Harmonization and analysis approaches for gene-environment interaction studies

Mariaelisa (Misa) Graff

University of North Carolina at Chapel Hill, Dept. Epidemiology

Guillings School of Global Public Health

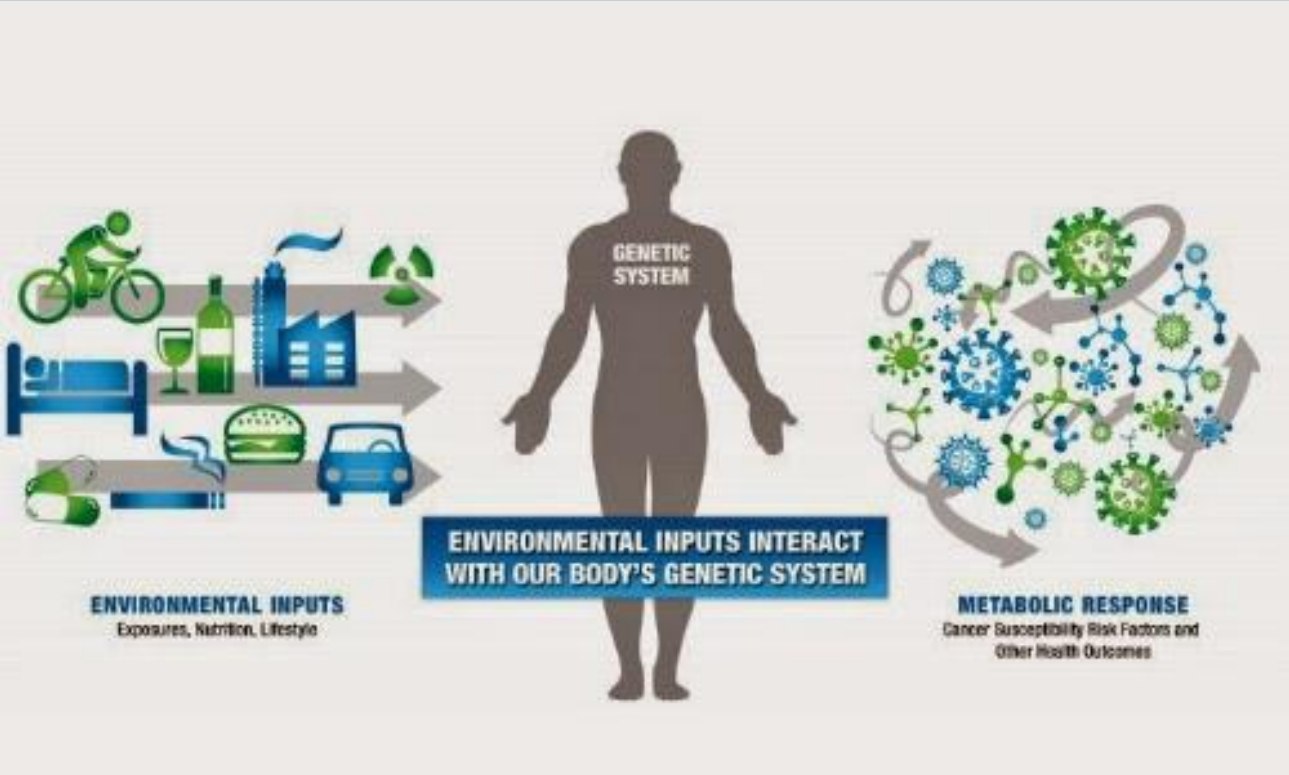
September 28, 2017

Overview of presentation

- What are gene-environment interactions?
- Harmonizing the environmental data among several studies
- Approaches and power for gene-environment interaction analyses

What are gene-environment interactions?

- Many diseases and traits result from a combination of a persons **genetic make-up** and exposure to the **environment.**

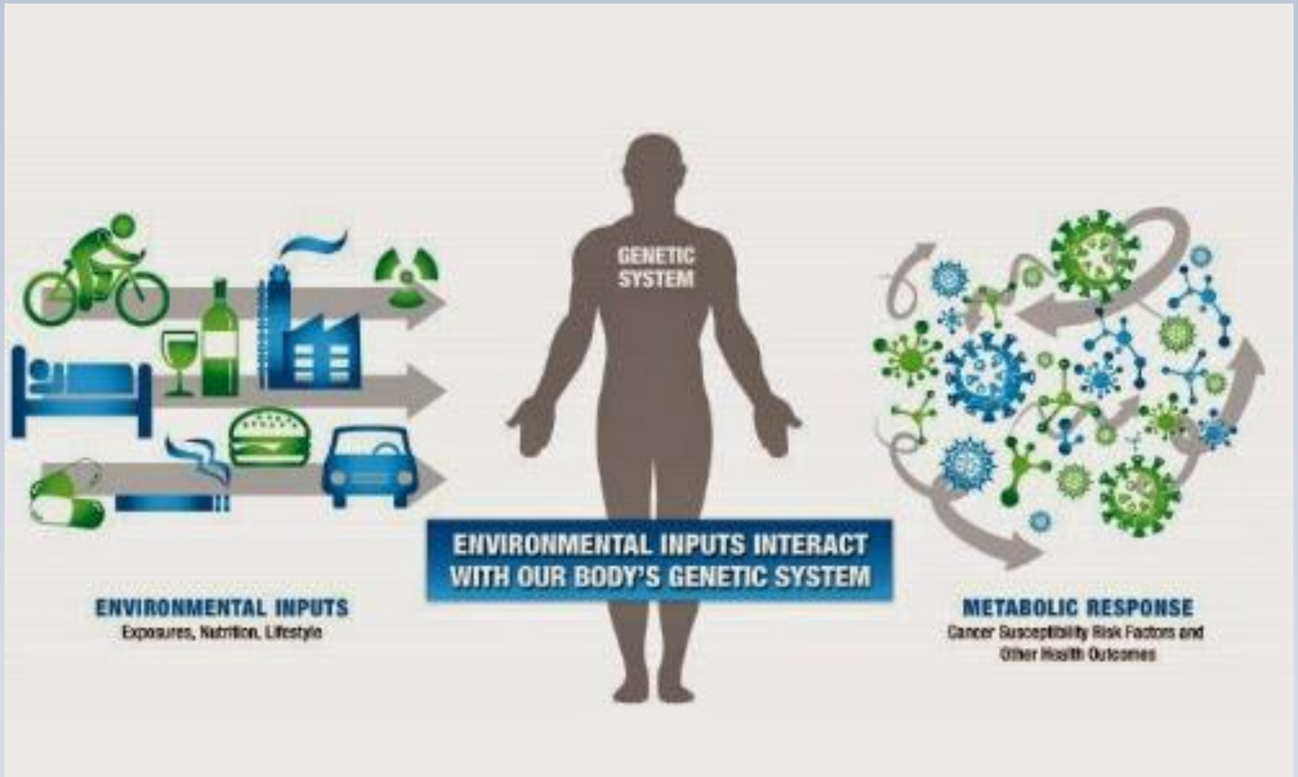

GENETIC SYSTEM
ENVIRONMENTAL INPUTS INTERACT WITH OUR BODY'S GENETIC SYSTEM
ENVIRONMENTAL INPUTS
Exposures, Nutrition, Lifestyle
METABOLIC RESPONSE
Cancer Susceptibility Risk Factors and Other Health Outcomes

- *Sensitivity* to environmental factors for a trait or disease may be inherited rather than the trait or disease itself being inherited.

- Understanding these sensitivities can give insight into different traits and diseases.

