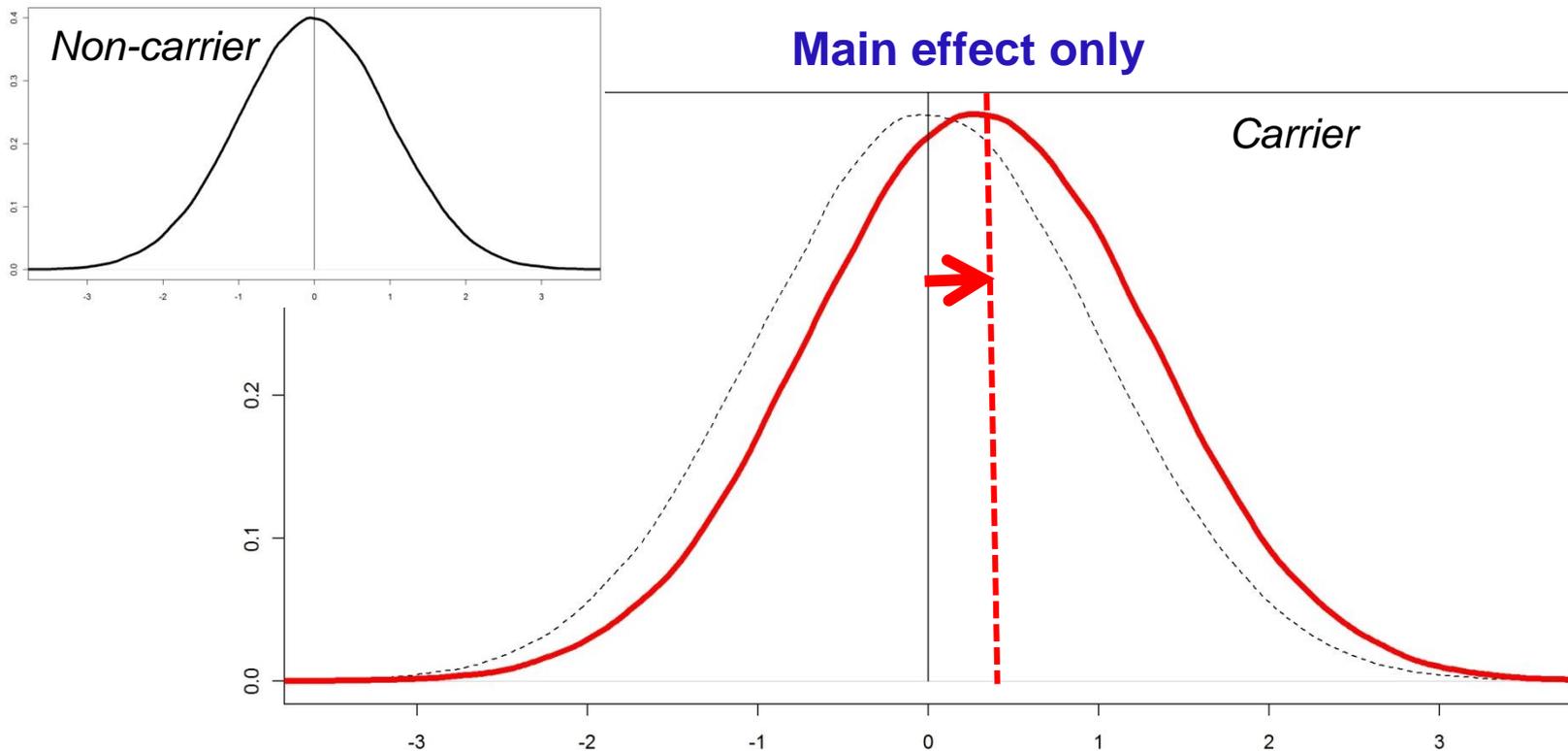
This allows us to scan for loci involved in $G \times E$ and $G \times G$ interactions without knowing or measuring the relevant E .



H. Aschard

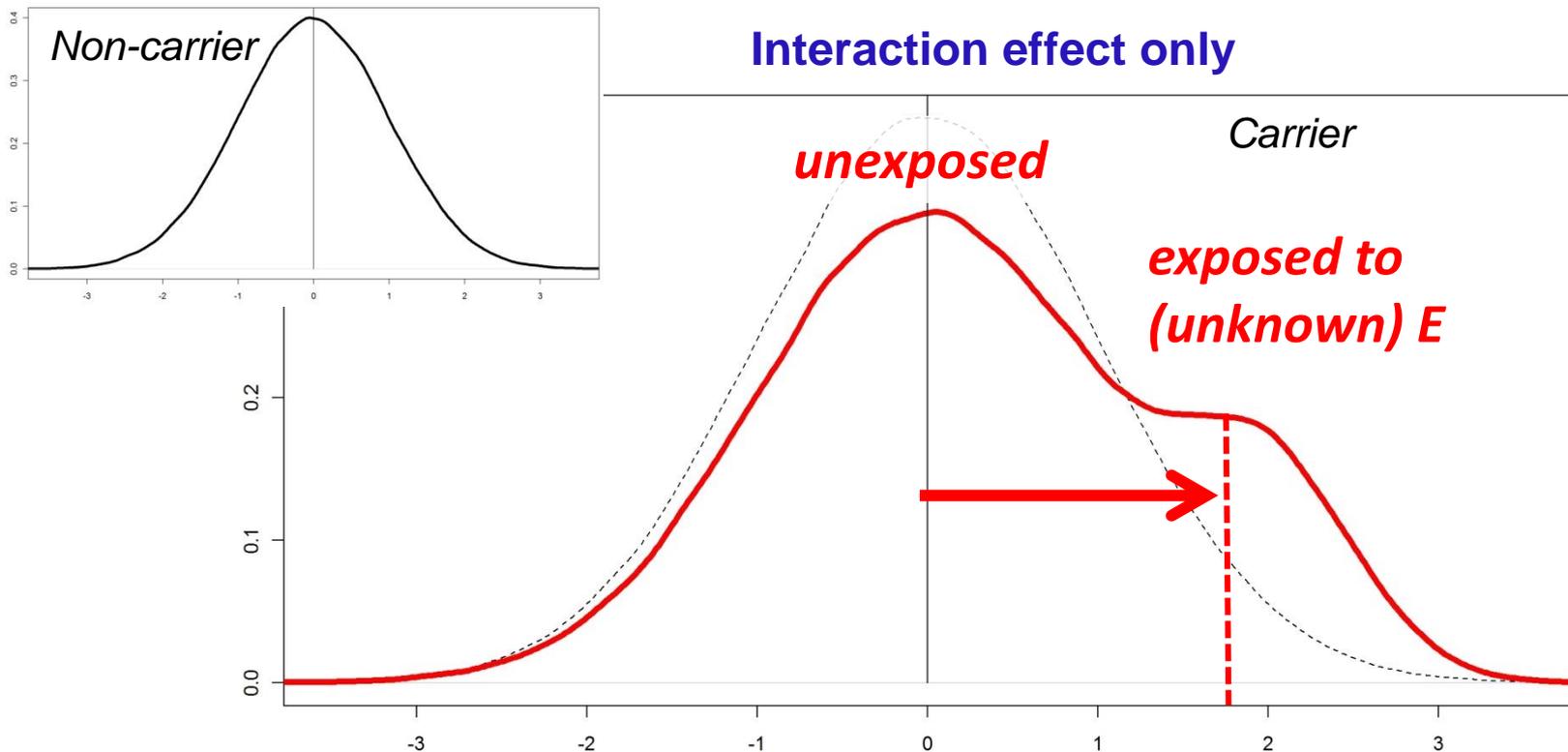
Method: Principles

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



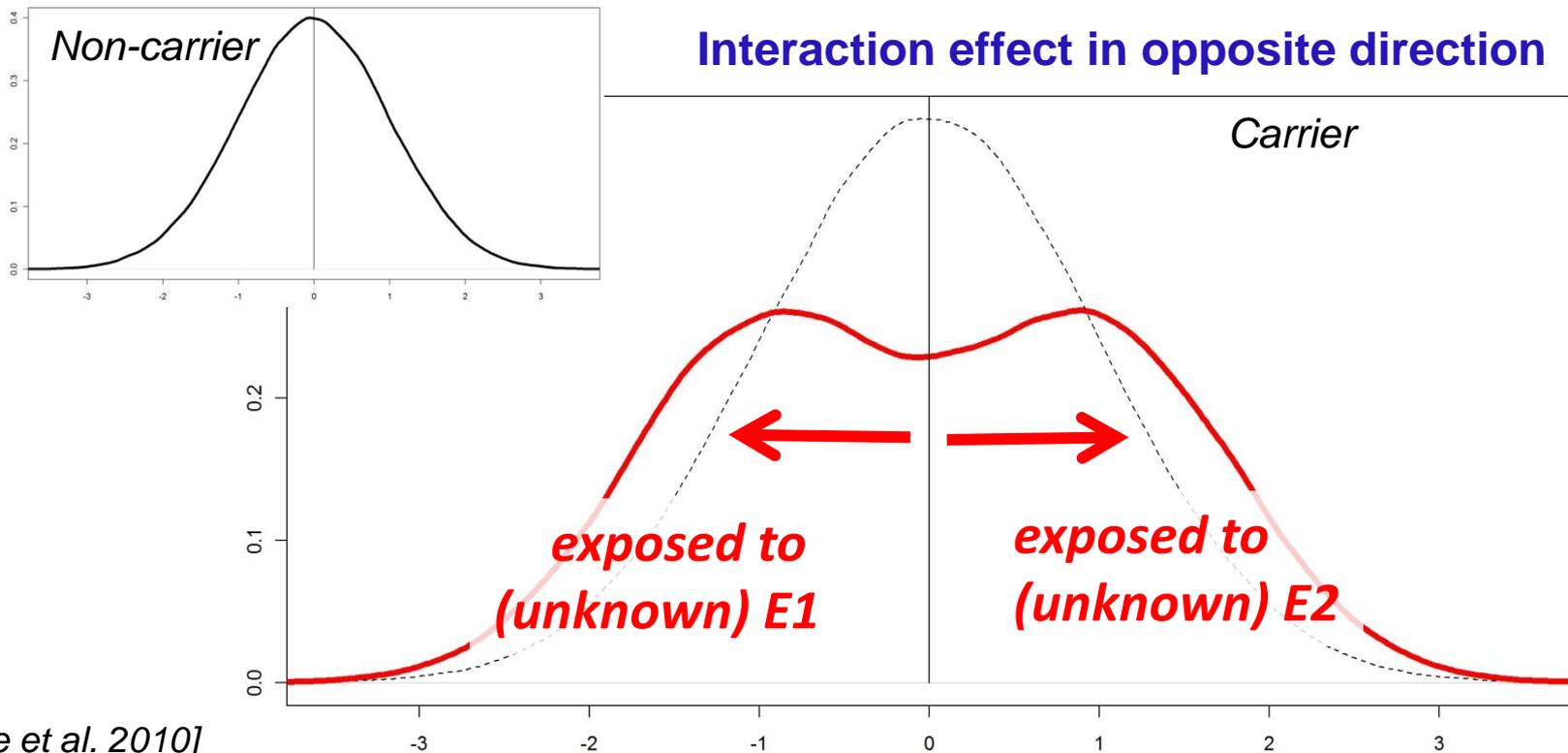
Method: Principles

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



Method: Principles

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



[Pare et al. 2010]

[Struchalin et al. 2010]

Method: Principles

LETTER

doi:10.1038/nature11401

***FTO* genotype is associated with phenotypic variability of body mass index**

A list of authors and their affiliations appears at the end of the paper.

