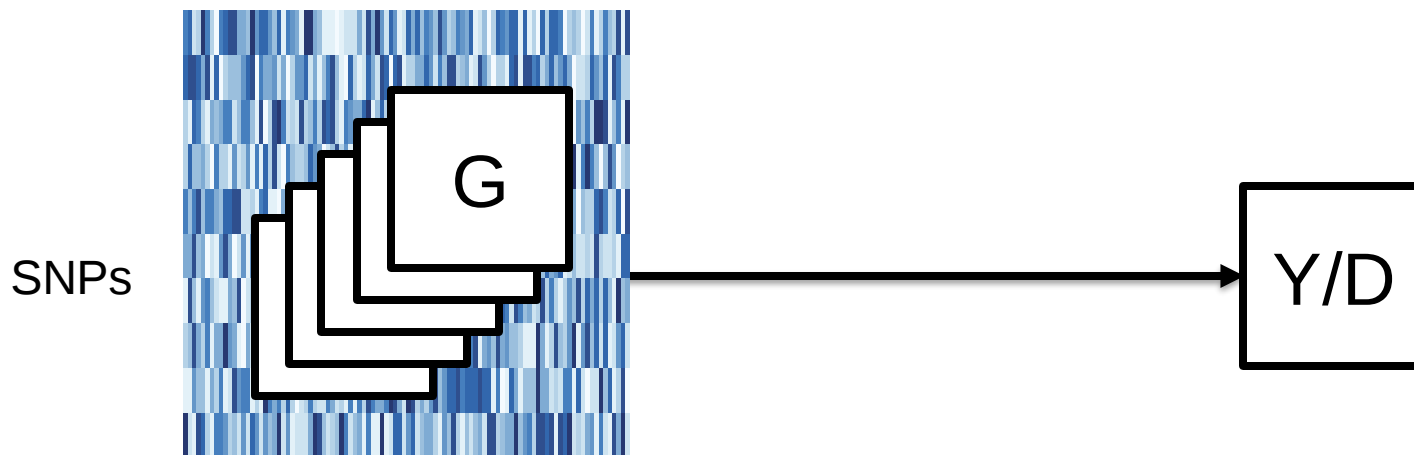


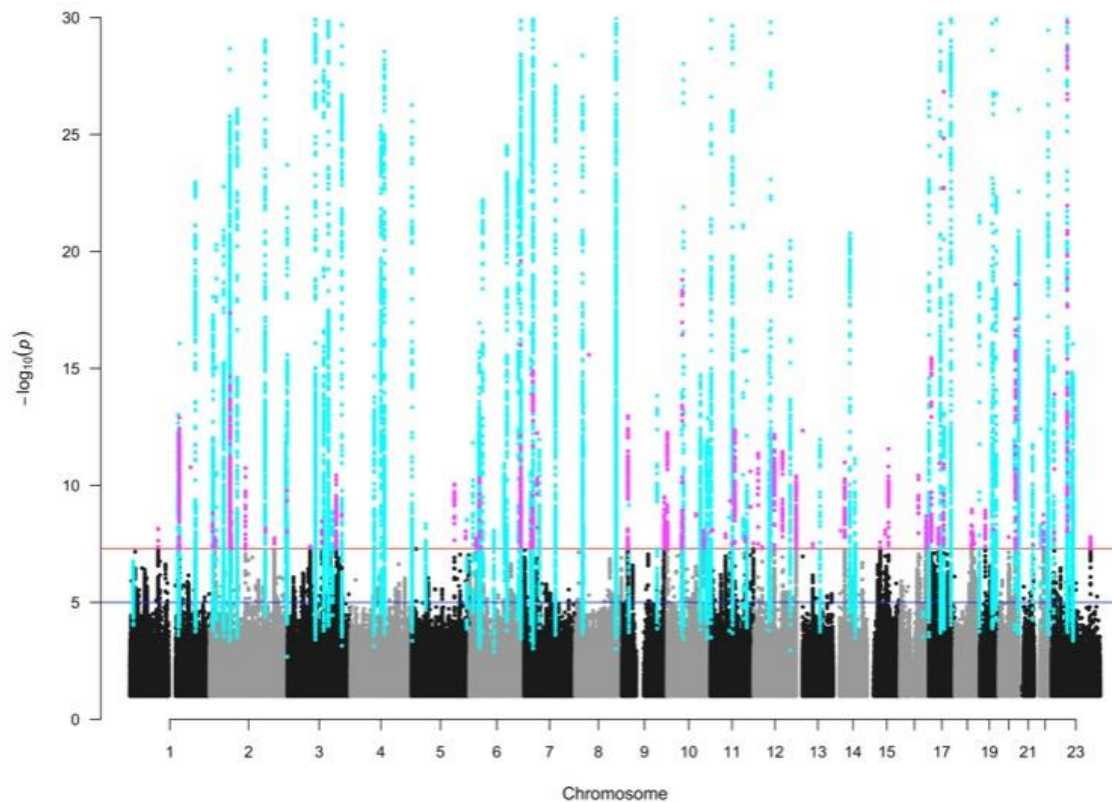
Statistical Approaches to Integrate ‘E’ with Genetics, Functional, and Multi-omic Data