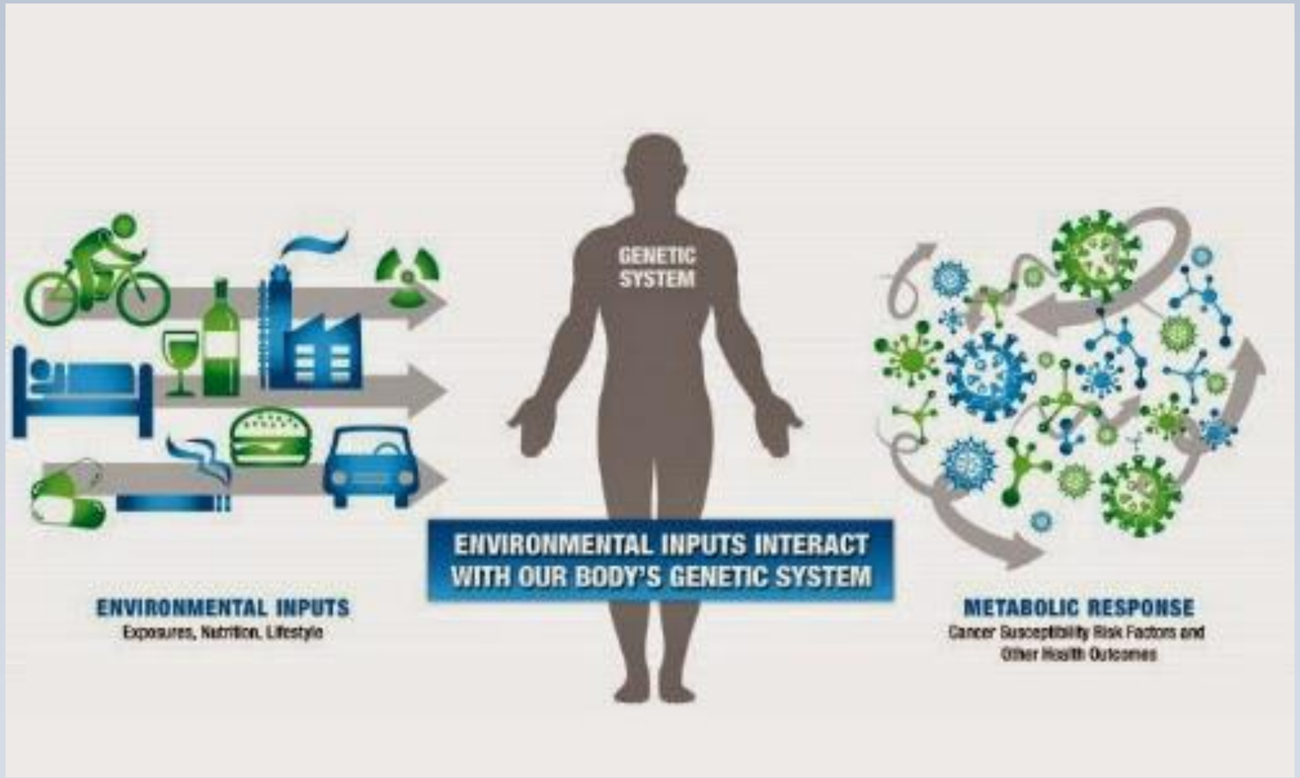
What are gene-environment interactions?

- **Genetic make-up** is commonly measured as a genotype or a single nucleotide polymorphism (SNP).
- It can also include a combination of several SNPs, gene expression, heritability, copy number variants, etc.


GENETIC SYSTEM
ENVIRONMENTAL INPUTS INTERACT WITH OUR BODY'S GENETIC SYSTEM
ENVIRONMENTAL INPUTS
Exposures, Nutrition, Lifestyle
METABOLIC RESPONSE
Cancer Susceptibility Risk Factors and Other Health Outcomes

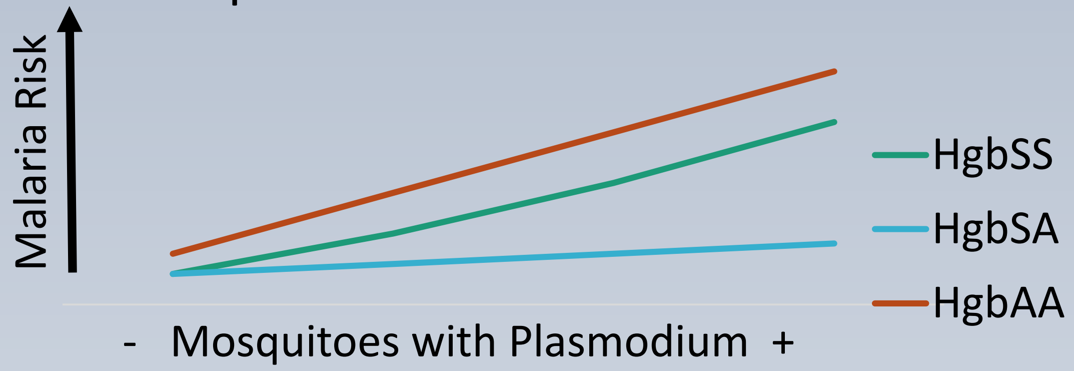
What are gene environment interactions?

- **Environment** refers to any non-genetic component:
  - A persons behaviors: e.g. sleeping, diet
  - Chemicals in the air: e.g. pollution, ozone
  - A treatment or medication
  - Biological trait or metabolite: e.g. BMI, LDL-cholesterol


GENETIC SYSTEM
ENVIRONMENTAL INPUTS INTERACT WITH OUR BODY'S GENETIC SYSTEM
ENVIRONMENTAL INPUTS
Exposures, Nutrition, Lifestyle
METABOLIC RESPONSE
Cancer Susceptibility Risk Factors and Other Health Outcomes

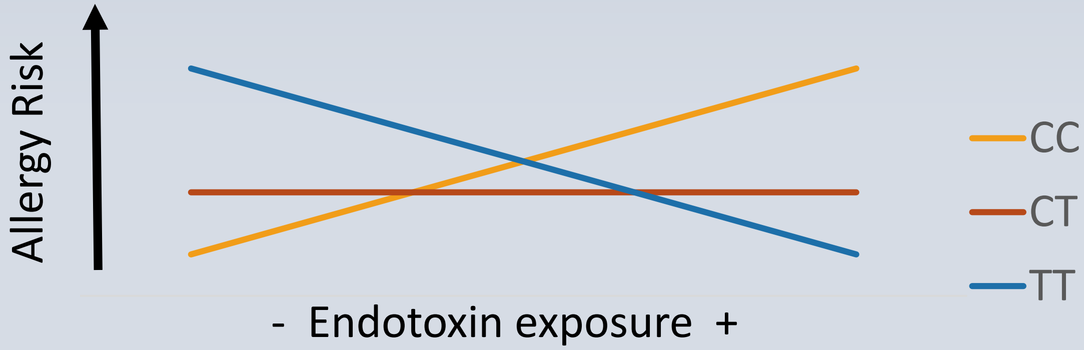
Presence of a gene-environment interaction

Heterozygote advantage of HgbS in presence of Plasmodium



PMID: 27852523; PMID: 19901265

Result of microbial load, CT14 and allergy risk



Eder et al *J Allergy Clin Immunol.* 2005; PMID: 16159630

Why do gene environment interactions occur?

- Individuals with different genotypes are affected differently by exposure to the same ***environmental*** *f*actors.
- Gene-environment interactions can result in different phenotypes.

Overview of presentation

- What are gene-environment interactions?
- Harmonizing the environmental data among several studies
- Approaches to gene-environment interaction analyses

Harmonizing the environmental variable is just as important as harmonizing the outcome variable

- What is the question being asked with respect to the environment?

- What sorts of data do the participating studies have with respect to the environmental component?

Examples of gene-environment interaction questions

- Is there a difference in the association between *amounts and intensities of physical activities* on biomarkers and changes in skeletal muscle gene expressions?
- Does *smoking* exacerbate an association of a genetic risk score of renin-angiotensin system gene polymorphisms and blood pressure?
- How does intake of *whole grain foods* interact with genetic variants to influence insulin and glucose levels?

Defining the environmental component

- What are "*amounts and intensities of physical activities*"?

Defining the environmental component

- What are "*amounts and intensities of physical activities*"?

Burpee Ab Shred

-10 Burpees
-20 Crunches
-10 Burpees
-20 Reverse Crunches
-10 Burpees - legs to the left
-20 Knee-Ins - alternate legs
-10 Burpees - legs to the right
-20 Knee-Ins - alternate legs
-30 second plank

Challenge: REPEAT!

GLASBERGEN
"Sweat is nature's way of showing you that your muscles are crying!"
Fatter wallet
Clearer head
NATIONAL WALK TO WORK

Clean house
Nice yard

Intensity!

Exercising commuters

Does exercising eyes and fingers count?

Every day?

Defining the environmental component

- What does *smoking* include?

Defining the environmental component

- What does *smoking* include?

Current smoking

Ever smoked

Smoked once

Age started smoking

Cigarettes per day

Age quit smoking

Defining the environmental component

- What are *whole grain foods?* How are they measured?

Defining the environmental component

- What are *whole grain foods?* How are they measured?

With Whole Grain First Ingredient
Lucky Charms
FROSTED TOASTED OAT CEREAL WITH MARSHMALLOWS
they're Magically Delicious!
More WHOLE GRAIN than any other ingredient!
pop-tarts
Whole Grain
Strawberry
Portion Distortion
What you're served
What's one serving
Fruits
Grains
Vegetables
Protein
My Plate - Grains Grou
Make 1/2 your grains 100% whole g

Harmonizing the environmental variable

- *Possible issues:*
  - Biologically invalid values?
  - Inconsistencies in the study data?
  - Missing data?
- *What to do:*
  - Which measurements are correct?
  - Should discrepant data values be excluded?
  - Look to understand as much about the variable and how it was measured as possible.
  - Are there algorithms or conversions that should be applied?