There is evidence across several species for genetic control of phenotypic variation of complex traits¹⁻⁴, such that the variance among phenotypes is genotype dependent. Understanding genetic control of variability is important in evolutionary biology, agricultural selection programmes and human medicine, yet for complex traits, no individual genetic variants associated with variance, as opposed to the mean, have been identified. Here we perform a

environmental sensitivity so that genotypes differ in phenotypic variance. Therefore, even if the environments, internal or external, are not directly measured, evidence for genetic control of variation can be quantified through an analysis of variability.

There is empirical evidence for genetic control of phenotypic variation in several species¹, including *Drosophila*¹³, snails¹⁴, maize¹⁵ and chickens³, and specific quantitative trait loci with an effect on variance

OPEN ACCESS Freely available online

PLOS GENETICS

Inheritance Beyond Plain Heritability: Variance-Controlling Genes in *Arabidopsis thaliana*

Xia Shen^{1,2}, Mats Pettersson³, Lars Rönnegård^{2,3}, Örjan Carlborg^{1,3*}

¹Department of Cell and Molecular Biology, Uppsala University, Uppsala, Sweden, ²Statistics Unit, School of Technology and Business Studies, Dalarna University, Borlänge, Sweden, ³Department of Animal Breeding and Genetics, Swedish University of Agricultural Sciences, Uppsala, Sweden

Abstract

The phenotypic effect of a gene is normally described by the mean-difference between alternative genotypes. A gene may, however, also influence the phenotype by causing a difference in variance between genotypes. Here, we reanalyze a publicly available *Arabidopsis thaliana* dataset [1] and show that genetic variance heterogeneity appears to be as common as normal additive effects on a genome-wide scale. The study also develops theory to estimate the contributions of variance differences between genotypes to the phenotypic variance, and this is used to show that individual loci can explain more than 20% of the phenotypic variance. Two well-studied systems, cellular control of molybdenum level by the ion-transporter *MOT1* and flowering-time regulation by the *FRI-FLC* expression network, and a novel association for *Leaf serration* are used to illustrate the contribution of major individual loci, expression pathways, and gene-by-environment interactions to the genetic variance heterogeneity.

Copyright © 2011 by the Genetics Society of America
DOI: 10.1534/genetics.111.127068

Detecting Major Genetic Loci Controlling Phenotypic Variability in Experimental Crosses

Lars Rönnegård^{*1} and William Valdar[†]

^{*}Statistics Unit, Dalarna University, SE-781 70 Borlänge, Sweden and [†]Department of Genetics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599-7265

Manuscript received January 20, 2011
Accepted for publication March 21, 2011

OPEN ACCESS Freely available online

PLOS GENETICS

Genomic Analysis of QTLs and Genes Altering Natural Variation in Stochastic Noise

Jose M. Jimenez-Gomez^{1*}, Jason A. Corwin^{2*}, Bindu Joseph^{2*}, Julin N. Maloof¹, Daniel J. Kliebenstein^{2*}

¹Department of Plant Biology, University of California Davis, Davis, California, United States of America, ²Department of Plant Sciences, University of California Davis, Davis, California, United States of America

Method: Testing for association

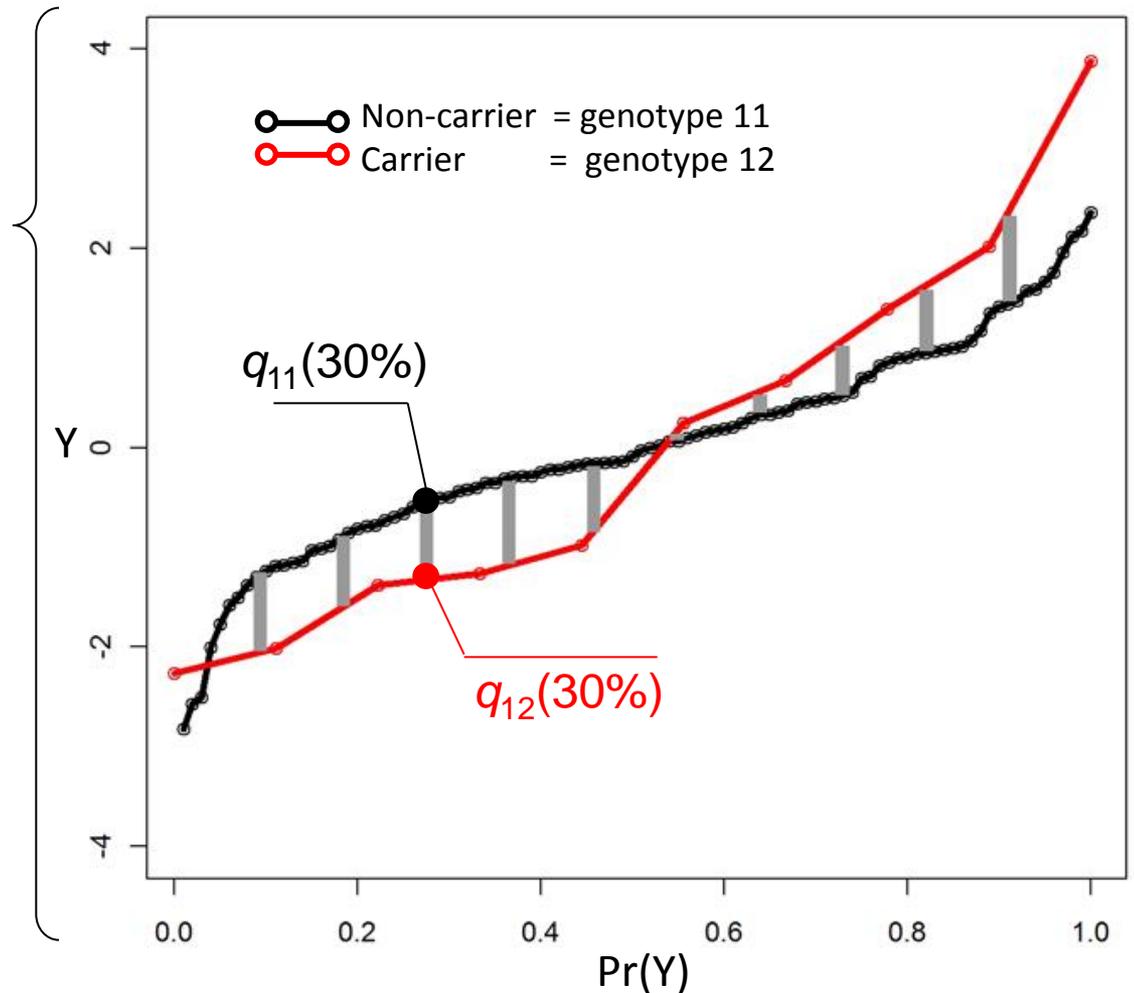
Sum of differences of quantiles across multiple points

$$S_{11 \leftrightarrow 12} = \sum_{i=1}^n (q_{11}^i - q_{12}^i)^2$$

$$T = S_{11 \leftrightarrow 12} + S_{12 \leftrightarrow 22} + S_{11 \leftrightarrow 22}$$

Significance of T derived by permutation

$$\text{p-val} = \frac{\text{count}(T_{\text{obs}} < T_{\text{perm}})}{N \text{ permutation} + 1}$$



(Partial) List of Important Topics I Do Not Have Time to Discuss

- Biased testing due to mis-modeling main effects (& fixes)
- Meta-analysis
- Impact of measurement error
- Confounders—when and how to adjust
- Study design (prospective, retrospective, oversampling)
- Considerations when characterization (clinically relevant interactions, biological interactions) rather than discovery is the goal

See Appendix...

Outline

- Definition and Notation
- Leveraging G×E Interactions to Discover Risk Markers
- **State of the science: cancer and obesity**
- Practicalities

Of the 407 articles, **307 articles** reported a significant gene-environment interaction.

Are these credible?

Of the 407 articles, **307 articles** reported a significant gene-environment interaction.

Are these credible?

Probably not.

1. Small sample sizes.
2. No correction for multiple testing/low priors.
3. Little in the way of replication.

What about large studies examining interactions between GWAS-identified markers and established risk factors?

DOI: 10.1093/jnci/djq265
 Advance Access publication on July 26, 2011.

© The Author 2011. Published by Oxford University Press. All rights reserved.
 For Permissions, please e-mail: journals.permissions@oup.com.

ARTICLE

Interactions Between Genetic Variants and Breast Cancer Risk Factors in the Breast and Prostate Cancer Cohort Consortium

Daniele Campa, Rudolf Kaaks, Loïc Le Marchand, Christopher A. Haiman, Ruth C. Travis, Christine D. Berg, Julie E. Buring, Stephen J. Chanock, W. Ryan Diver, Lucie Dostal, Agnes Fournier, Susan E. Hankinson, Brian E. Henderson, Robert N. Hoover, Claudine Isaacs, Mattias Johansson, Laurence N. Kolonel, Peter Kraft, H-Min Lee, Catherine A. McCarty, Kim Overvad, Salvatore Panico, Petra H.M. Peeters, Elio Riboli, María José Sánchez, Fredrick R. Schumacher, Guri Skeie, Daniel O. Stram, Michael J. Thun, Dimitrios Trichopoulos, Shumin Zhang, Regina G. Ziegler, David J. Hunter, Sara Lindström, Federico Canzian

		Cases	Controls
Travis et al.	Lancet (2010)	7,160	10,196
Milne et al.	Br Can Res (2010)	26,349	32,208
Campa et al.	JNCI (2011)	8,576	11,892

We confirmed the association of 14 SNPs with breast cancer risk ($P_{\text{total}} = 2.07 \times 10^{-4}$ – 3.26×10^{-7}). Three SNPs (*LSP1*-rs3817198, *COL1A1*-rs2075555, and *RNF146*-rs2180341) did not show association with breast cancer risk. After accounting for multiple testing, no statistically significant interactions were detected between the 17 SNPs and the nine risk factors. We also confirmed that SNPs in *EGFR2* and *TNRC9* were associated with greater

No statistical evidence of interaction was observed beyond that expected by chance

identified by several genome-wide association studies (GWAS) or studies of specific candidate single-nucleotide polymorphisms (SNPs) (1–10). Genetic variants that showed strong statistically significant associations with breast cancer risk (odds ratios [ORs] = 1.15–1.45, $P < 5 \times 10^{-7}$) were identified in fibroblast growth factor receptor 2 (*FGFR2*). The vicinity is referred to all genes, not only

graph are located directly within the mentioned genes, some others are near the genes. TOX high mobility group box family member 3 (*TOX3*; also known as *TNRC9*), mitogen-activated protein kinase kinase kinase 1 (*MAP3K1*), caspase 8 (*CASP8*), lymphocyte-specific protein 1 (*LSP1*), collagen type I alpha 1 (*COL1A1*), cytochrome c oxidase assembly homolog 11 (*COX11*),