David Conti, PhD
University of Southern California

Genomewide Association Studies



GWAS

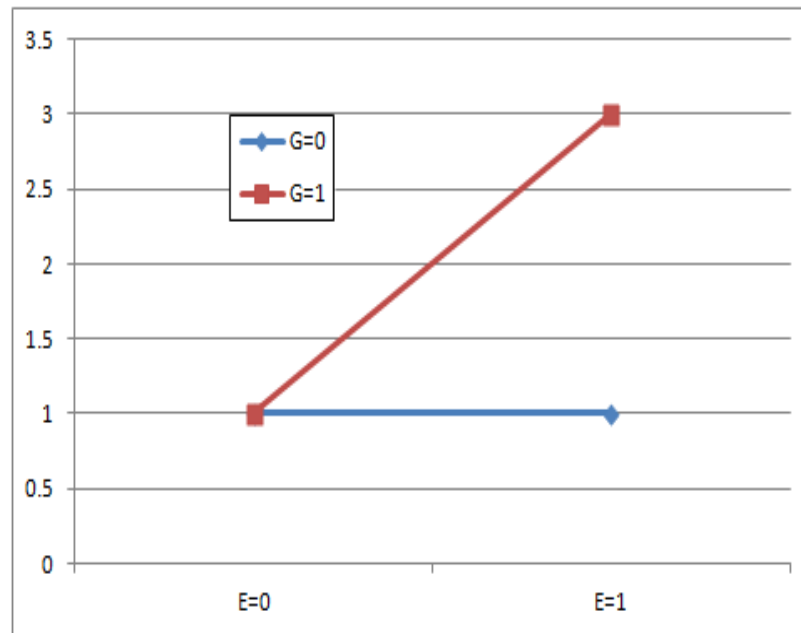


What Are We Missing?

- SNPs with modest marginal effect that might be important in one or more subgroups?

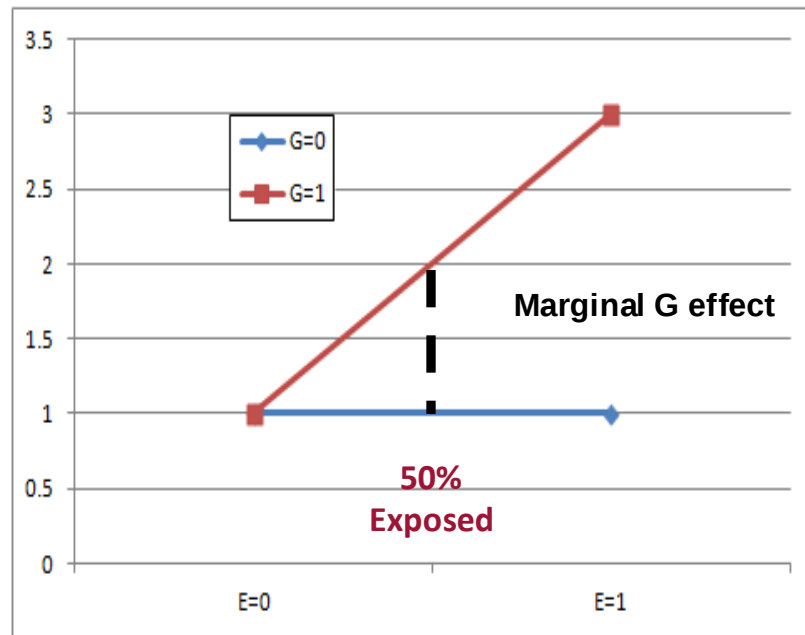
What Are We Missing?

- SNPs with modest marginal effect that might be important in one or more subgroups?



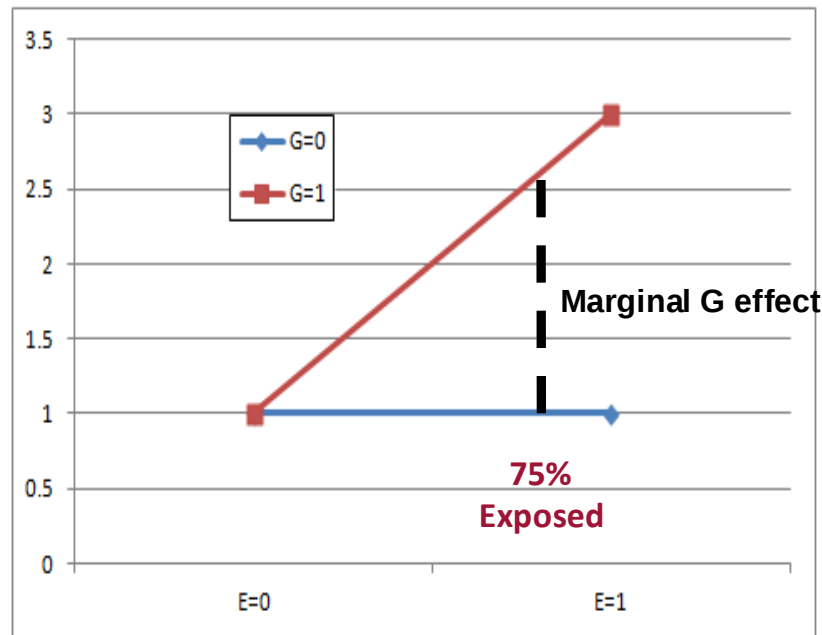
What Are We Missing?