Example of harmonizing physical activity (PA) in a SNPxPA genome-wide meta-analysis of adiposity traits

- Participating studies used various different ways of measuring and quantifying environmental exposures
- Gene x environment interactions generally have small effects. Need large sample sizes to maximize power.

→ How can we harmonize heterogeneous PA data to maximize power for detecting GxPA interactions in 60 cohorts?

M Graff, RA Scott, AE Justice, KL Young, et al. PLoS Genet. 2017 Apr 27;13(4):e1006528. PMID: 28448500

Heterogeneity of PA data

I) Types of PA

- Leisure-time PA
    - Recreational
    - Domestic
- Occupational PA
- Commuting PA

II) PA measurements

- Objective measurement (e.g. accelerometer based)
- Subjective measurement (questionnaires)
    - Categorical (e.g. ‘Do you spend most working hours sitting?’)
    - Continuous (questions on PA duration/frequency)

Options for Harmonizing PA

Harmonizing PA across all cohorts

- From the onset it seemed that to maximize sample size, only crude harmonization by dichotomizing PA would be feasible

Harmonizing PA in subsets of studies

- Used a subset of studies to test the best way to dichotomize PA
- Meta-analyzed studies that use the same PA measure (most commonly moderate-to-vigorous LTPA h/wk)
- Meta-analyzed cohorts with objective PA data

Dichotomous PA variable
→Which PA cut-off to choose?

- Results from harmonizing PA in subsets of studies
  - 1) *FTO*xPA interaction seen when comparing sedentary vs. other individuals
  - 2) Benefits of increasing PA greatest in sedentary individuals
  - 3) Sedentary individuals easy to identify in most cohorts

→ Dichotomized by *sedentary individuals* vs. *others*

Definition of sedentariness:

- sitting at work AND
- <1 h/wk of moderate-to-vigorous leisure-time/commuting PA

Choosing homogeneous PA cut-off

1) Studies with categorical PA measure

- Limited options → Choosing the most appropriate cut-off
- Problem: Categories may not correspond well across studies

2) Studies with continuous PA measure: 2 options

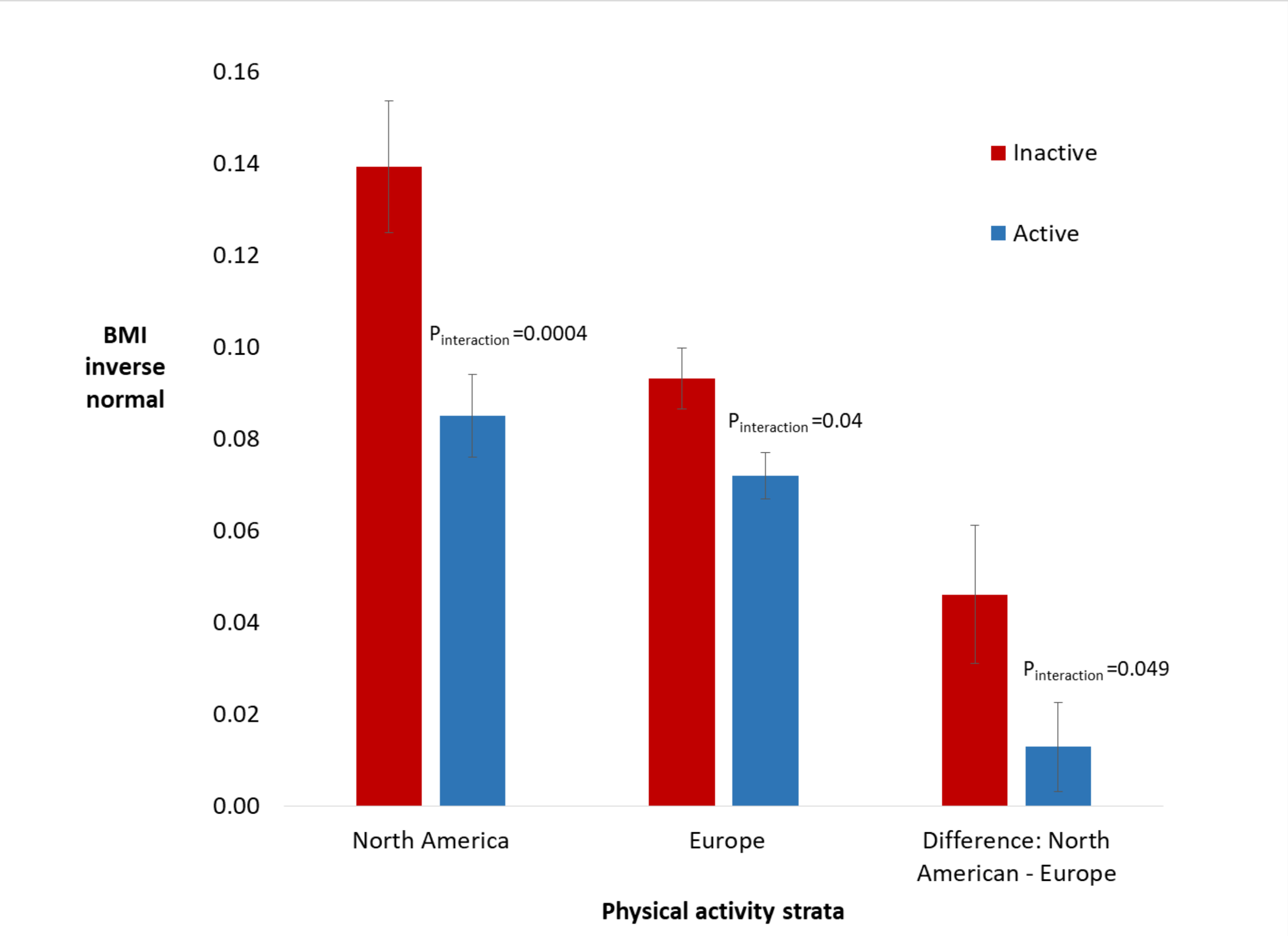
- A) *Absolute PA cut-off*
  - (e.g. sedentary = individuals with <300 MET-min/wk of moderate-to-vigorous LTPA)
  - Problem: Coverage of PA behaviors differs between questionnaires → Absolute values not comparable
- B) *Relative PA cut-off*
  - (e.g. sedentary = individuals in the lowest quintile of PA distribution)
  - Problem: Does not account for differences in PA levels between populations

Summary: Harmonization in SNPxPA genome-wide meta-analysis

- Meta-analyzed all cohorts with genetic data and PA
- Used all available PA data (occupational, leisure-time, commuting)
- Dichotomous PA variable (sedentary vs. others)
- Choice of cut-off within individual cohorts:
  - Studies with categorical PA measure: chose the most appropriate category for sedentary behavior
  - Studies with continuous PA measure: sedentary = lowest sex-specific quintile of PA distribution