Ger
can

Ruth C
Travis

Summary
Background
Study Design
Results
Conclusions

Meth
10 196
2435-1
LSP1-
rs3817198
breast
height

Findings
Genome-wide
association
studies
162-4

Interp
signifi

Fundin

Intro
Genes
of spe
cific
loci
affect
breast
cancer
risk

At
intro
air
car
As
Me
nu
po
JM
3p

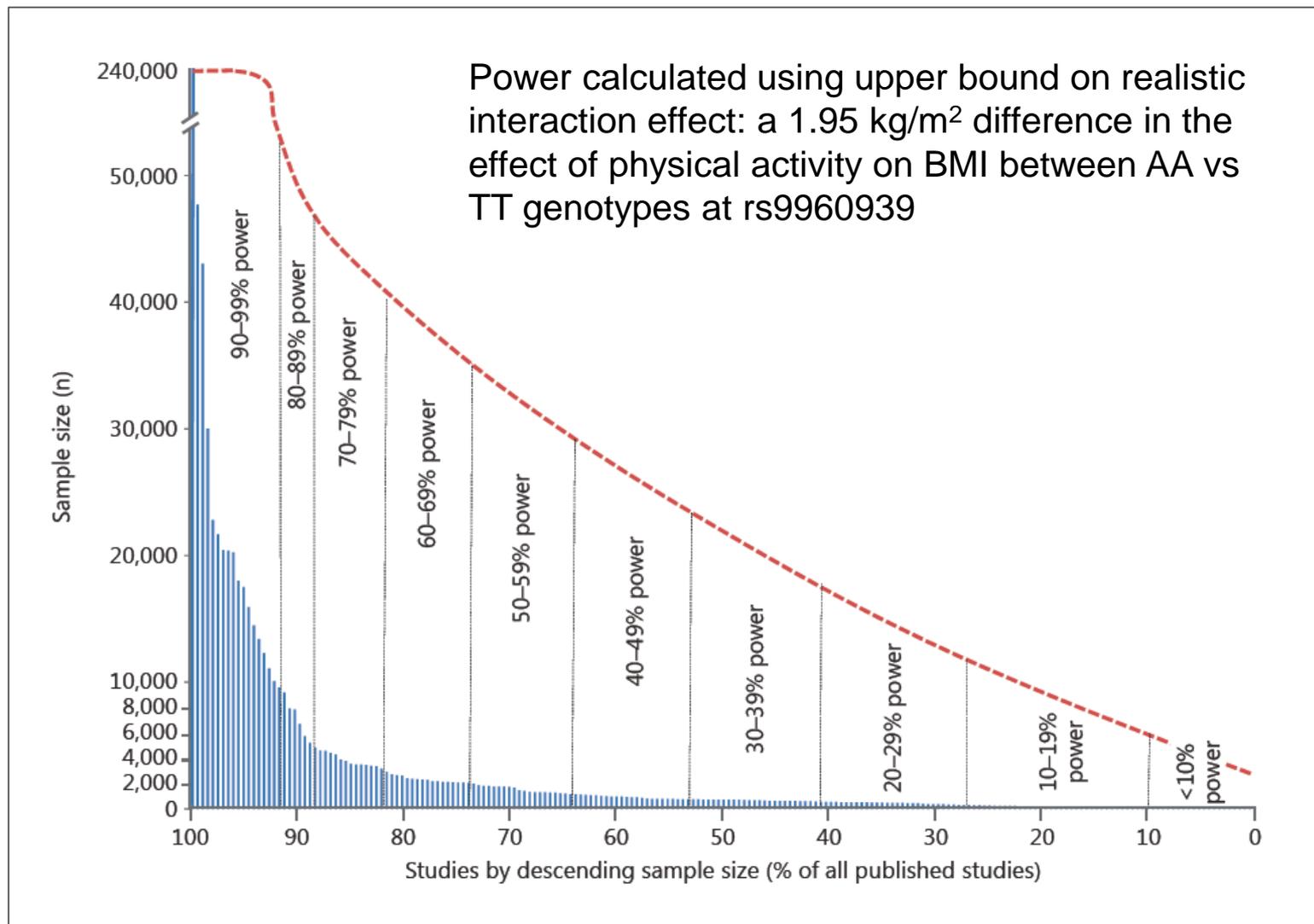
Meth
Partic
Durin
age 51

www.jnci.org

Replicated, Credible Interactions*

Site	Interaction	OR _{GE}
Breast	<i>Bupkes</i>	-
Prostate	<i>Nichts</i>	-
Colon	<i>It's Complicated</i>	-
Bladder	<i>NAT2</i> and smoking	1.29
Esophogeal	<i>ALDH2</i> and drinking	1.31
Lung	<i>CHRNA3/5</i> and smoking	1.21

*Departures from a multiplicative odds model



Color version available online

Fig. 1. Relationship between sample sizes of studies of gene × lifestyle interactions in obesity (n = 210 studies) published since 1995 and power to detect the interaction effect reported by Andreasen et al. [59] for *FTO* (rs9960939) × physical activity on BMI (dashed curve).

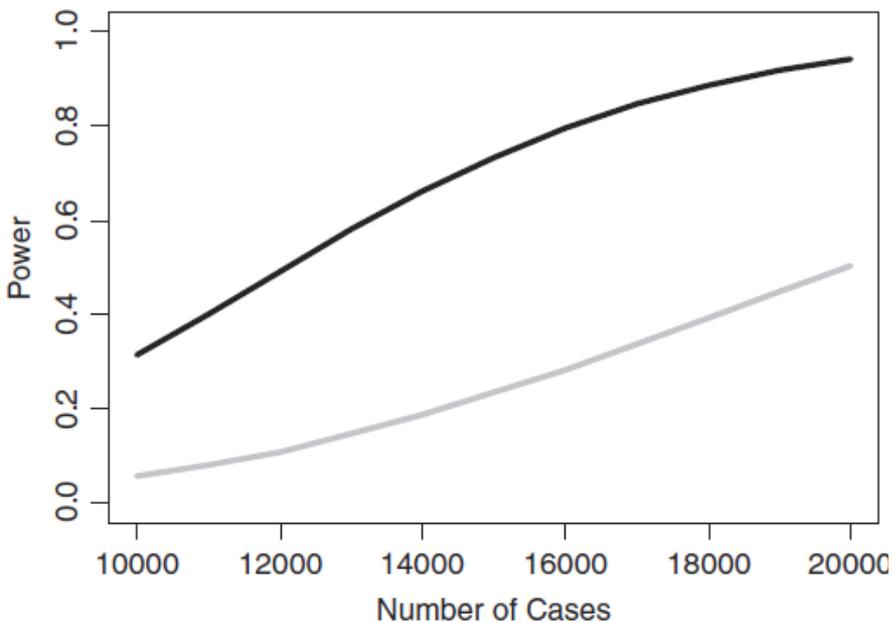
Outline

- Definition and Notation
- Leveraging G×E Interactions to Discover Risk Markers
- State of the science: cancer and obesity
- **Practicalities**

Practicalities (among many)

- Sample size
- Harmonization
- Range of exposure

Exposure Prevalence : 33%



Exposure Prevalence : 10%

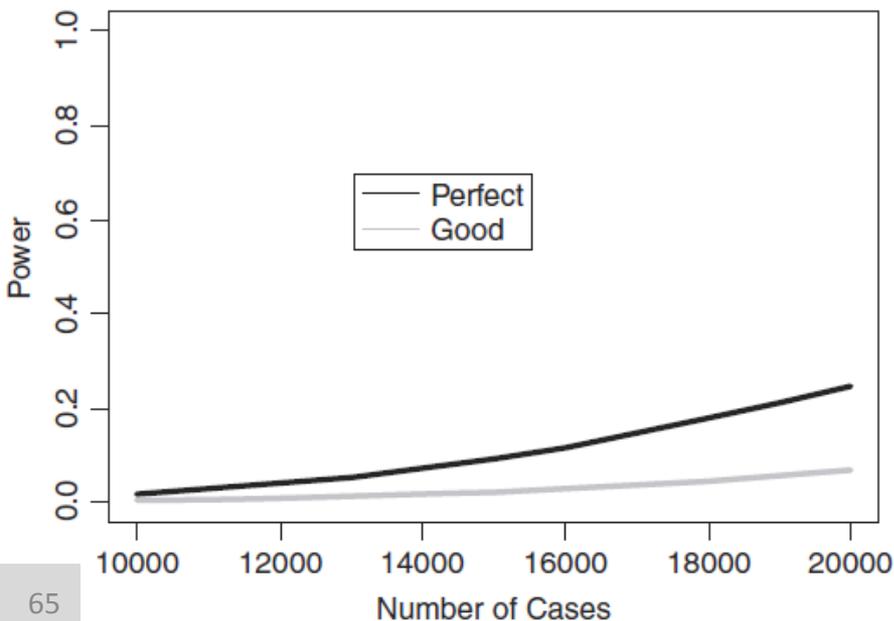


Fig. 2. Power of a case-control study to detect a gene-environment interaction (departure from a multiplicative odds model) when the binary exposure is measured perfectly or via a good proxy with 77% specificity and 99% sensitivity (roughly analogous to self-reported versus measured overweight status). This figure illustrates several points: (a) large sample sizes are needed to detect gene-environment interactions; (b) even modest misclassification can greatly decrease the power of tests for gene-environment interaction (and the relative decrease is greater for rare exposures); yet (c) a large study using the proxy can have greater power than a smaller study using the perfect measure. This last point is important when the perfect measure is prohibitively expensive or only available on a small fraction of samples, while the good measure is relatively inexpensive or already available on many samples. Power calculations were performed using the methods described in Lindström et al. [2009], assuming a rare disease (prevalence 1 in 1,000), no main effect for the binary genetic factor (with 20% prevalence), an odds ratio of 1.5 for the exposure, an interaction odds ratio of 1.35, and a Type I error rate of 5×10^{-8} .

FTO, Physical Activity and BMI

Kilpelainen et al. (2011). PLoS Medicine. 8(11). e1001116

- Meta-analysis of 218,166 European-ancestry subjects
- Risk of Obesity (BMI ≥ 30 vs. BMI < 25 kg/m²) 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



New Delhi



Trivandrum

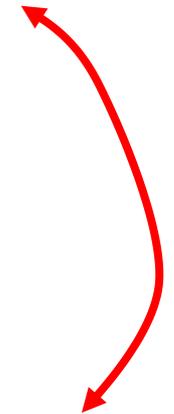


Participant characteristics by region

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 _{trend}	Interaction by PA
Overall	1,209	+1.61 cm (0.67, 2.55)	0.0008	0.009
New Delhi				
Overall	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				
Overall	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 et al
(2011), Obesity

The flipside: can increase power to detect interaction by increasing the range of genetic exposure measured and tested

Most work has focused on pairwise interactions. Considering aggregate evidence for interaction may be useful.

Lindstrom et al. (CEBP 2012) find evidence that effect of a prostate-cancer SNP score differs by age; Qi et al. (NEJM 2012) show the effect of a BMI SNP score differs by intake of sugar-sweetened beverages.

But these approaches assume you have a defined set of SNPs with common biological effect and known allelic effects (i.e. you know which allele is likely deleterious).

Recap

Why study genes and environment?