- SNPs with modest marginal effect that might be important in one or more subgroups?
- **Size of the marginal G effect depends on prevalence of exposure**



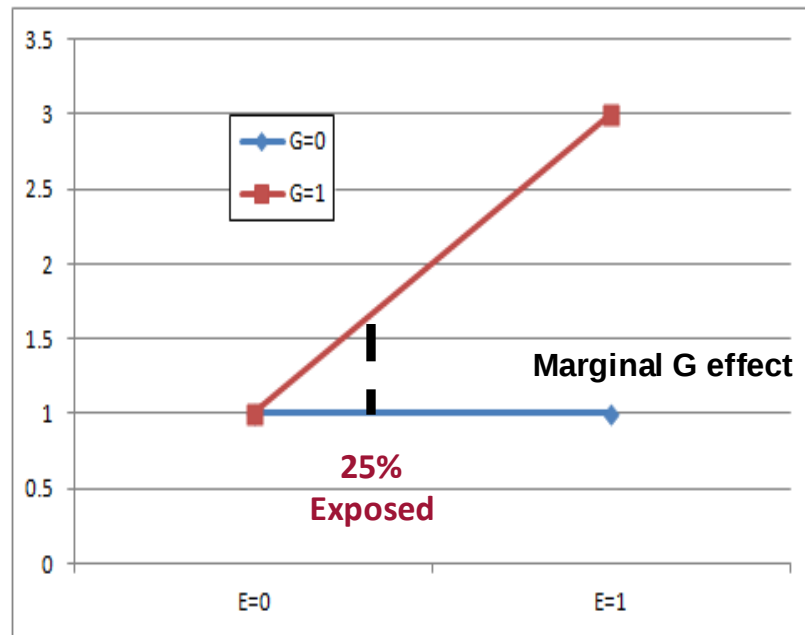
What Are We Missing?

- SNPs with modest marginal effect that might be important in one or more subgroups?
- **Size of the marginal G effect** depends on prevalence of exposure

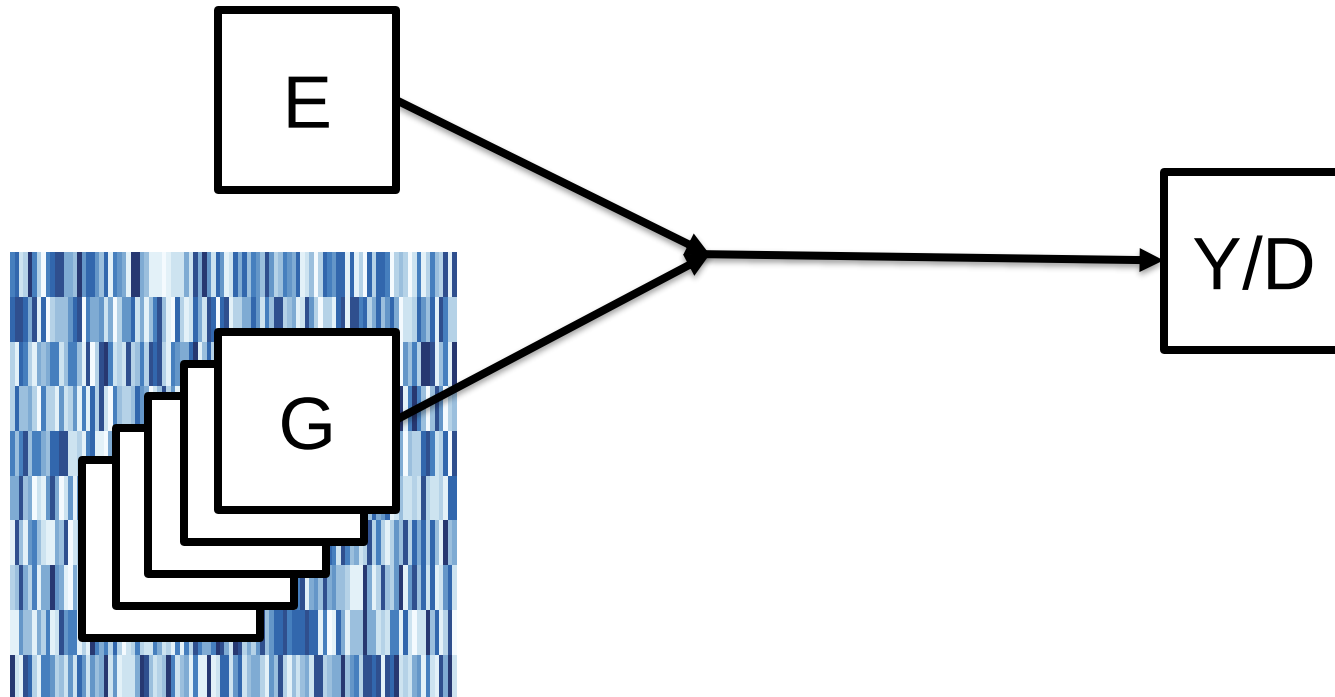


What Are We Missing?

- SNPs with modest marginal effect that might be important in one or more subgroups?
- **Size of the marginal G effect** depends on prevalence of exposure



Genomewide Interactions (GWIS)



Improving GWIS Efficiency: The Basic Idea

- For logistic regression of a case control sample:

$$\text{logit}(\text{Pr } D=1|G, E) = \alpha + \beta_G G + \beta_E E + \beta_{G \times E} G * E$$

the test of $H_0: \beta_{G \times E} = 0$ has low power

Improving GWIS Efficiency: The Basic Idea

- For logistic regression of a case control sample:

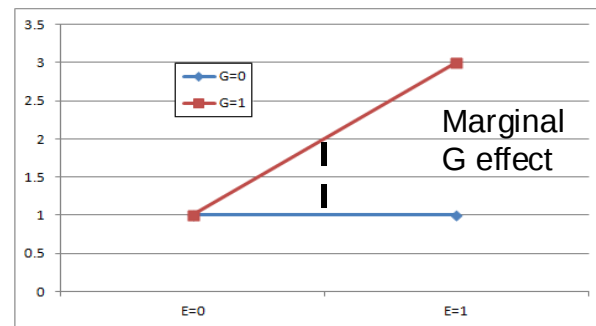
$$\text{logit}(\text{Pr } D=1|G, E) = \alpha + \beta_G G + \beta_E E + \beta_{G \times E} G * E$$

the test of $H_0: \beta_{G \times E} = 0$ has low power

- There is *additional information* in a case-control sample about GxE interaction that is not used in the above test