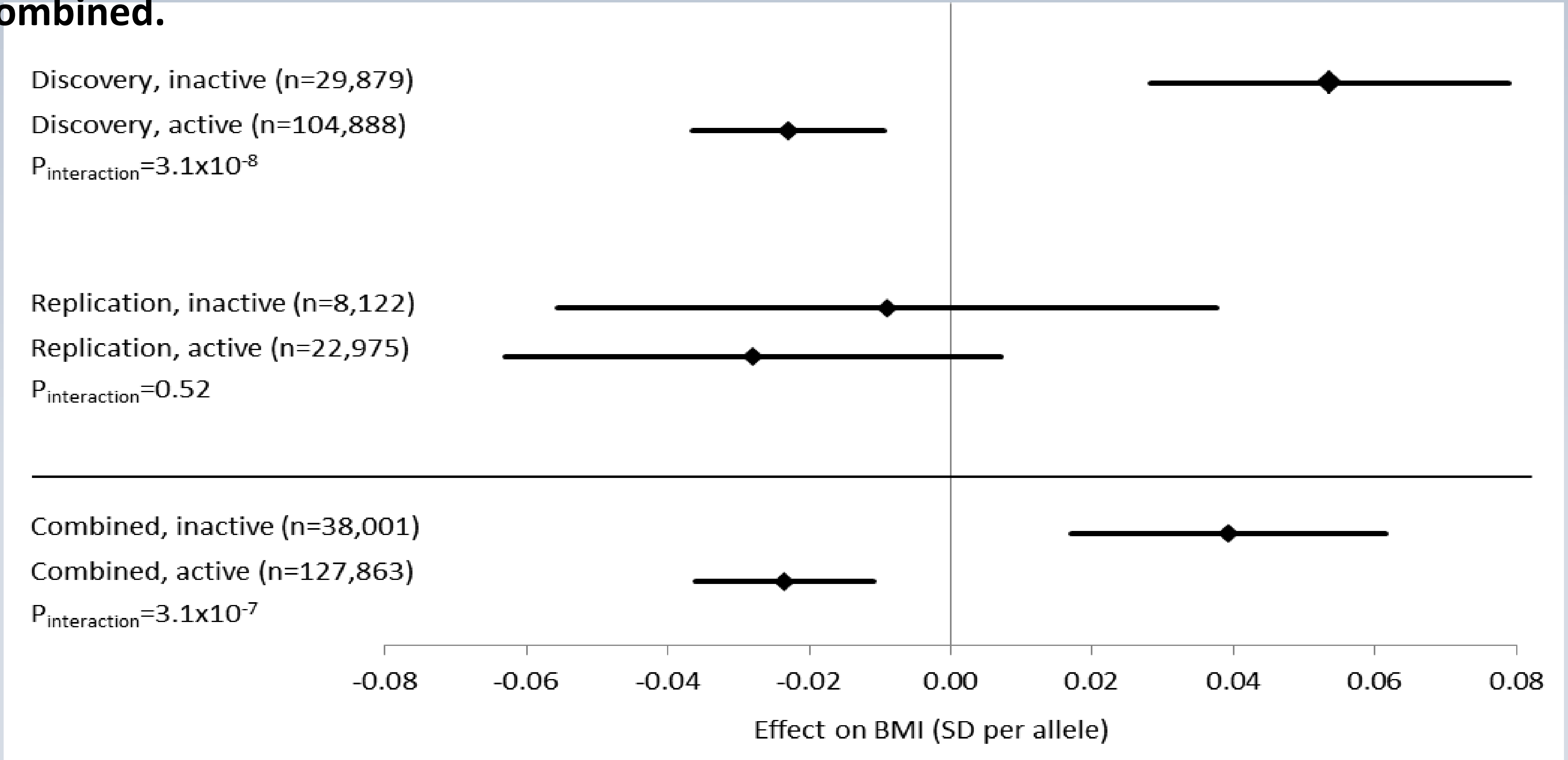
Summary: Harmonization in SNPxPA genome-wide meta-analysis of adiposity traits

- Findings: FTO x PA interaction with BMI
- No new interactions

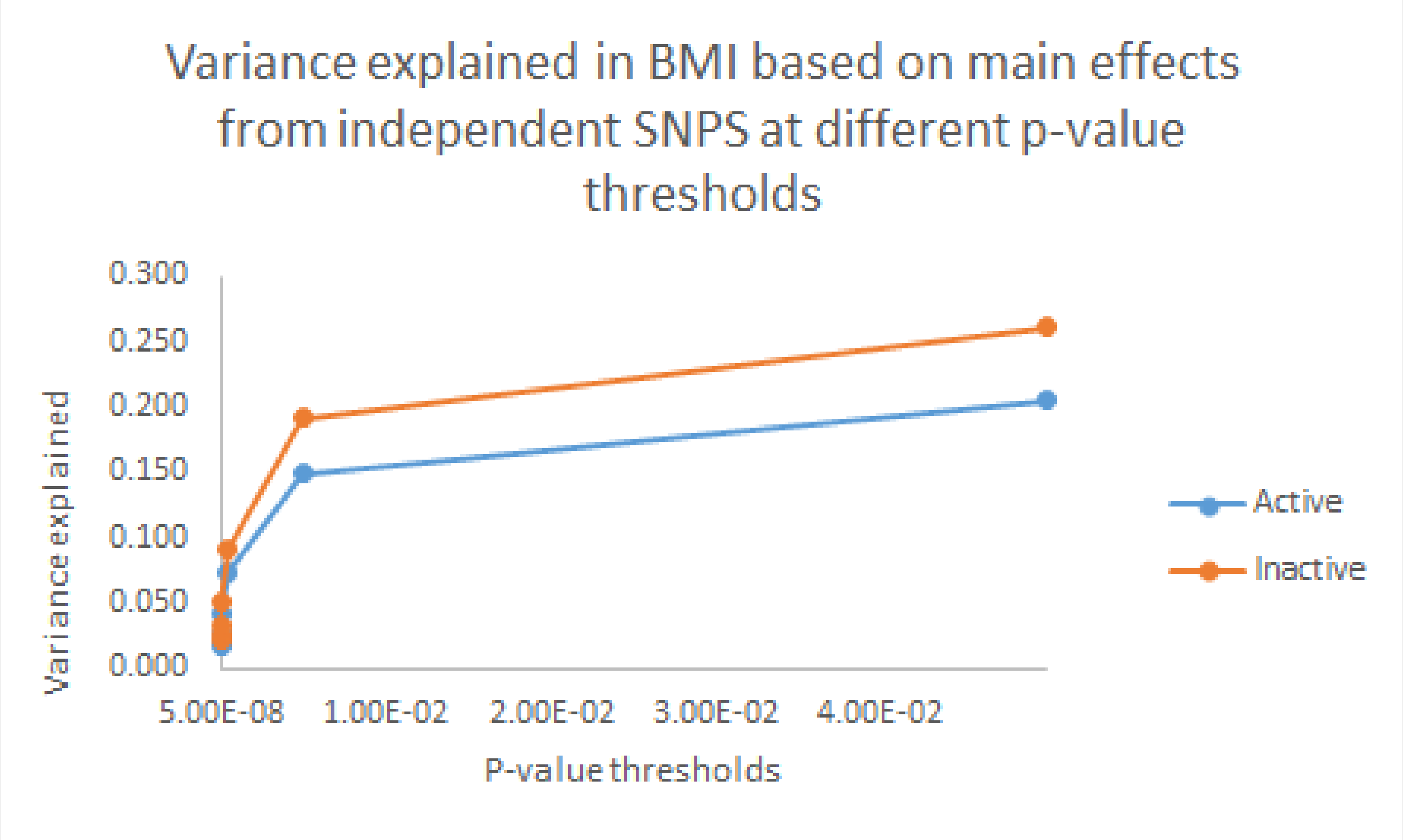


Summary: Harmonization in SNPxPA genome-wide meta-analysis

Interaction between the *CDH12* locus and physical activity on BMI in the discovery genome-wide meta-analysis (n=134,767), in the independent replication sample (n=31,097), and in the discovery and replication samples combined.



Summary: Harmonization in SNPxPA genome-wide meta-analysis

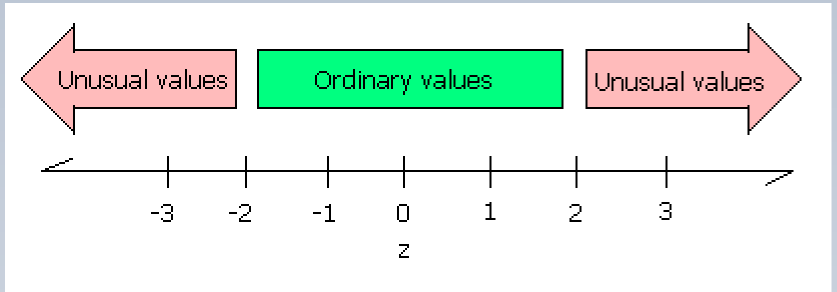


Harmonizing the environmental variable

- When meta-analyzing with several studies, its important to understand how the data is defined and measured.
- Poorly defined or measured variables can lead to increased error.
- Poorly harmonized variables can lead to increased error.
- There are tools that can help with harmonization

(e.g. PhenX).

Other ways to combine variables across different studies



- Possible options:
  - Inverse normalize or transform variable of interest to Z-scores.
  - Meta-analyze summary results using sample size (or weights) and p-values across several studies.
- Benefit - variables do not have to be the same.
- Drawback - may not be able to calculate a meaningful effect estimate.

Overview of presentation

- What are gene-environment interactions?
- Harmonizing the environmental data among several studies
- Approaches and power for gene-environment interaction analyses

Approaches to gene-environment interaction analyses
Statistical Framework : Joint interaction model

Approach 1) Single regression model that includes both the genetic (SNP), Environment (E), and Genetic (SNP) x Environment (E) interaction effects.

- All exposed and unexposed together with an interaction term:

  $Y = \beta_0 + \beta_E E + \beta_G SNP + \beta_{GE} E * \mathrm{SNP} + \beta_C C + e$

*Outcome = intercept + E + SNP + SNP * E + covariates*

Approaches to gene-environment interaction analyses
Statistical Framework : Joint interaction model

- **Approach 1)** $Y = \beta_0 + \beta_E E + \beta_G SNP + \underline{\beta_{GE} E * \mathrm{SNP}} + \beta_C C + e$

  - **Question 1:** Is there a significant interaction effect ($\beta_{GE} E * \mathrm{SNP}$)?
    - Test this using the Wald test statistic. It follows a chi-squared distribution with 1 DF under H_0: $\beta_{GE} = 0$.
    - *Most powerful in a cross-over interaction, when the association of the SNP and outcome flips in divergent environments*.