- Leverage assumed effect modifiers to increase power
- Provide insights into biological mechanism
- Improve risk prediction and prognostic models and strategies for risk prediction

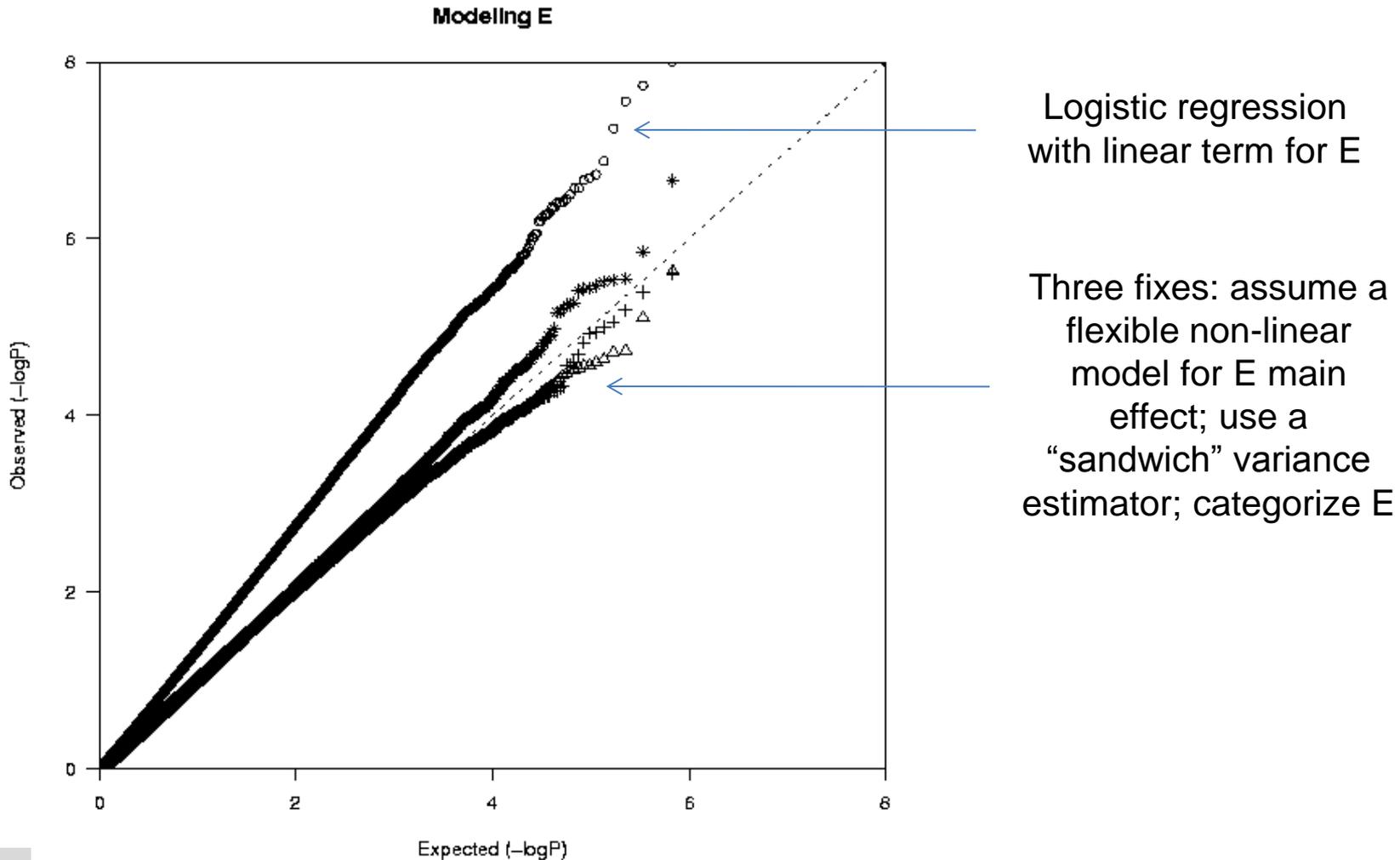
Statistical interaction *per se* generally offers at most circumstantial evidence to address any of these goals

Challenges

- The study of gene-environment interaction arguably combines the toughest aspects of both environmental and genetic epidemiology
 - From genetic epidemiology: problems associated with high-dimensional data with sparse and small effects
 - From environmental epidemiology: problems associated with measurement error, range and timing of exposure
- And sample sizes needed to reliably detect gene-environment interaction are typically quite large

Appendix

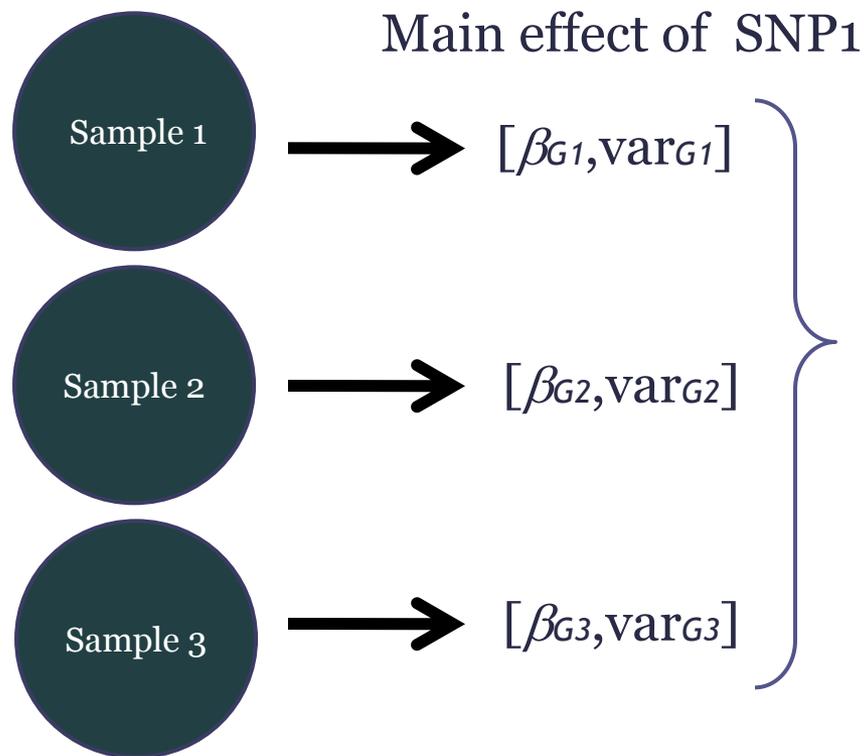
Misspecification of the main effect of E can lead to inflated Type 1 error rate



Methods for Meta-Analysis

Methods

Meta-analysis of a single parameter



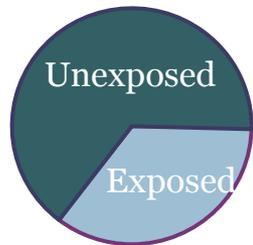
β_G can be estimated using an inverse variance weighted sum

Overall
$\frac{\left(\sum \frac{\hat{\beta}_{G.i}}{\text{var}_{G.i}} \right)^2}{\sum \frac{1}{\text{var}_{G.i}}}$

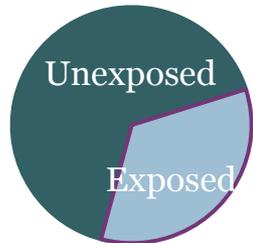
> The weighted sum is following a 1df chi2 under the null hypothesis

Methods

Meta-analysis of a multiple parameters



Main effect of SNP1 $\Rightarrow [\beta_{G1}, \text{var}_{G1}]$ Int. effect of SNP1 and E $[\beta_{GE1}, \text{var}_{GE1}]$



$\Rightarrow [\beta_{G2}, \text{var}_{G2}]$ $[\beta_{GE2}, \text{var}_{GE2}]$



$\Rightarrow [\beta_{G3}, \text{var}_{G3}]$ $[\beta_{GE3}, \text{var}_{GE3}]$

■ Unexposed subject ■ Exposed subject

$\beta = (\beta_G, \beta_{GE})$ can be estimated by MLE where $l(\beta)$ is equal to:

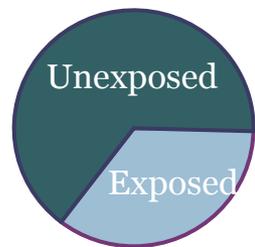
$$-\frac{1}{2} \sum (\beta - \hat{\beta}_i)^T \Sigma_i^{-1} (\beta - \hat{\beta}_i)$$

> Test for association using a Score test or Wald test

[Aschard et al. Hum Hered]
[Manning et al. Genet Epidemiol]

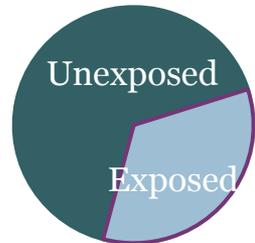
Methods

Meta-analysis of a multiple parameters



Main effect of SNP1 Int. effect of SNP1 and E

→ $[\beta_{G1}, \text{var}_{G1}]$ $[\beta'_{G1}, \text{var}_{GE1}]$



→ $[\beta_{G2}, \text{var}_{G2}]$ $[\beta'_{GE2}, \text{var}_{GE2}]$



→ $[\beta_{G3}, \text{var}_{G3}]$ $[\beta'_{GE3}, \text{var}_{GE3}]$

■ Unexposed subject ■ Exposed subject

$\beta = (\beta_G, \beta_{GE})$ can be estimated by summing effect on strata

Unexposed	+	Exposed
$\frac{\left(\sum \frac{\hat{\beta}'_{G.i}}{\text{var}'_{G.i}} \right)^2}{\sum \frac{1}{\text{var}'_{G.i}}}$		$\frac{\left(\sum \frac{\hat{\beta}_{G.i}}{\text{var}_{G.i}} \right)^2}{\sum \frac{1}{\text{var}_{G.i}}}$

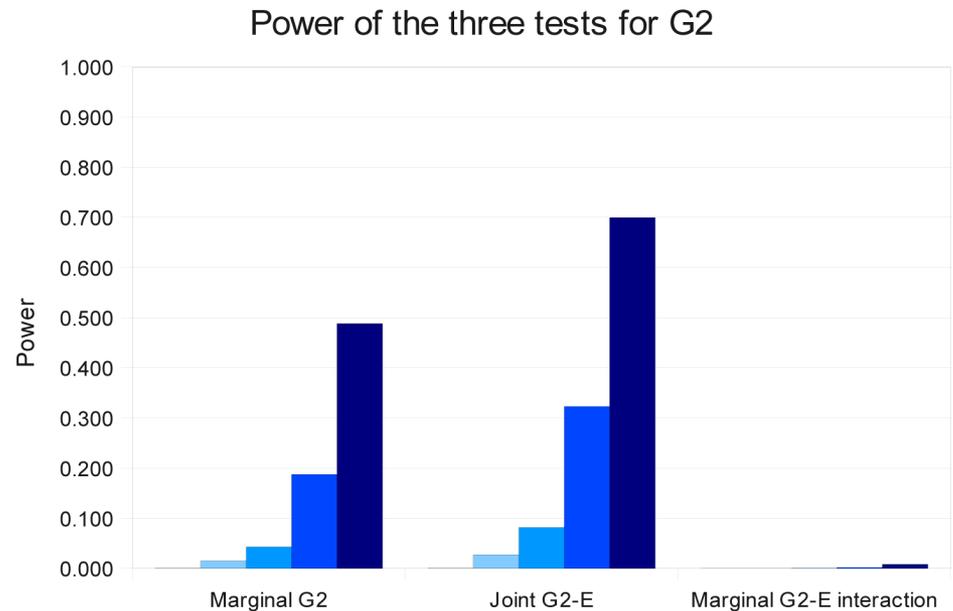
> The sum is following a 2df chi2 under the null hypothesis

Results

Power to detect G2 function of sample size

For a single realization of \mathbf{Y} , we compute the power of the 3 tests while increasing sample size

NHS CHD	1145 +
NHS T2D	3110 +
NHS BrCa	2285 +
HPFS CHD	1311 +
HPFS T2D	2310 = 10,161



> Regardless of the test used, large sample will be needed to reliably detect genes with subtle gene-environment interaction patterns.