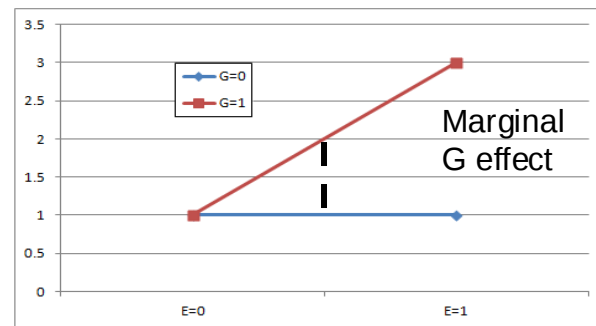
In the Presence of GxE...

- Induced “Marginal”:
 - G to D association →



In the Presence of GxE...

- Induced “Marginal”:
 - G to D association →
 - G to E association
 - ‘case-only’ style association
 - Observed in combined case-control sample if cases are oversampled relative to population prevalence
- Can we use this extra info to construct more efficient GW interaction scans?

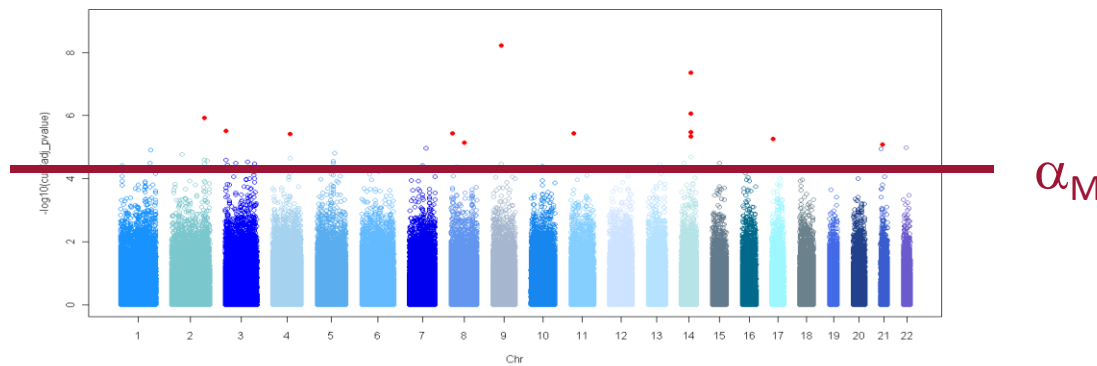


2-step Approach: **DG**|GxE

- Step 1: Genomewide screen of M SNPs using ‘marginal-effect’ test on all subjects

$$\text{Logit}[\text{Pr}(\text{D}=1 \mid G)] = \mu_0 + \mu_1 G$$

- Test $H_0: \mu_1=0$ for each SNP at α_M level



2-step Approach: **DG**|GxE

- **Step 1:** Genomewide screen of M SNPs using ‘marginal-effect’ test on all subjects

$$\text{Logit}[\Pr(D=1 \mid G)] = \mu_0 + \mu_1 G$$

- **Test H_0 :** $\mu_1=0$ for each SNP at α_M level

- **Step 2:** For m SNPs with Step-1 $p < \alpha_M$, standard GxE analysis:

$$\text{Logit}[\Pr(D=1 \mid G, E)] = \beta_0 + \beta_G G + \beta_E E + \beta_{G \times E} G \times E$$

- **Test H_0 :** $\beta_{G \times E}=0$ for the m SNPs at α/m level

2-step Approach: **EG**|GxE

- Step 1: Genomewide screen of M SNPs using 'E vs. G' test on all subjects

$$\text{Logit}[\Pr(E=1 \mid G)] = \gamma_0 + \gamma_1 G$$

- Test $H_0: \gamma_1=0$ for each SNP at α_M level

2-step Approach: **EDGE**

- Step 1: Genomewide screen of **M** SNPs using both 'D vs. G' and 'E vs. G' information
 - T_{EG} based on **E vs G** (*Murcray et al.*)
 - T_{DG} based on **D vs G** (*Kooperberg & LeBlanc*)
 - Screening Test: $T_{EgDg} = T_{EG} + T_{DG}$ (2-df test)

2-step Approach: **EDGE**

- Step 1: Genomewide screen of M SNPs using both 'D vs. G' and 'E vs. G' information
 - T_{EG} based on E vs G (Murcray et al.)
 - T_{DG} based on D vs G (Kooperberg & LeBlanc)
 - Screening Test: $T_{EgDg} = T_{EG} + T_{DG}$ (2-df test)

- Step 2: For m SNPs with Step-1 $p < \alpha_M$, standard GxE analysis:

$$\text{Logit}[\Pr(D=1 \mid G, E)] = \beta_0 + \beta_G G + \beta_E E + \beta_{G \times E} G \times E$$