Approaches to gene-environment interaction analyses
Statistical Framework : Stratified model

Approach 2) If environment is dichotomous, we can use a 'stratified' framework that carries out the genetic main-effect analyses separately within the exposed and unexposed groups.

- *Exposed(E1):* $Y = \beta_0^{(1)} + \beta_G^{(1)}\mathrm{SNP} + \beta_c^{(1)}\mathrm{C} + e$

  (E1) Outcome = intercept + SNP$_{E1}$ + covariates

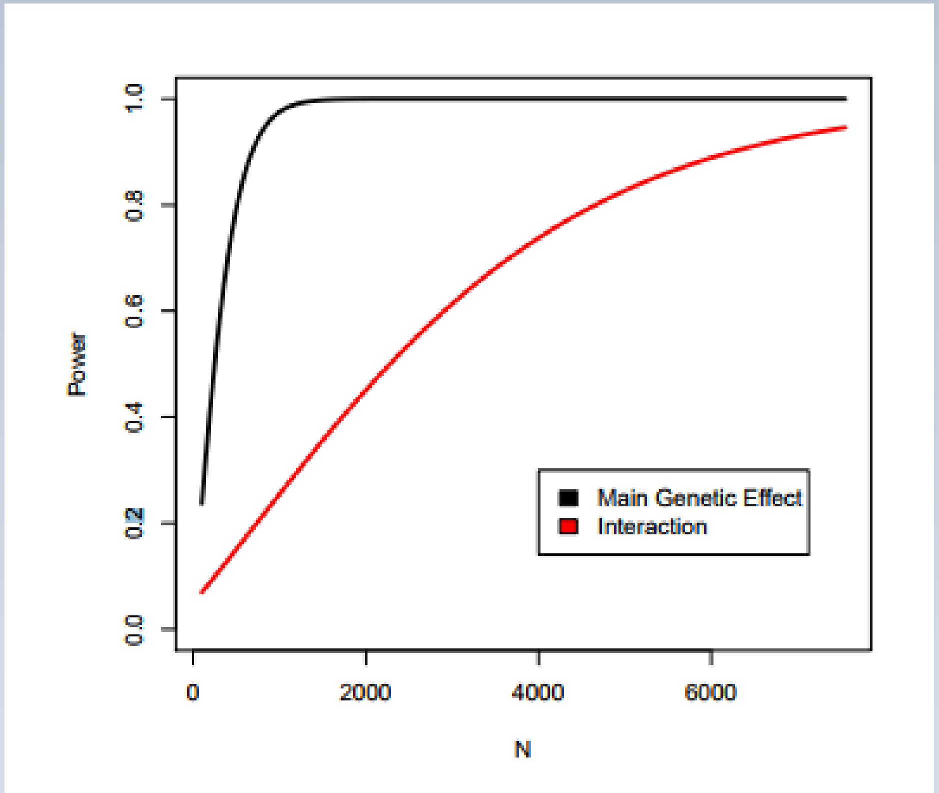
- *Unexposed(E0):* $Y = \beta_0^{(0)} + \beta_G^{(0)}\mathrm{SNP} + \beta_c^{(0)}\mathrm{C} + e$

  (E0) Outcome = intercept + SNP$_{E0}$ + covariates

Approaches to gene-environment interaction analyses
Statistical Framework : Stratified model

- **Approach 2)**
  - **Question 1**: Is there a significant difference in SNP effect between the 2 exposure groups ($\beta_G^{(1)}$SNP − $\beta_G^{(0)}$SNP)?
  - Calculate a z-statistic: $Z_{diff} = \frac{\beta_G^{(1)}\text{SNP} - \beta_G^{(0)}\text{SNP}}{\sqrt{SE(^{(1)}_{G})^2 + SE(^{(0)}_{G})^2 - 2rSE(^{(1)}_{G})SE(^{(0)}_{G})}}$
    - Follows a standard normal distribution with 1DF under H_0: $\beta_{GE} = 0$
    - r= Spearman rank correlation, $\beta_G^{(1)}$SNP and $\beta_G^{(0)}$SNP

Power is low in GxE meta-analyses: requires large sample sizes



- Power as function of sample size: *<u>α = 0.05 level</u>*, disease pop. risk of 0.01%, SNP with MAF of 0.25, environment with prevalence of 20%, both main SNP and interaction effect are 1.25 (OR).

Power in GxE meta-analyses: alternate strategies

- For gene discovery, leverage the interaction by combining with the main effect; 2DF test.
- Case-only analysis
- Combined several SNPs in a risk score (Multi-SNP by E Testing)
- Select only certain SNPs to test - Which SNPs to Test?
  - SNPs with main effects
  - SNPs in candidate genes or pathways (functional groups)
  - Two-stage screening
    - SNPs that meet a suggestive significance (e.g. P<5e-6) in stage 1, the combine with a 2nd stage of results

Approaches to gene-environment interaction analyses
Statistical Framework : Joint interaction model

- **Approach 1)** $Y = \beta_0 + \beta_E E + \underline{\beta_G SNP + \beta_{GE} E * \text{SNP}} + \beta_C C + e$

  - **Question 2:** Do we find significance if we add the main effect with the interaction effect ($\beta_G SNP + \beta_{GE} E * \text{SNP}$)?
    - Wald test statistic, chi-squared distribution with 2 DF under H_0: $\beta_G = \beta_{GE} = 0$
    - *This is powerful in detecting associations with a suggestive main effect that is stronger in a given environment over another*.
    - ***Primarily useful for gene discovery:*** significance does not necessarily inform interaction.

Kraft et al. 2007 Hum Herid 63:111-9. Huang et al. 2011, Genome Med 3:42.

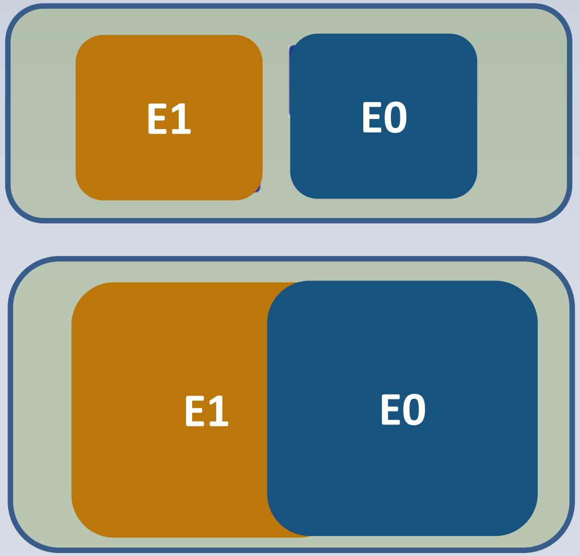
Approaches to gene-environment interaction analyses
Statistical Framework : Stratified model

- **Approach 2)** Question 2:

Do we find significance if we add the main effect with the difference in effect ?

$(\beta_G^{(1)}\mathrm{SNP} + \beta_G^{(0)}\mathrm{SNP}) + (\beta_G^{(1)}\mathrm{SNP} - \beta_G^{(0)}\mathrm{SNP})$

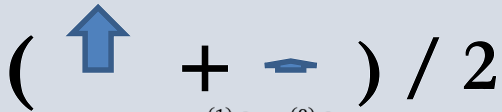
follows a 2 DF chi-squared distribution under H_0: $\beta_G = \beta_{GE} = 0$ when the two strata are independent.



diff

$$t_{diff} = \frac{\beta_G^{(1)}G - \beta_G^{(0)}G}{\sqrt{SE({}_G^{(1)})^2 + SE({}_G^{(0)})^2 - 2rSE({}_G^{(1)})SE({}_G^{(0)})}}$$

main



$$t_{mainG} = \frac{\beta_G^{(1)}G + \beta_G^{(0)}G}{\sqrt{SE({}_G^{(1)})^2 + SE({}_G^{(0)})^2 + 2rSE({}_G^{(1)})SE({}_G^{(0)})}}$$