Testing for Additive Interaction in Case-Control Data

When to report additive or
multiplicative interaction

Testing for departures from additivity
on the absolute risk scale when you
have case-control data

Relative Excess Risk due to Interaction

- We can use a clever trick to test for non-additivity
 - $I_{11} - (I_{01} - I_{00}) - (I_{10} - I_{00}) - I_{00} = 0 \Rightarrow RR_{11} = RR_{10} + RR_{01} - 1$
 - $RERI = RR_{11} - RR_{10} - RR_{01} + 1$
- This is no longer a generalized linear model
 - Can't fit using standard logistic regression software, e.g.
 - Have to use custom code (e.g. PROC NLMIXED)

Likelihood Ratio Test

```
proc nlmixed data=twosnp;  
  if (g eq 0) and (e eq 0) then eta=a;  
  if (g eq 0) and (e eq 1) then eta=a+b2;  
  if (g eq 1) and (e eq 0) then eta=a+b1;  
  if (g eq 1) and (e eq 1) then eta=a+log(exp(b1)+exp(b2) - 1);  
  ll = caco*eta + (1-caco)*log(1+exp(eta));  
  model caco ~ general(ll);  
  parms a b1 b2=0;  
run;
```

Null Model

(interaction constrained to be additive on risk scale)

```
proc nlmixed data=twosnp;  
  if (g eq 0) and (e eq 0) then eta=a;  
  if (g eq 0) and (e eq 1) then eta=a+b2;  
  if (g eq 1) and (e eq 0) then eta=a+b1;  
  if (g eq 1) and (e eq 1) then eta=a+b3;  
  ll = caco*eta + (1-caco)*log(1+exp(eta));  
  model caco ~ general(ll);  
  parms a b1 b2=0;  
run;
```

Alternative Model

(interaction not constrained)

Compare $-2 \log L_{\text{null}} + 2 \log L_{\text{alt}}$ to chi-square 1 d.f.

Testing for additive interactions using case-control data is less straightforward. Under the null hypothesis of additivity on the absolute scale, $RR_{G_1G_2} = RR_{G_1} + RR_{G_2} - 1$, where $RR_{G_1G_2}$ is the relative risk for a woman with genotype G_1 at locus 1 and G_2 at locus 2, and RR_{G_1} (RR_{G_2}) is the marginal relative risk for genotype G_1 (G_2). Thus testing whether the Relative Excess Risk due to Interaction ($RERI$) = $RR_{G_1G_2} - (RR_{G_1} + RR_{G_2} - 1) = 0$ is equivalent to testing for additive interaction.^{105,107} Testing $RERI=0$ can be done by fitting the alternative model [E2] and constructing an appropriate point estimate and confidence interval for $RERI$ using the fitted odds ratios $OR_{G_1G_2} = \exp[\beta_{G_1} G_1 + \beta_{G_2} G_2 + \beta_{G_1G_2} G_1 G_2]$ etc., or by comparing [E2] to the constrained, non-linear logistic model

$$\text{log odds of breast cancer (given } G_2=0) = \alpha + \beta_X'X + \beta_{G_1} G_1 \tag{E4.a}$$

$$\text{log odds of breast cancer (given } G_1=0) = \alpha + \beta_X'X + \beta_{G_2} G_2 \tag{E4.b}$$

$$\text{log odds of breast cancer (} G_1 \neq 0, G_2 \neq 0) = \alpha + \beta_X'X + \log [\exp(\beta_{G_1} G_1) + \exp(\beta_{G_2} G_2) - 1]. \tag{E4.c}$$

(This is equivalent to the linear odds model described in Richardson and Kaufman¹⁰⁷ and can be fit using nonlinear function maximizers in standard software packages, e.g. PROC NL MIXED in SAS or the nlm() function in R.) **There are two potential drawbacks to using the RERI approach to testing for additive interaction in this context. First, we will rely on the odds ratio approximation to the relative risk.¹⁰⁸ Considering breast cancer is a relatively rare disease and the individual allelic relative risks are small, the odds ratio should be a good approximation to the relative risk. Second, if $\beta_X \neq 0$ the RERI varies across strata defined by the covariates X ; so the estimated RERI derived by the procedures described above does not necessarily estimate the RERI in any particular stratum, rather it represents an average RERI.¹⁰⁹ (Tests for the null that $RERI=0$ for all strata have appropriate Type I error, however.)**

- 105. Greenland S, Rothman K. Concepts of interaction. In: Rothman K, Greenland S, eds. *Modern Epidemiology*. Philadelphia: Lippincott Williams & Wilkins, 1998.
- 106. Greenland S. Interactions in epidemiology: relevance, identification, and estimation. *Epidemiology* 2009;20(1):14-7.
- 107. Richardson DB, Kaufman JS. Estimation of the relative excess risk due to interaction and associated confidence bounds. *Am J Epidemiol* 2009;169(6):756-60.
- 108. Kalilani L, Atashili J. Measuring additive interaction using odds ratios. *Epidemiol Perspect Innov* 2006;3:5.
- 109. Skrdal A. Interaction as departure from additivity in case-control studies: a cautionary note. *Am J Epidemiol* 2003;158(3):251-8.

Recommendations for presenting analyses of effect modification and interaction

Mirjam J Knol^{1*} and Tyler J VanderWeele^{2,3}

Int J Epidemiol (2012)

- Present effect measures for each GxE category
- Present tests for both additive and multiplicative int.

Impact of departures from gene-environment independence on “case-only” style tests in the context of GWAS

The price for the increased power for the case-only test is increased Type I error rate if $OR(G-E | D=0) \neq 1$, i.e. if G and E are associated in controls.

How could this happen?

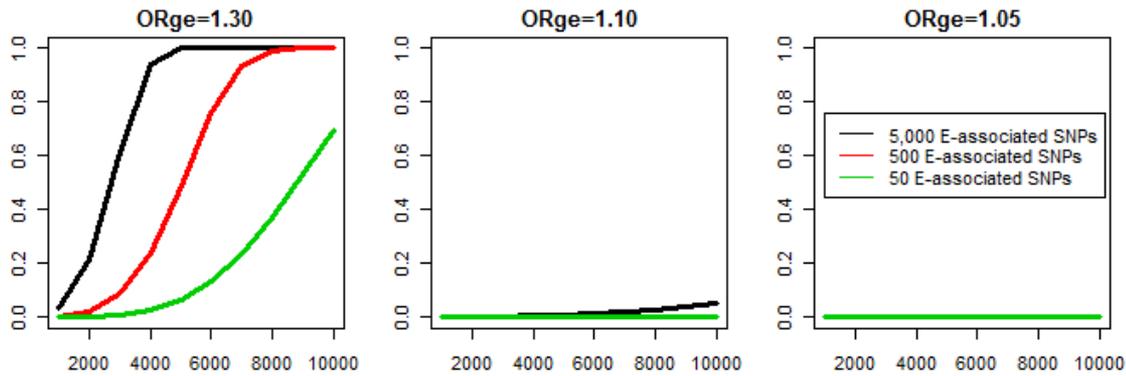
1. Population stratification
2. “E” is an intermediate on the $G \rightarrow D$ pathway

How likely is this?

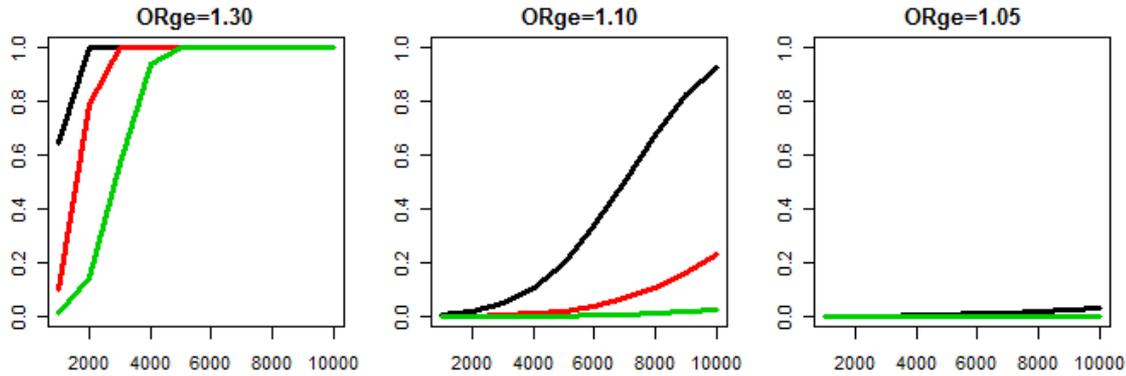
1. Population stratification could affect many markers, but can also be controlled at design and analysis stage
2. A small number of markers out of the many many markers tested in a GWAS will affect E, and those may be known.

Genome-wide Type I error rate for case-only test

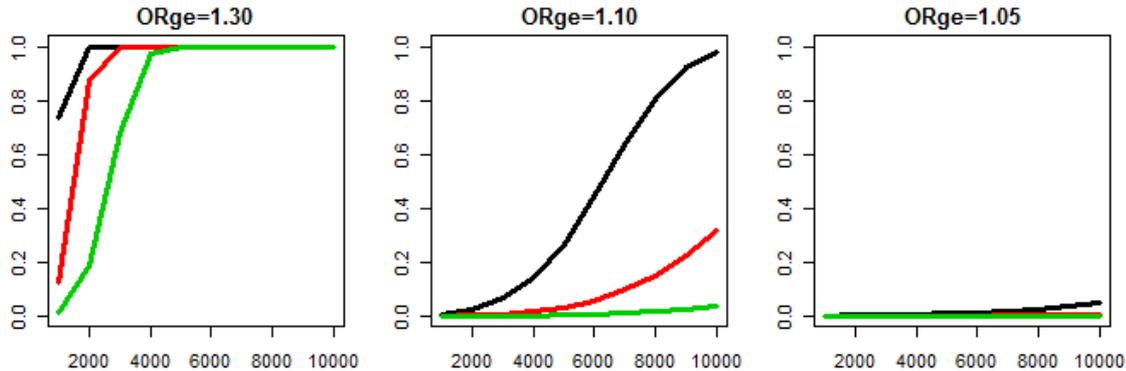
Exposure
Prevalence
5%



Exposure
Prevalence
25%



Exposure
Prevalence
40%



Three Reasonably Up-to-date
Overviews of Statistical Methods
for GxE Interactions for Binary
Outcomes—

*And an Interesting Observation about
What Can Happen when G-E Correlation
and G-E Interactions Go in Opposite
Directions*

Testing Gene-Environment Interaction in Large Scale Case-Control Association Studies: Possible Choices and Comparisons

Mukherjee B, Ahn J, Gruber SB, Chatterjee N.