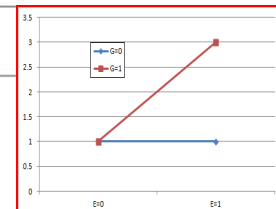
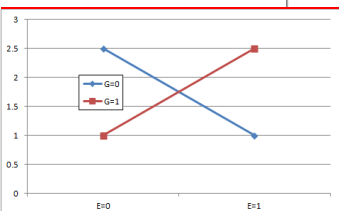
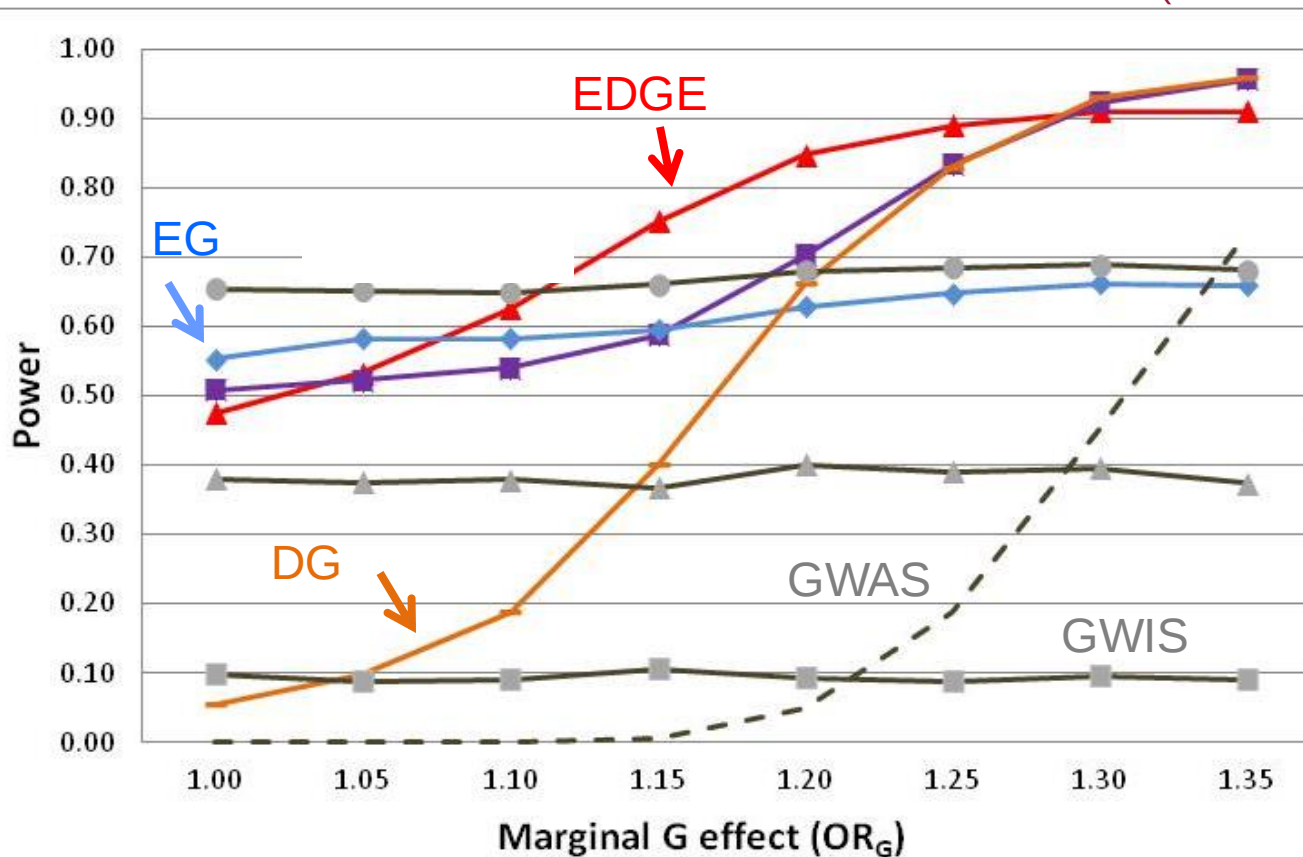
- Test $H_0: \beta_{G \times E} = 0$ for the m SNPs at α/m level

2-step
"Subset" Testing

Genomewide Power to Detect

$OR_{G \times E} = 1.5$ (N=3,500 cases, 3,500 controls)

(Gauderman et al., 2013)



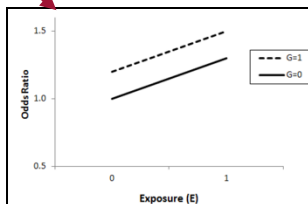
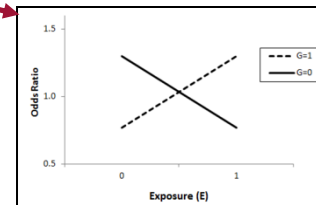
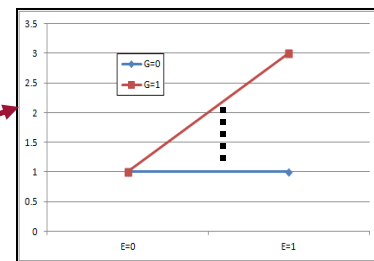
Another way to combine information:

The “2-df” joint test

$$\text{Logit}(\text{Pr } D=1|G, E) = a + \beta_G G + \beta_E E + \beta_{G \times E} G \times E$$

$H_0: \beta_G = \beta_{G \times E} = 0$ (Joint 2-df test of G, Gx E;)

- Can identify loci with ...
 - A Gx E effect and induced marginal G effect
 - A Gx E effect but no G effect
 - A G effect but no Gx E effect



The “3-df” Joint Test

$$\text{Logit}[\Pr(G=1 \mid D, E, \mathbf{C})] = \beta_0 + \beta_D D + \beta_E E + \beta_{D \times E} D \times E + \beta_C \mathbf{C}$$