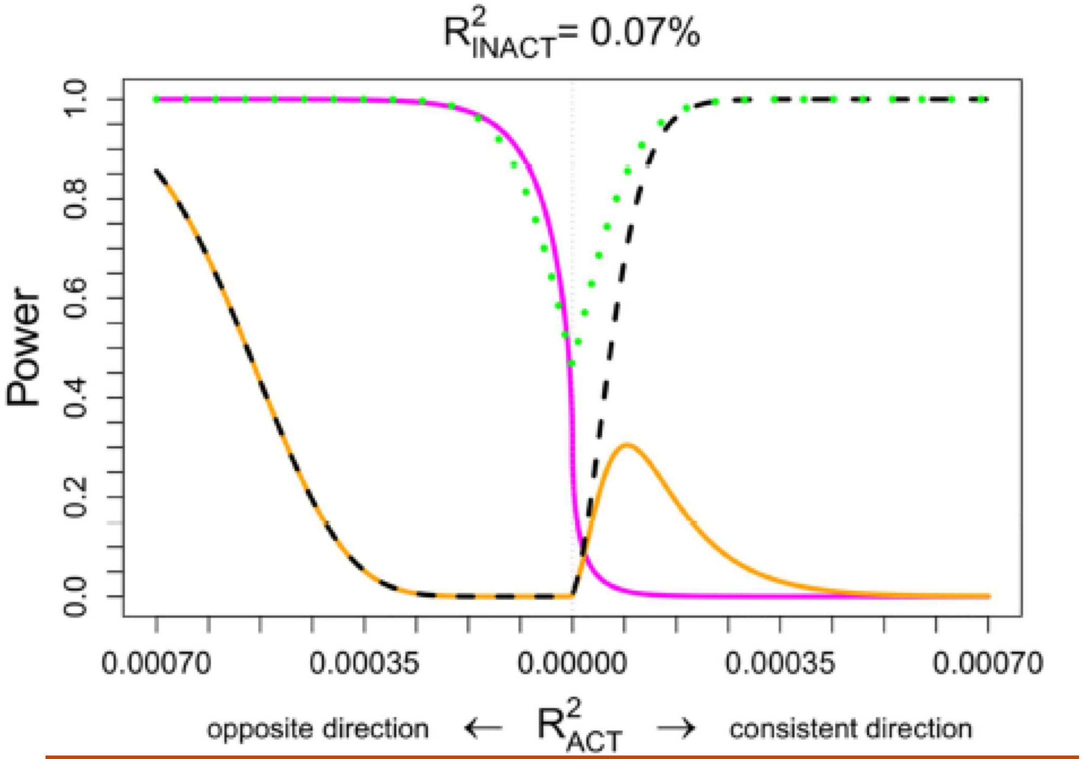
2df

$$X^2_{joint} = \left[\frac{\beta_G^{(1)}G}{SE({}_G^{(1)})}\right]^2 + \left[\frac{\beta_G^{(0)}G}{SE({}_G^{(0)})}\right]^2$$

Aschard H, et al. Hum Hered. 2010; 70(4):292–30, PMID: 21293137.

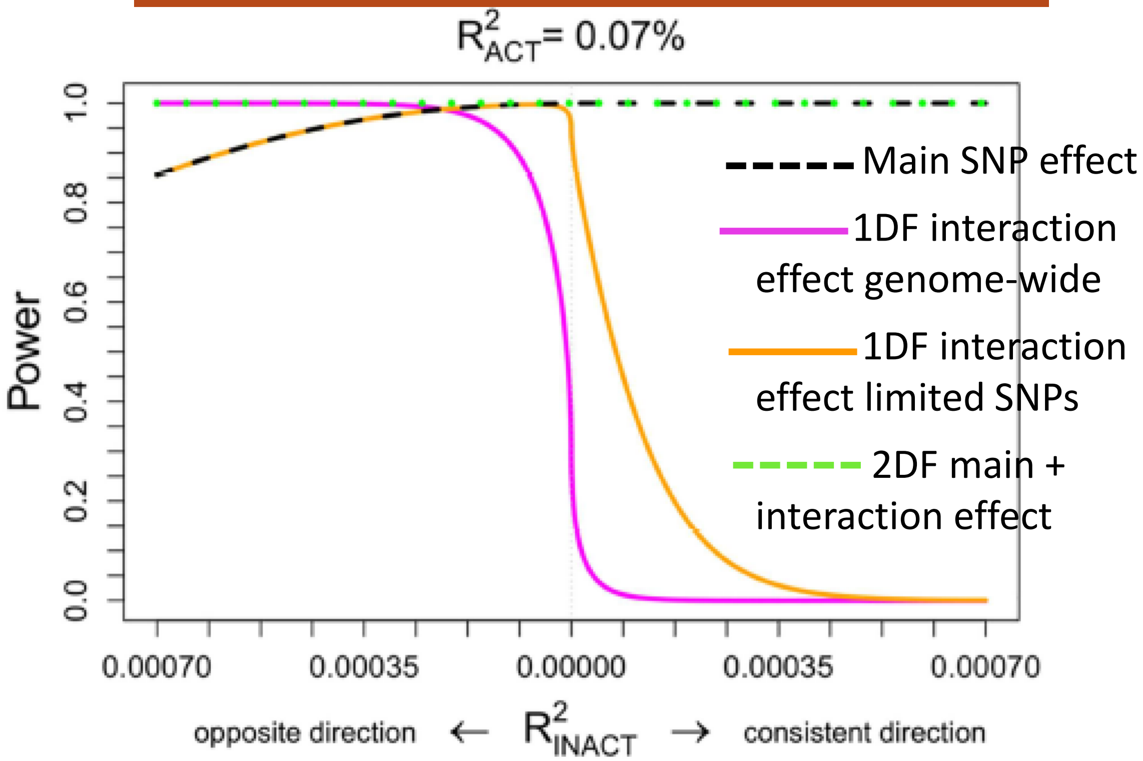
Approaches to gene-environment interaction analyses
Power for 1DF or 2DF tests

Inactive group, N=45,000
Fixed medium effect size



Effect size varies from – to +
Active group, N=155,000

Active group, N=155,000
Fixed medium effect size



Effect size varies from – to +
Inactive group, N=45,000

Approaches to gene-environment interaction analyses
Statistical Frameworks

- **Approach 1)** Joint interaction model
  - Traditional approach
  - Allows for use of continuous environment variable
  - Only need to run the model 1 time
- **Approach 2)** Stratified model
  - Maybe simpler to run depending on the software being used
  - Allows for comparisons of summary statistics
  - Can assess the genetic effects in one group separately

Comparison between 2 Statistical Frameworks:

- ***An Empirical Comparison of Joint and Stratified Frameworks for Studying G × E Interactions: Systolic Blood Pressure and Smoking in the CHARGE Gene-Lifestyle Interactions Working Group.*** Sung et al. *Genet Epidemiol.* 2016 PMID: 27230302

Comparison between 2 Statistical Frameworks:

- Outcome: systolic blood pressure
- 2 environmental exposures:
  - Current versus no smoking
  - Ever versus no smoking
- Data from summary association results:
  - 20 cohorts, European ancestry
  - Family-based and population-based cohorts
  - Cohort sample sizes range from N= 456 to N= 22,983
  - Cohorts analyzed data both ways: using the joint interaction model and stratified model
- Filtering variants:
  - A) **In joint model** remove all variants with minor allele count (MAC)<10 in smokers or non-smokers; **in stratified model** removed variants with MAC<10 based on stratum only
  - B) **in both models**, removed all variants with MAC<10 in smokers or non-smokers