Am J Epidemiol (2012)

Gene-environment interactions in genome-wide association studies: A comparative study of tests applied to empirical studies of type 2 diabetes

Cornelis MC, Tchetgen Tchetgen E, Liang L, Chatterjee N, Hu FB, Kraft P

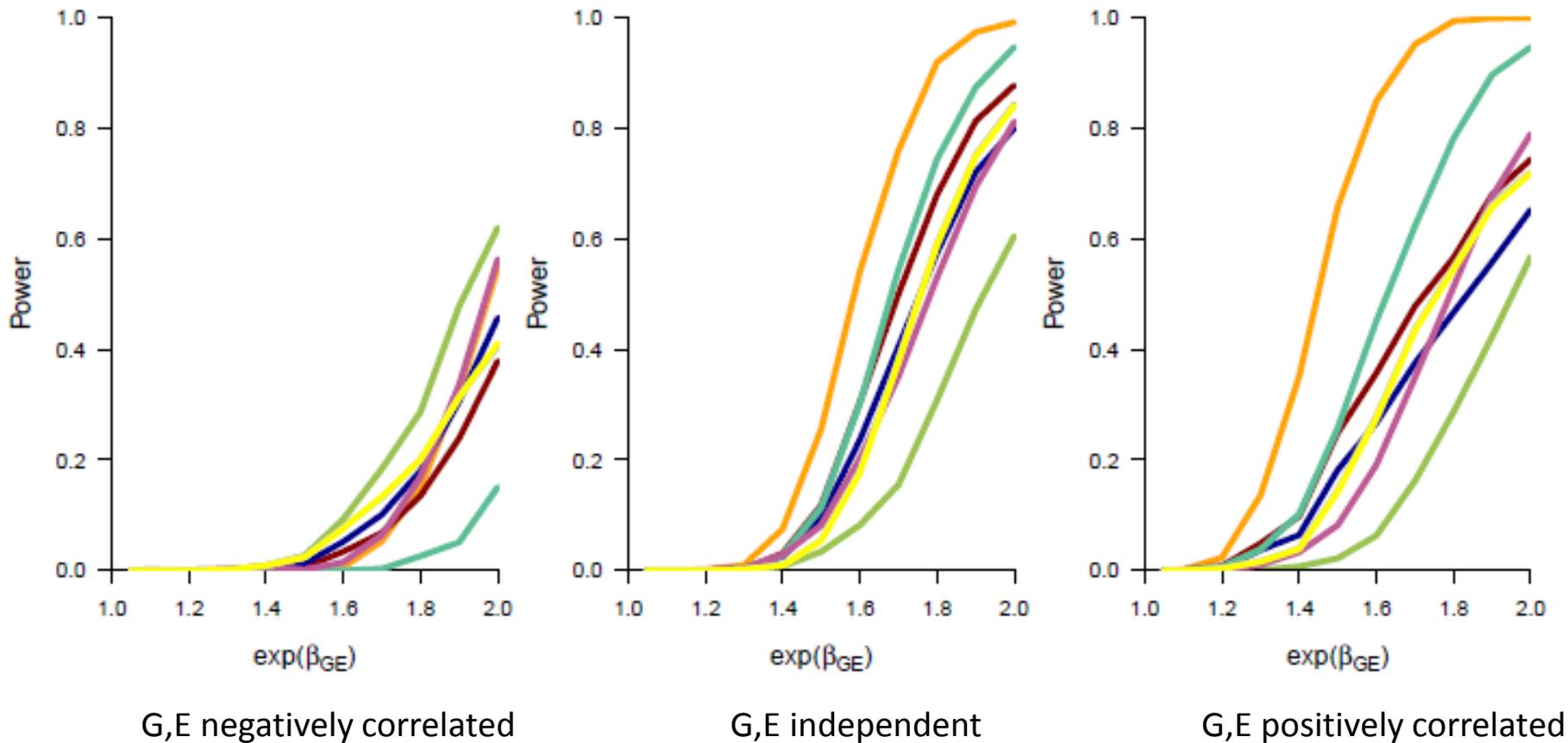
Am J Epidemiol (2012)

GE-Whiz! Ratcheting Gene-Environment Studies up to the Whole Genome and the Whole Exposome.

Thomas DC, Lewinger JP, Murcray CE, Gauderman WJ

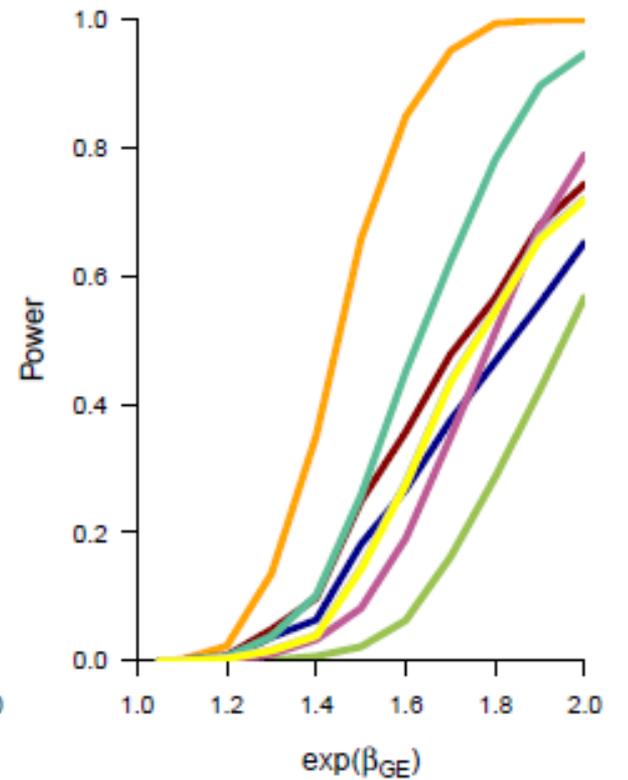
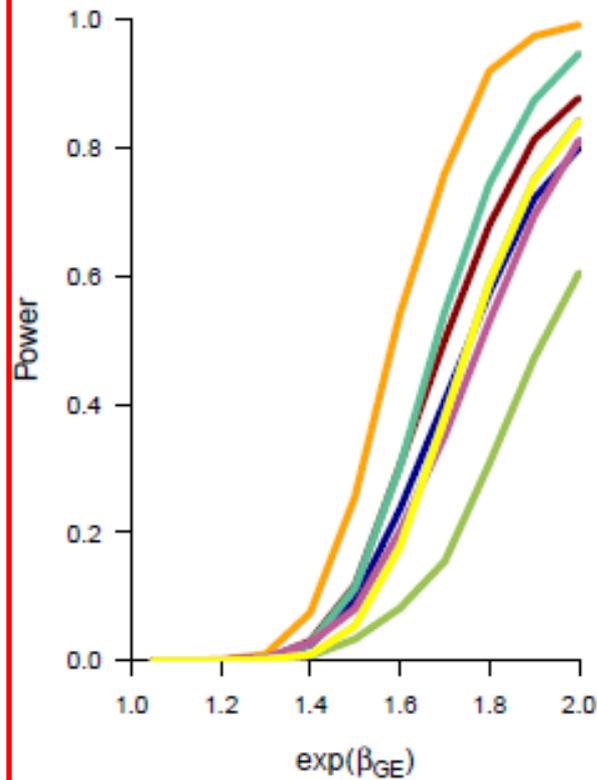
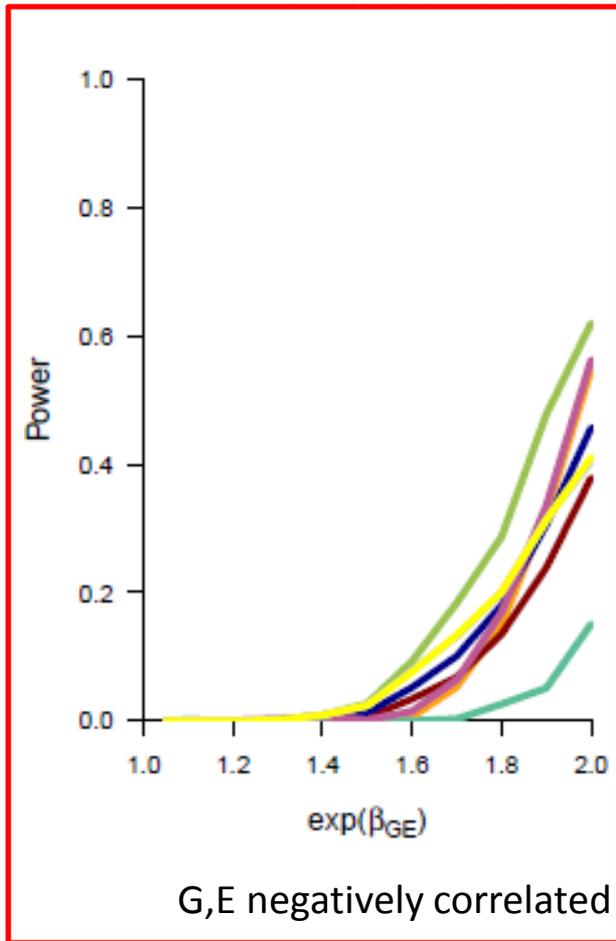
Am J Epidemiol (2012)

Power for 8 different approaches



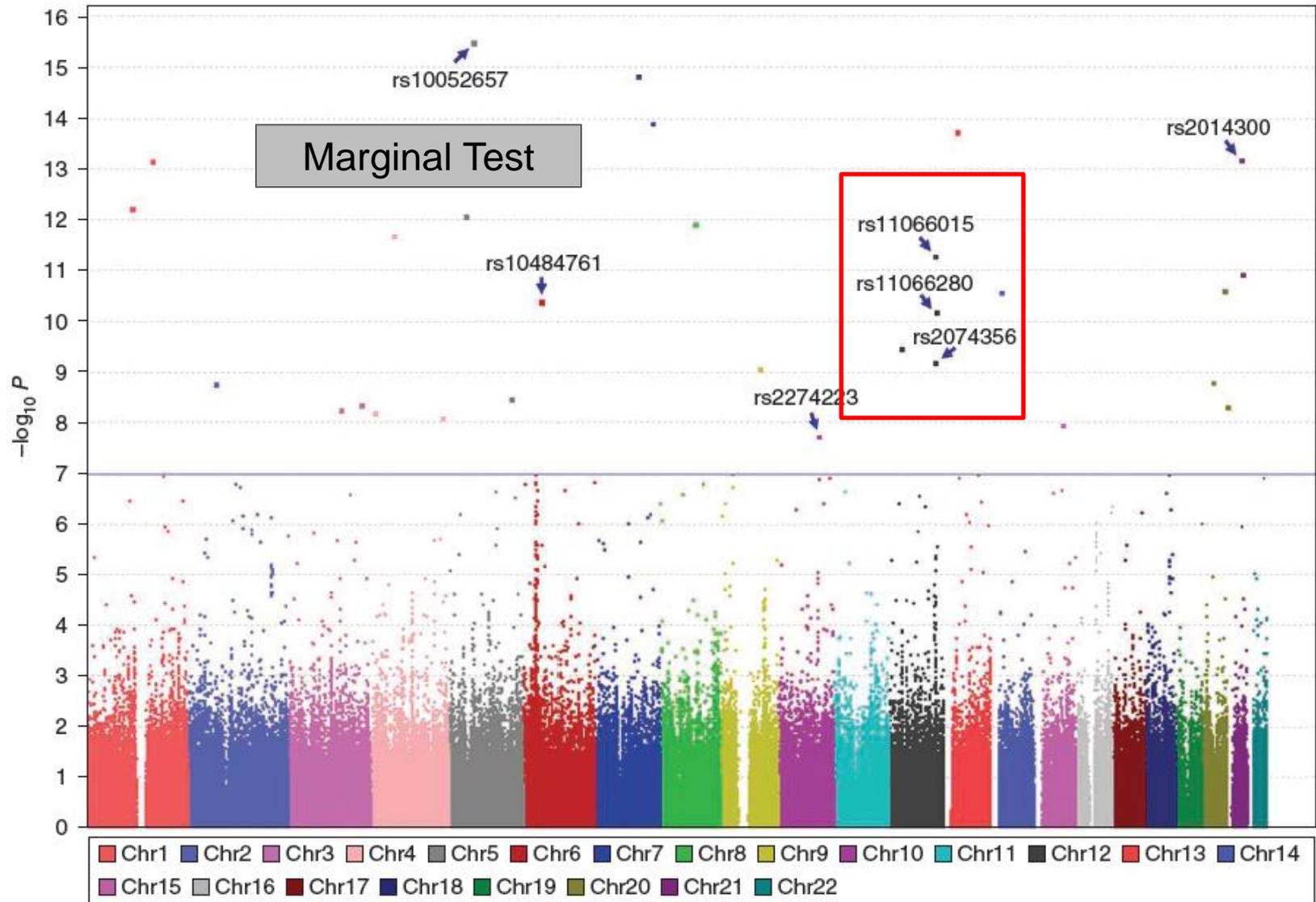
Color code: CC — CO — TS ($\alpha_1 = 5 \times 10^{-4}$) — TS ($\alpha_1 = 5 \times 10^{-2}$) — EB — EB2 — AIC — BMA —