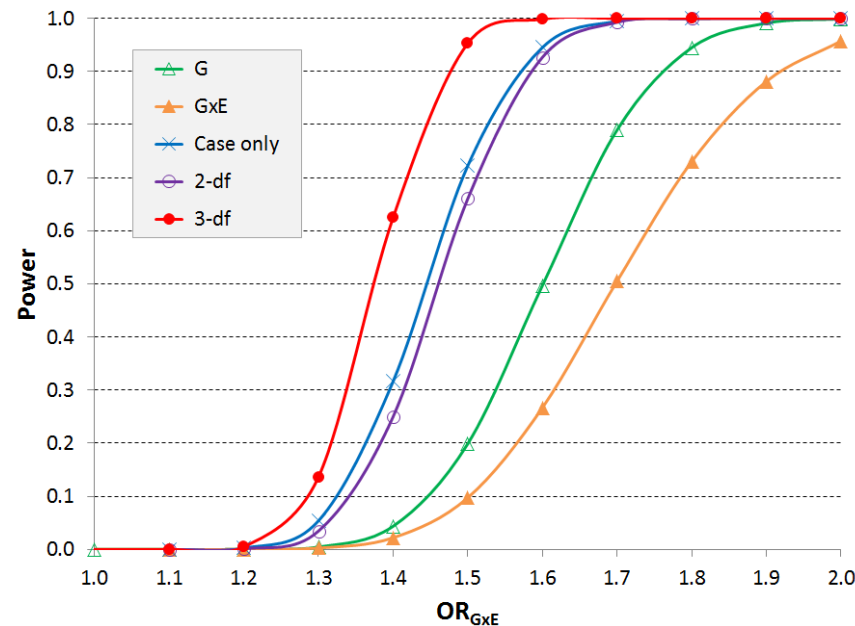
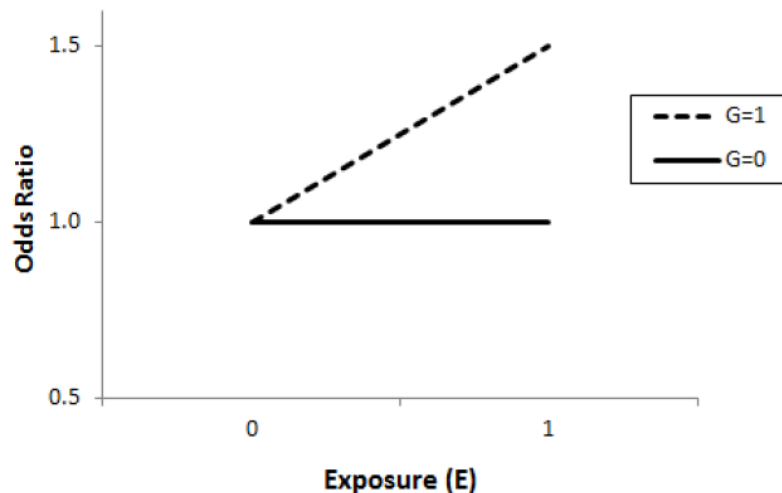
$$H_0: \beta_D = \beta_E = \beta_{D \times E} = 0$$

What is it testing?

- Marginal G vs. D association (standard GWAS)
 - Marginal G vs. E association (“case-only” style G x E)
 - G x E interaction (standard GWIS)
-
- Potentially powerful for discovery

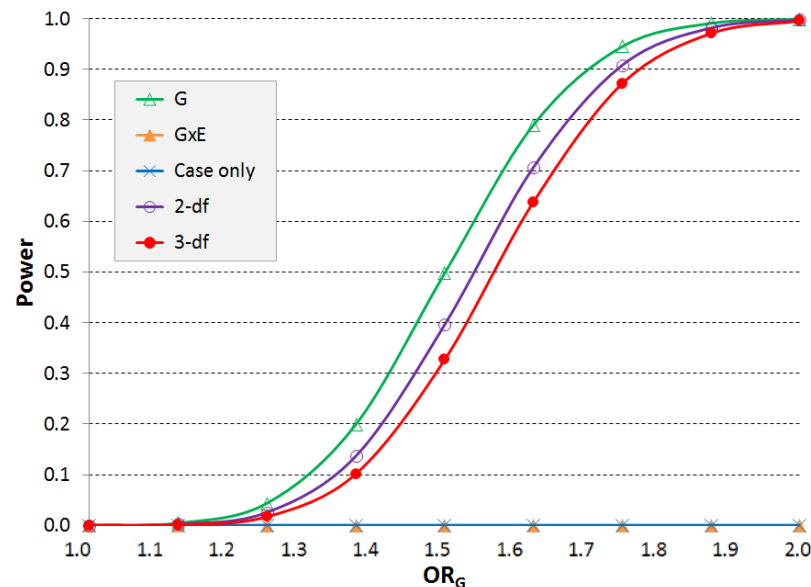
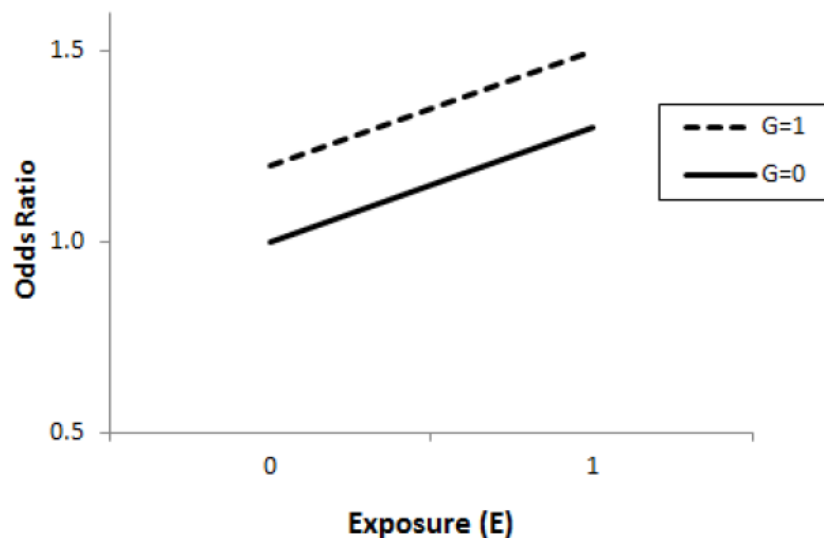
The “3-df” Joint Test: Power