Sung et al. *Genet Epidemiol.* 2016 PMID: 27230302

Comparison between 2 Statistical Frameworks: cohort results

Family-based cohorts

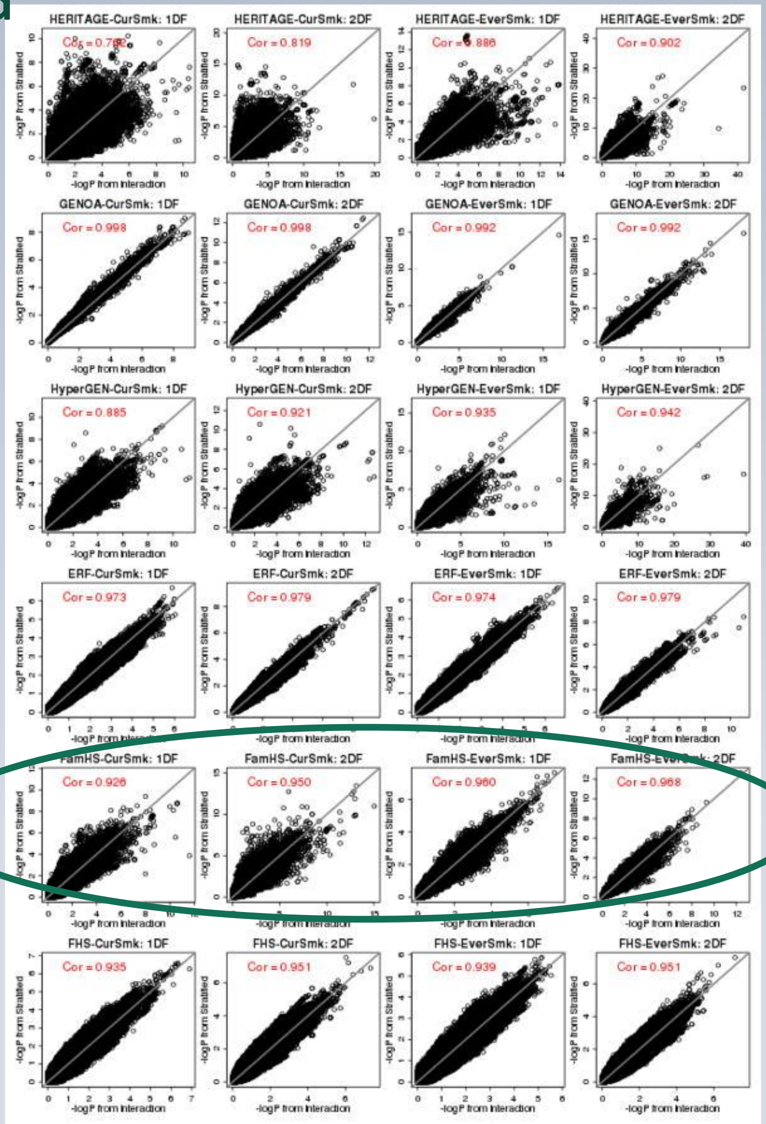
Current vs no

1DF 2DF

Ever vs no

1DF 2DF

Stratified model



N=500

N=1000

N=1200

N=2500

N=3500

N=8000

Current smk
N=500
Ever Smk
N=1600

Joint interaction model

Population-based cohorts

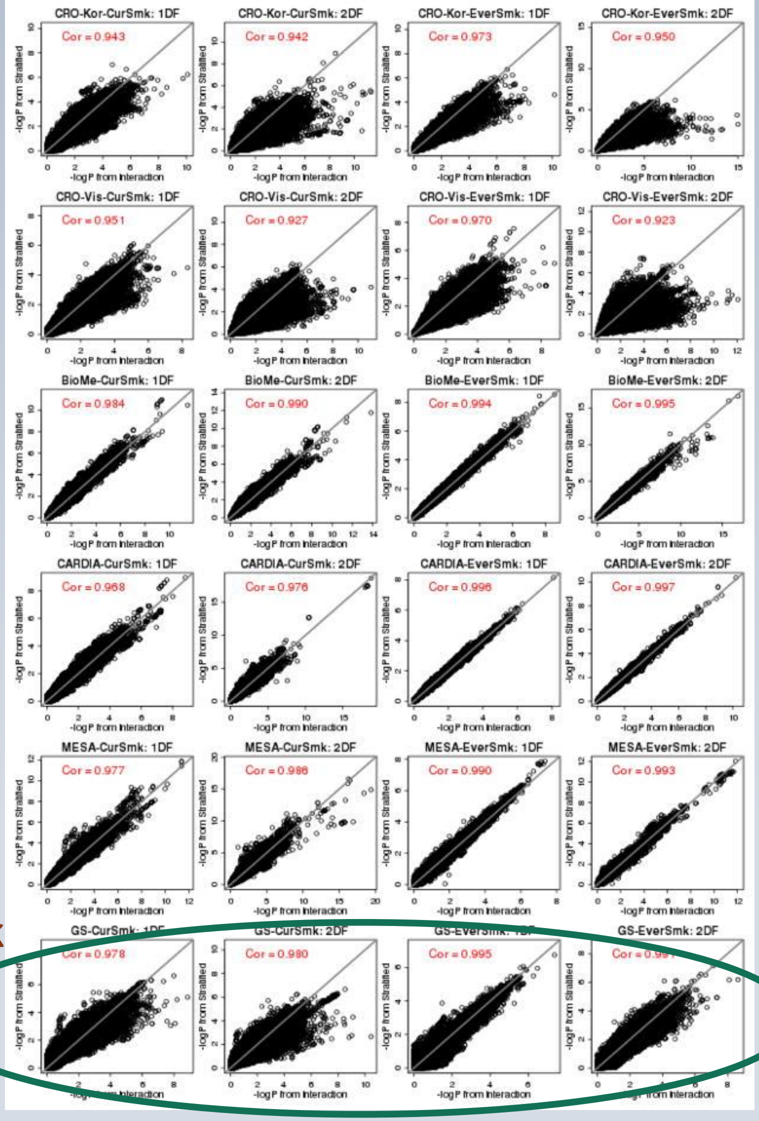
Current vs no

1DF 2DF

Ever vs no

1DF 2DF

Stratified model



N=500

N=500

N=1500

N=1600

N=2600

N=6500

Current smk
N=1000
Ever Smk
N=3100

Joint interaction model

Comparison between 2 Statistical Frameworks: Meta-analysis results

Filtering A

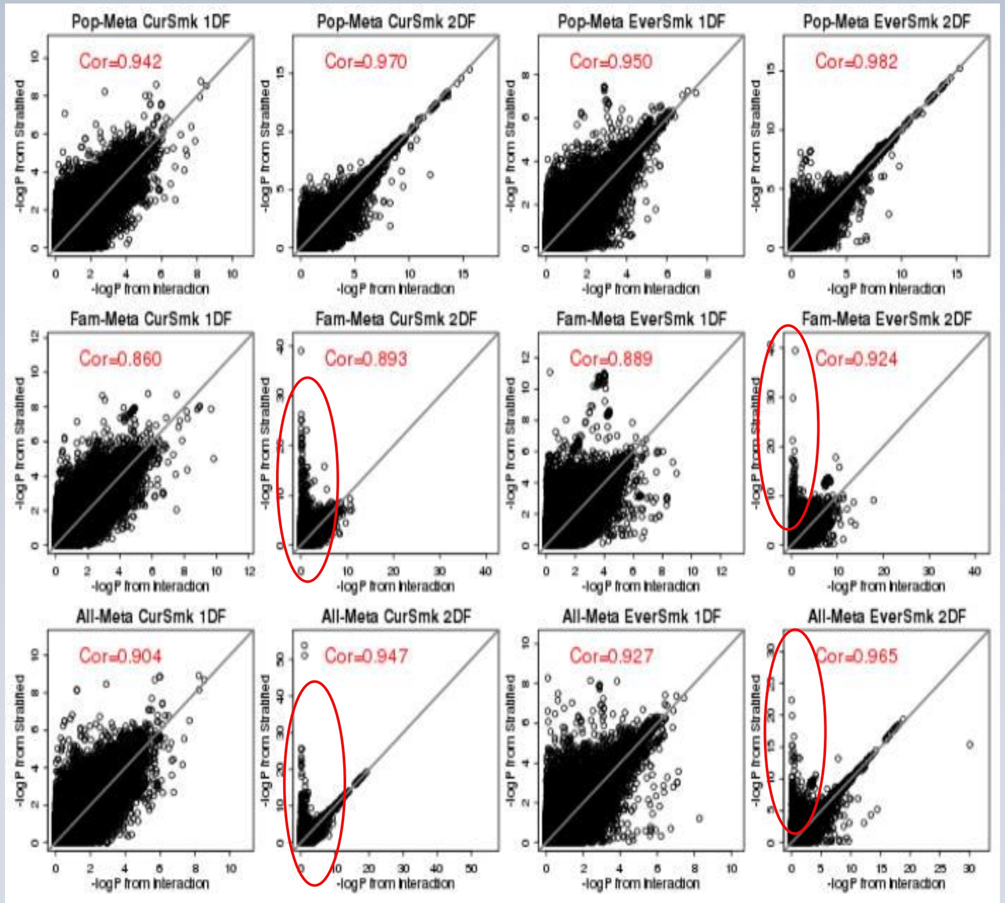
Current vs no: 1DF, 2DF | Ever vs no: 1DF, 2DF