Power for 8 different approaches

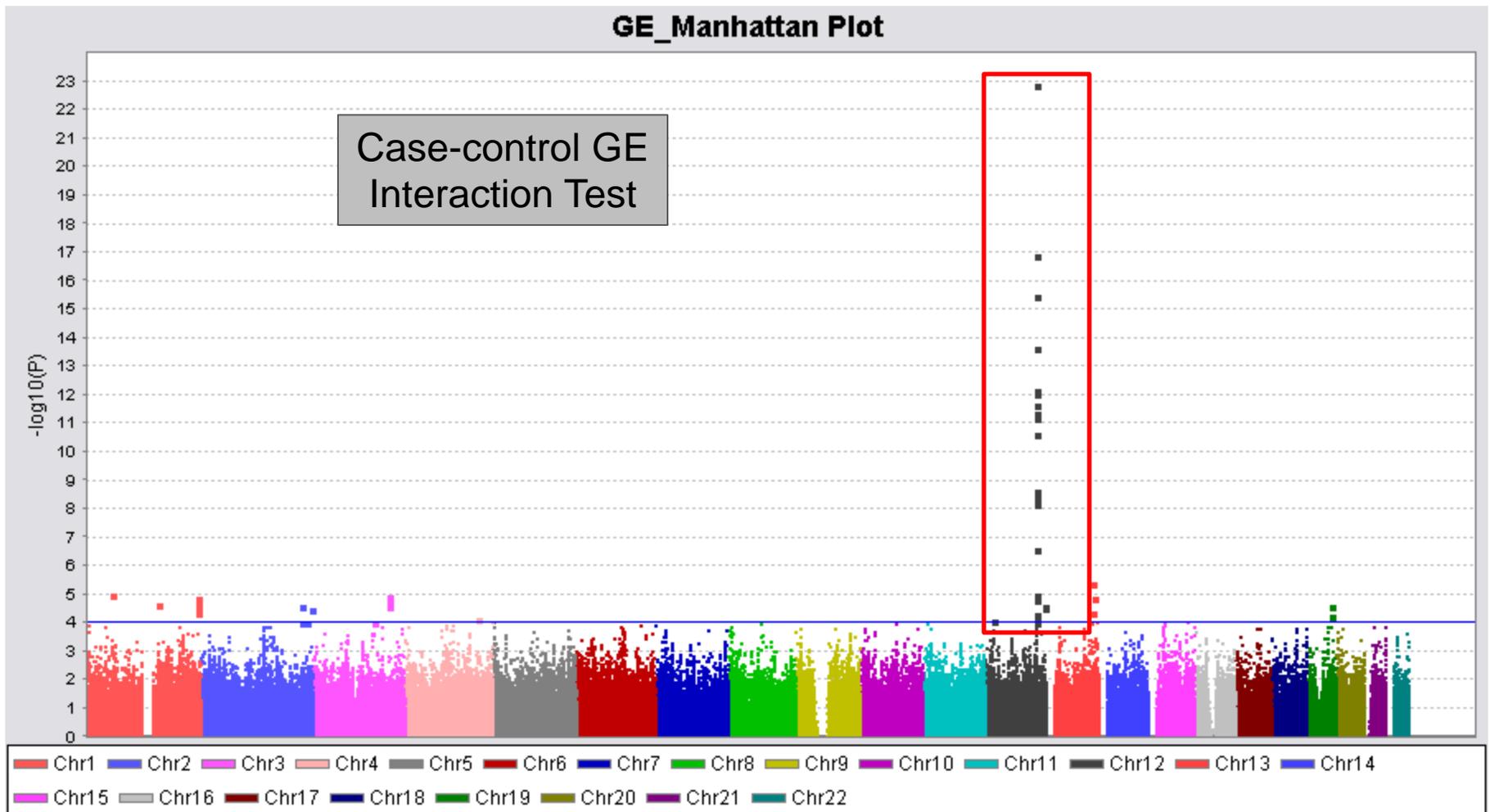


Color code: CC — CO — TS ($\alpha_1=5 \times 10^{-4}$) — TS ($\alpha_1=5 \times 10^{-2}$) —
 EB — EB2 — AIC — BMA —

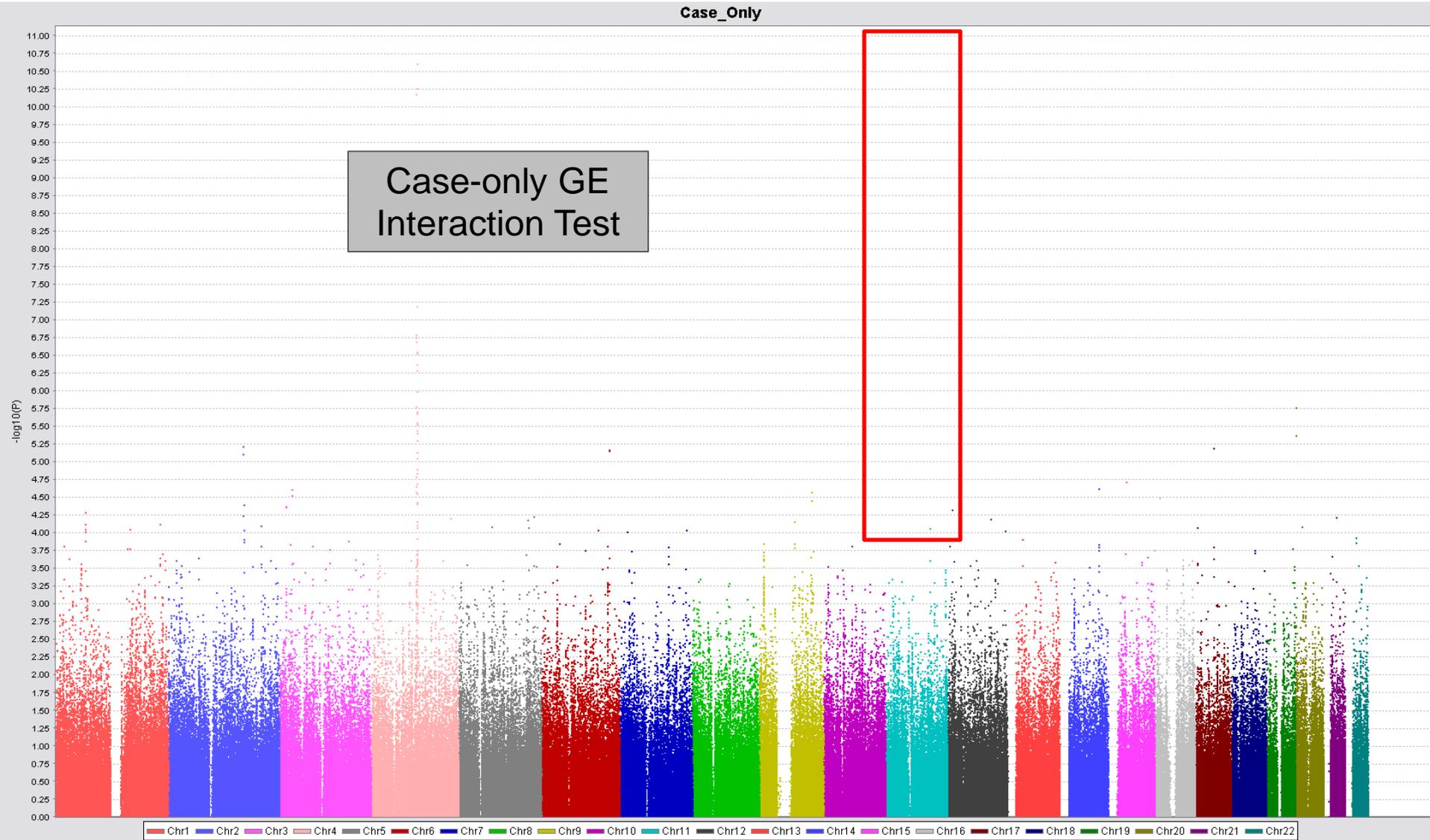
Example: ESCC, *ALDH2* and Alcohol Intake



Example: ESCC, *ALDH2* and Alcohol Intake



Example: ESCC, *ALDH2* and Alcohol Intake



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

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

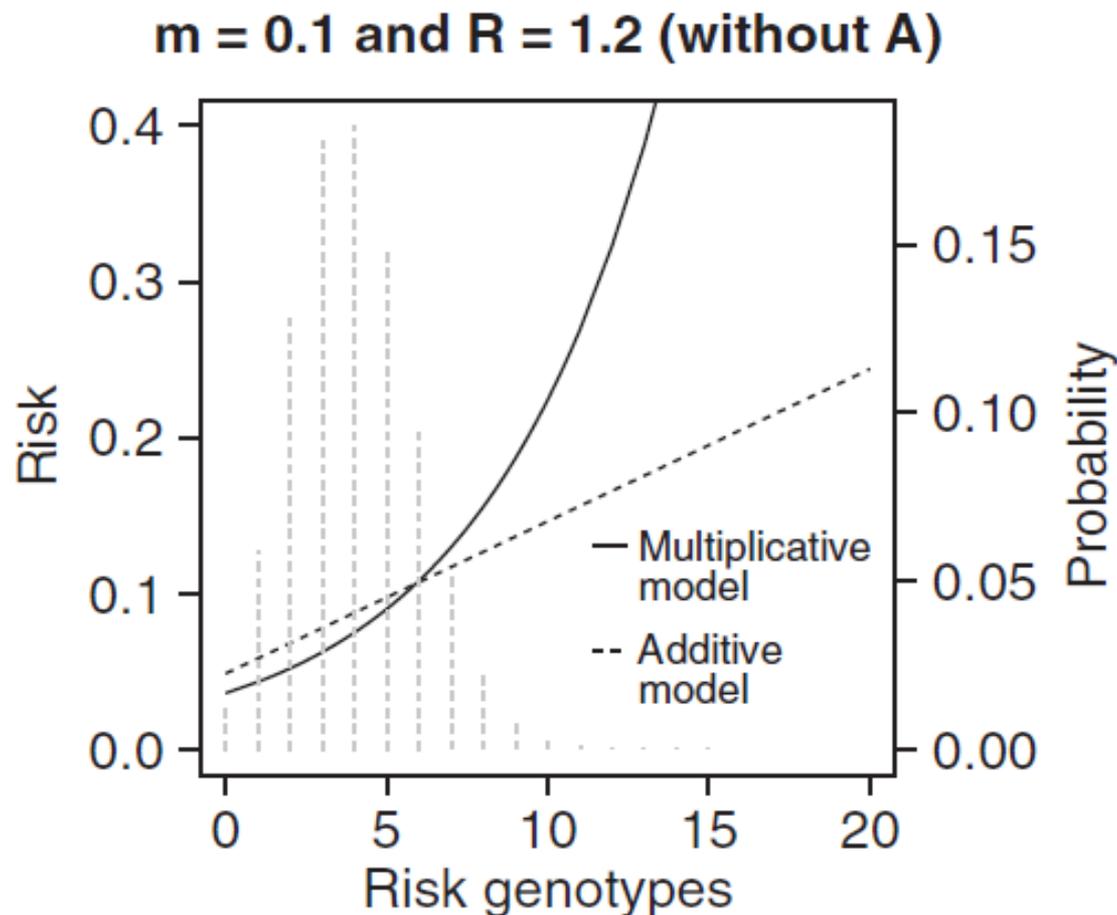
	Genome-wide Significant?	
	<i>ALDH2</i> ($\alpha=5\times 10^{-8}$) rs11066015 ^a	<i>ADH</i> rs3805322 ^b
Standard case-control test	Yes	no
Case-only test	No	Yes
Empirical Bayes test	Yes	no
Hybrid two-step approach	Yes	no
Cocktail 1	Yes	Yes
Cocktail 2	Yes	Yes

^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 $p_A=6.0\times 10^{-14}<\alpha_A$ and $p_M=7.3\times 10^{-8}<\alpha_M$, and the standard test of $G\times E$ 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 ($p<10^{-16}$).

^b 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 $G\times E$ interaction at the second stage did not meet the second stage threshold ($\approx 4.2\times 10^{-4}$); 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 ($\approx 4.2\times 10^{-4}$).

When “no interaction” is the more interesting result!

“No (supra- or sub-) multiplicative interaction”
can still have dramatic consequences.



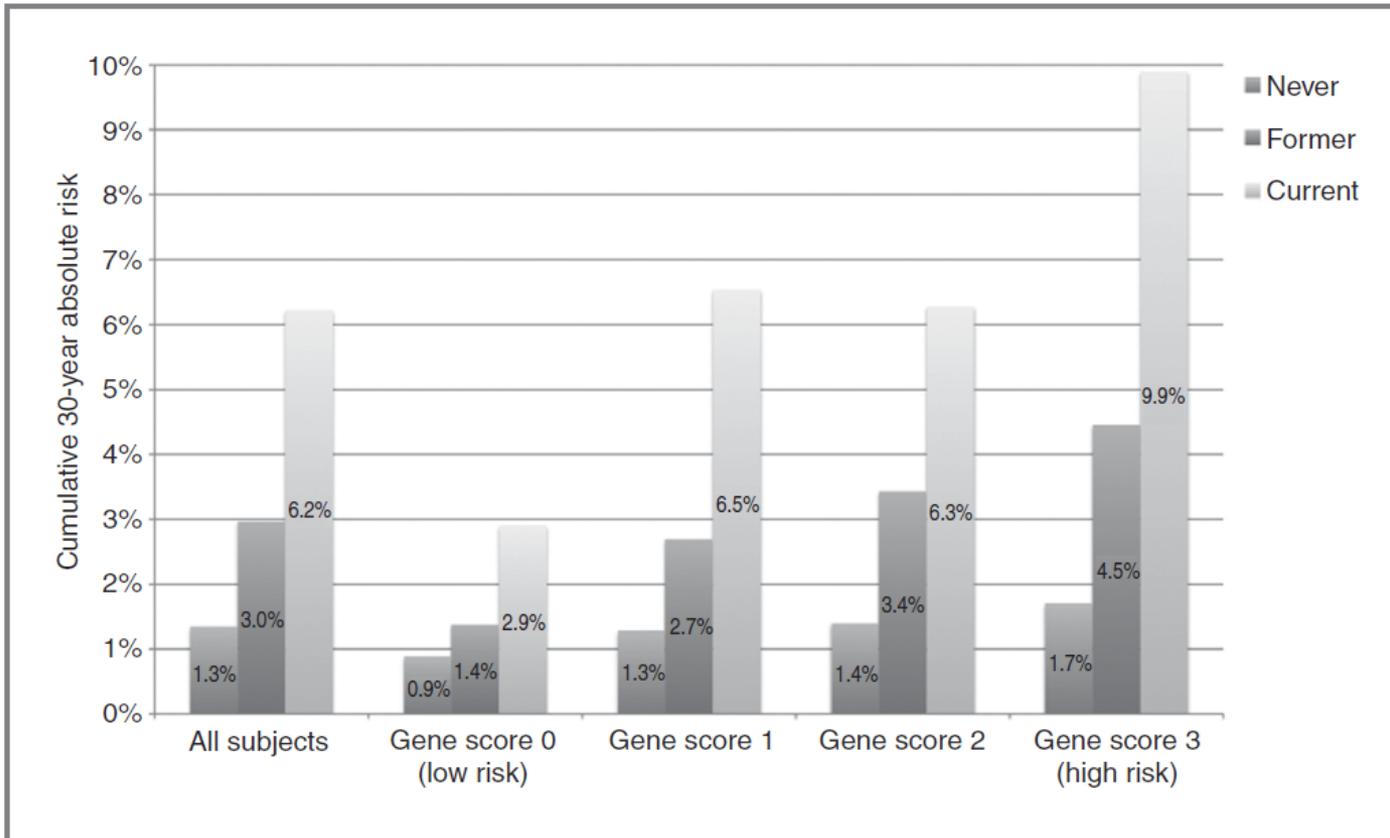
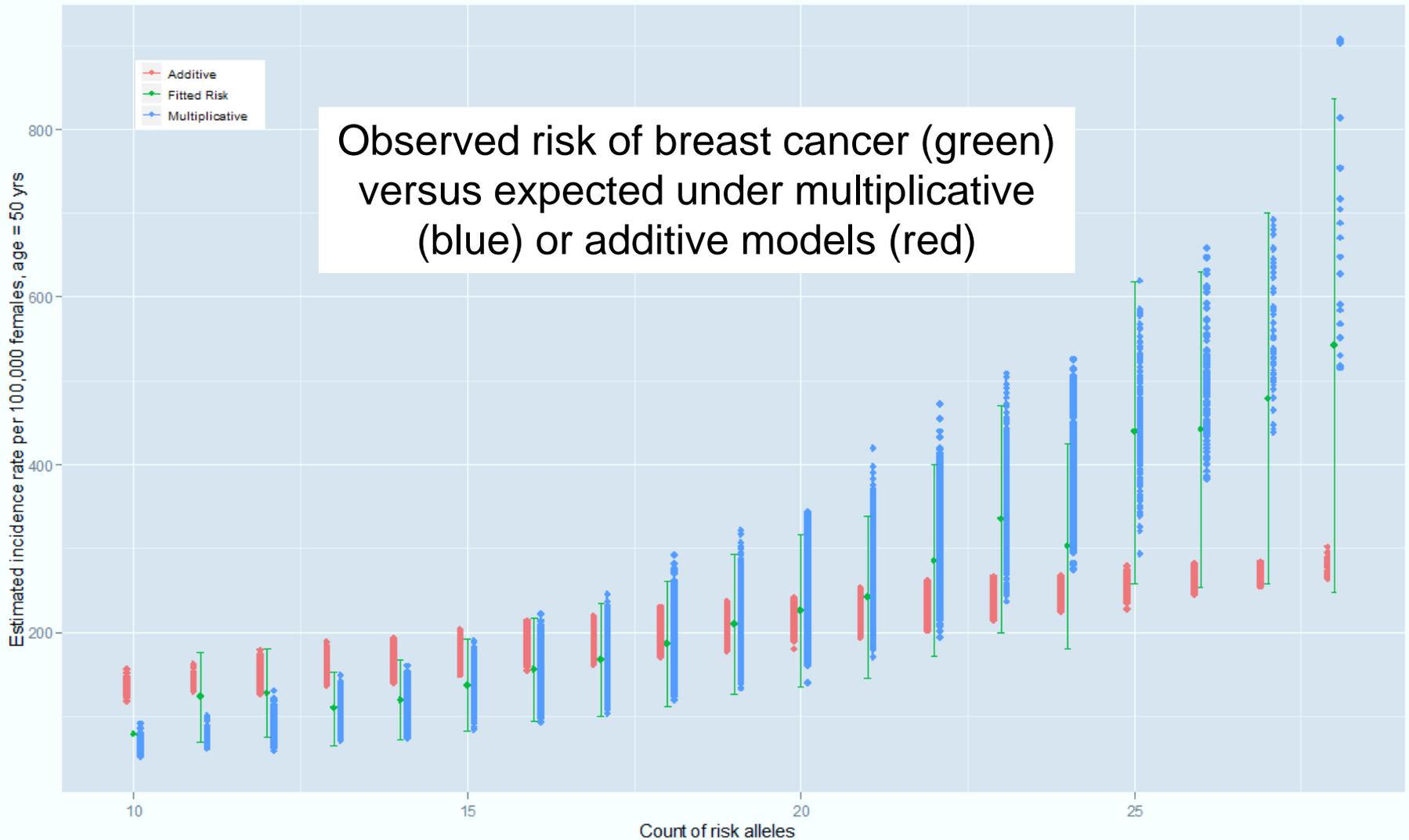


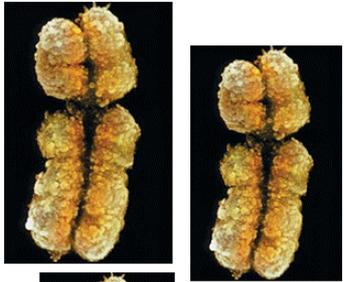
Figure 1. Cumulative 30-year absolute risk for bladder cancer in a 50-year-old male in the United States, overall and by quartiles of a polygenetic genetic score.

Benefit of smoking (=reduction in 30 year cumulative cancer risk) much greater among those in highest quartile of genetic burden versus lowest (8.2 vs 2.0). Clearly interesting, although the test for multiplicative interaction between genetic risk score and smoking was non-significant.



Challenge: the uncertainty of the risk estimates is greatest in the tails, which is where we are most likely to identify individuals who would benefit from genetic information

Limits on etiologic inference



???

$$\mu = a + b_g G + b_e E + b_{ge} GE$$

$$\mu = a + b_g G + b_e E$$

$$\log(\mu) = a + b_g G + b_e E + b_{ge} GE$$

$$\log(\mu) = a + b_g G + b_e E$$

$$\log(\mu/(1-\mu)) = a + b_g G + b_e E + b_{ge} GE$$

$$I(G,E;D) - [I(G;D) + I(E;D)]$$

Without assumptions—often strong and untestable assumptions—inferences for or against particular mechanistic models cannot be made, as multiple, qualitatively different mechanistic models are consistent with the observed pattern of statistical interaction

Sorting out true from false positives, balancing against false negatives

Maintain epistemological modesty. I.e. don't place too much faith in specific priors—never mind unfalsifiable hypotheses or post-hoc explanations

Interaction Between the Serotonin Transporter Gene (*5-HTTLPR*), Stressful Life Events, and Risk of Depression

A Meta-analysis

Neil Risch, PhD

Richard Herrell, PhD

Thomas Lehner, PhD

Kung-Yee Liang, PhD

Lindon Eaves, PhD

Josephine Hoh, PhD

Andrea Griem, BS

Maria Kovacs, PhD

Jurg Ott, PhD

Kathleen Ries Merikangas, PhD

Results In the meta-analysis of published data, the number of stressful life events was significantly associated with depression (OR, 1.41; 95% CI, 1.25-1.57). No association was found between *5-HTTLPR* genotype and depression in any of the individual studies nor in the weighted average (OR, 1.05; 95% CI, 0.98-1.13) and no interaction effect between genotype and stressful life events on depression was observed (OR, 1.01; 95% CI, 0.94-1.10). Comparable results were found in the sex-specific meta-analysis of individual-level data.

Conclusion This meta-analysis yielded no evidence that the serotonin transporter genotype alone or in interaction with stressful life events is associated with an elevated risk of depression in men alone, women alone, or in both sexes combined.

JAMA. 2009;301(23):2462-2471

www.jama.com

Lessons learned from the study of marginal genetic effects

- Candidate genes have typically not been associated with relevant traits (priors are still low)
- “Moving the goalposts” can generate confusion and divert resources from more promising avenues
- Now strong statistical evidence for association and precise replication are required up front
- Priors for particular gene-environment interactions will be even smaller
- The ability and temptation to “move the goalposts” will be higher for gene-environment interactions