Population-based cohorts

Family-based cohorts

All cohorts

Stratified model

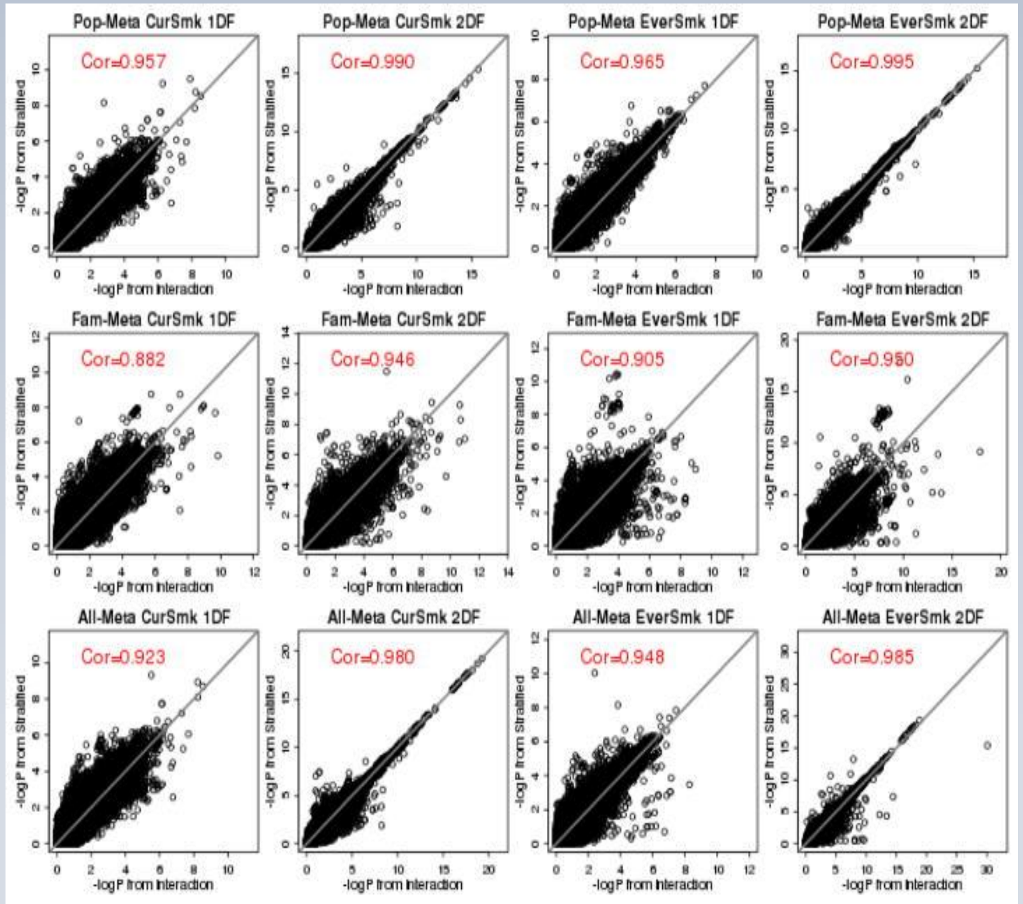


Joint interaction model

Filtering B

Current vs no: 1DF, 2DF | Ever vs no: 1DF, 2DF

Stratified model



Joint interaction model

Sung et al *Genet Epidemiol.* 2016 PMID: 27230302

Comparison between 2 Statistical Frameworks:

- In cohort-specific analyses, good agreement depended on
  - 1) balance between sample sizes of the two strata,
  - 2) total sample size.
- In meta-analyses, agreement depended on
  - 1) the minor allele frequency,
  - 2) inclusion of family-based cohorts in meta-analysis,
  - 3) filtering scheme.
- Stratified framework is more appropriate for population-based cohorts.
- For family-based cohorts, there is less agreement between the two frameworks.
- The stratified framework is unable to fully account for family structures across strata.
  - Spearman rank correlation coefficient in the 1 DF test may partly correct for any correlation between the strata. In contrast, the 2 DF test does not take into account any relatedness across the strata.

Sung et al *Genet Epidemiol.* 2016 Jul; 40(5): 404–415. PMID: 27230302

Summary

- Gene-environment interactions play an important role in the pathobiology of traits and disease.
- Harmonizing the environment variable(s) is essential when working with lots of different kinds of data and /or studies.
- There are different statistical models to use to detect gene-environment interactions.
- Power in gene-environment studies is low and requires large sample sizes.
    - Leveraging the gene-environment interaction and/or limiting the number of SNPs and tests can be alternate ways to deal with low power.

Acknowledgements

- Thomas Winkler
- Anne Justice
- Lindsay Fernandez-Rhodes
- Kristin Young
- Mary Feitosa
- Ingrid Borecki
- Iris Heid
- Ulrike Peters
- Zoltan Kutalik
- Adrienne Cupples
- DC Rao
- Penny Gordon-Larsen
- Tuomas Kilpeläinen
- Ruth Loos
- Kari North
- CHARGE, PAGE, GIANT consortia

NIH/NHGRI funding:

- R01 HD057194 Gene-Environment Interactions and Weight Gain
- 1U01HG004803 Genetic Epidemiology of Casual Variants Across the Life Course

Questions?
Comments?

Accounting for the environment in genetic analyses

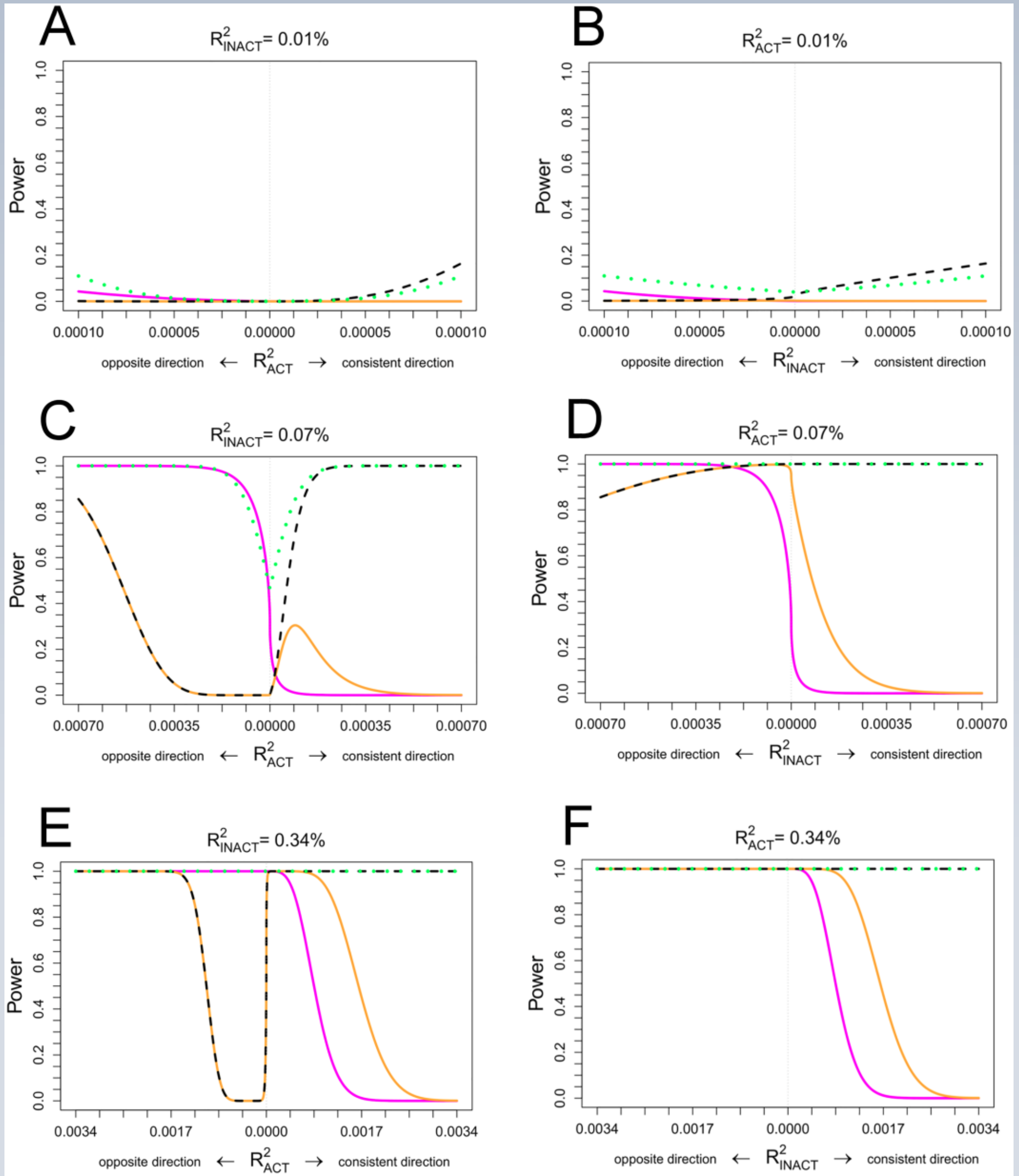
By accounting for certain environmental conditions we might be able to detect additional new genetic loci associated with a disease or trait.

$$Y = \beta_0 + \beta_G \mathrm{E} + \beta_G \mathrm{SNP} + \beta_c \mathrm{C}$$

Outcome = intercept + E + SNP + covariates

Power in GxE meta-analyses

- summary


A
R^2_{INACT}= 0.01%
B
R^2_{ACT}= 0.01%
C
R^2_{INACT}= 0.07%
D
R^2_{ACT}= 0.07%
E
R^2_{INACT}= 0.34%
F
R^2_{ACT}= 0.34%
Power
opposite direction ← R^2_{ACT} → consistent direction
opposite direction ← R^2_{INACT} → consistent direction