Pure GxE



The “3-df” Joint Test: No Free Lunch

No GxE



GWIS Discoveries Using Efficient Methods

Authors	Exposure	GxE discovery Method(s)
Jordahl <i>et al.</i>	Alcohol	2-step
Aglago <i>et al.</i>	BMI	2-step, 3df
Diez-Obrero <i>et al.</i>	Calcium	2-step
Dimou <i>et al.</i>	Diabetes	2df, 3df
Bouras <i>et al.</i>	Folate	1df
Papadimitriou <i>et al.</i>	Fruit, Veggie, Fiber	3df
Stern <i>et al.</i>	Red meat	2-step, 3df
Tian <i>et al.</i>	HRT	2-step, 2df
Drew <i>et al.</i>	NSAIDs/aspirin	1df, 2-step, 3df
Peoples <i>et al.</i>	Physical Activity	1df, 2-step
Carreras-Torres <i>et al.</i>	Smoking	1df, 3df

G x

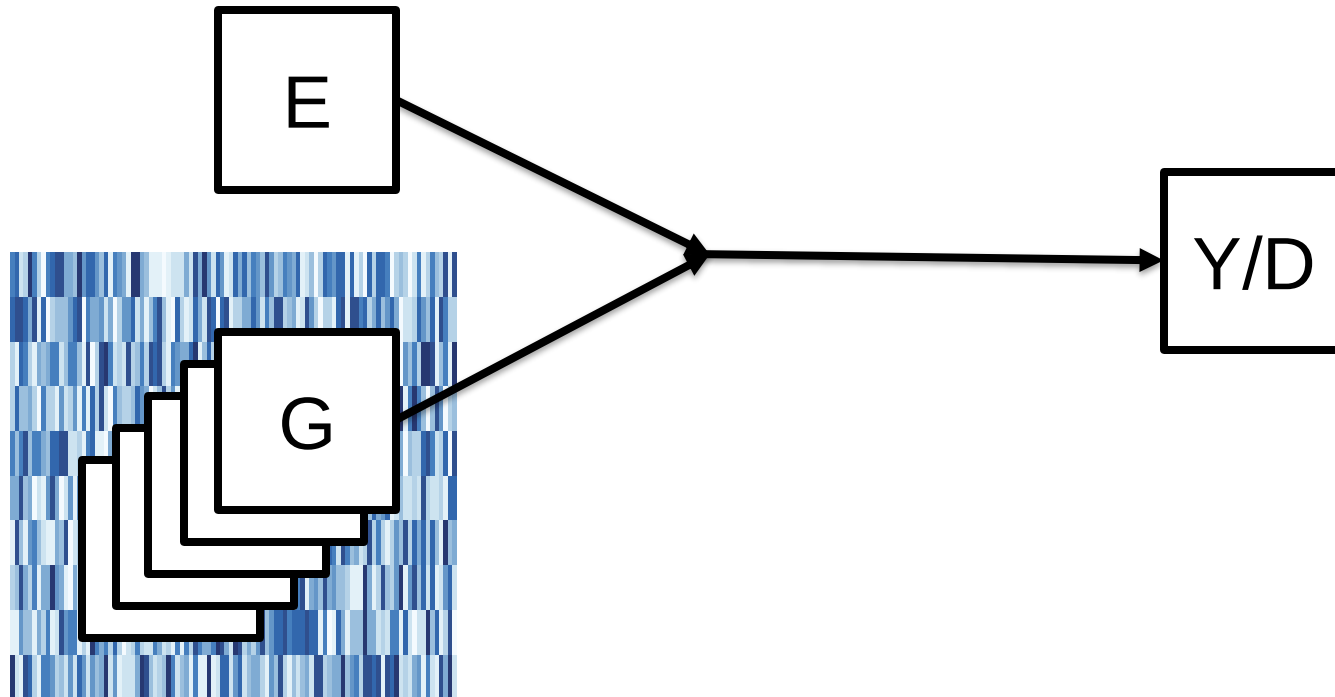


→ CRC

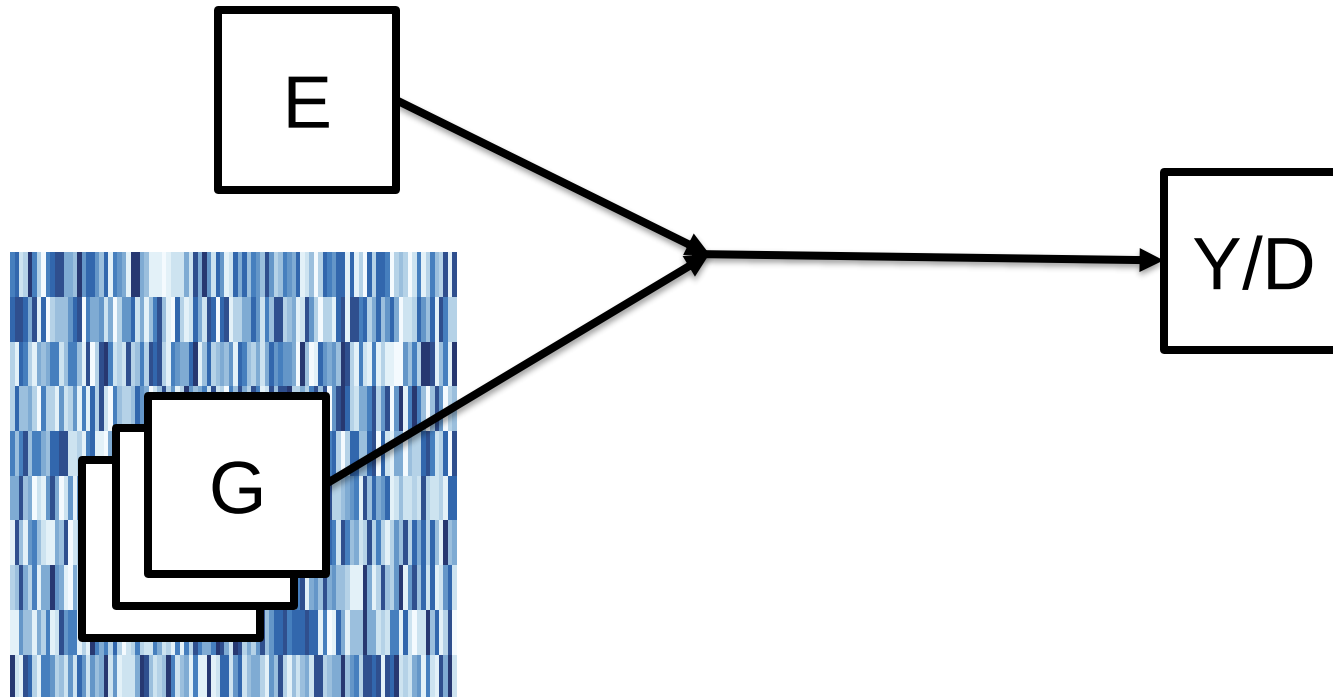
All analyses used
GxEscanR



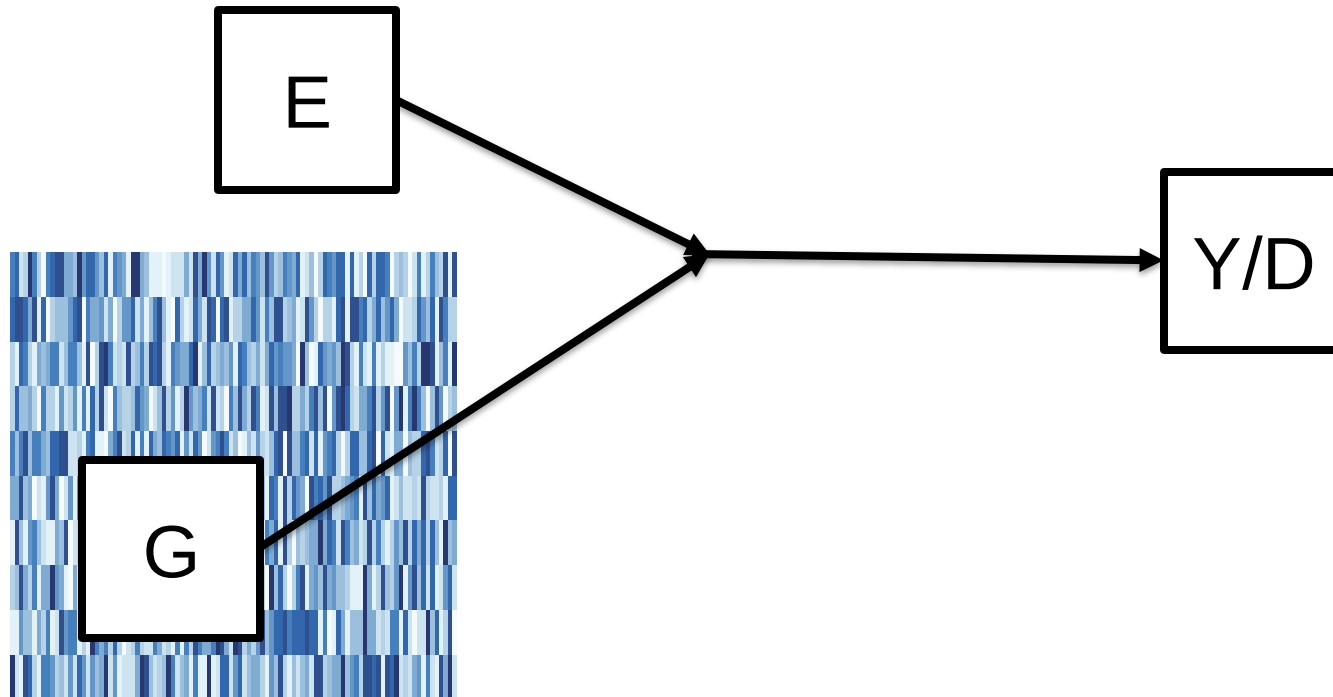
Many Single-Marker Interactions



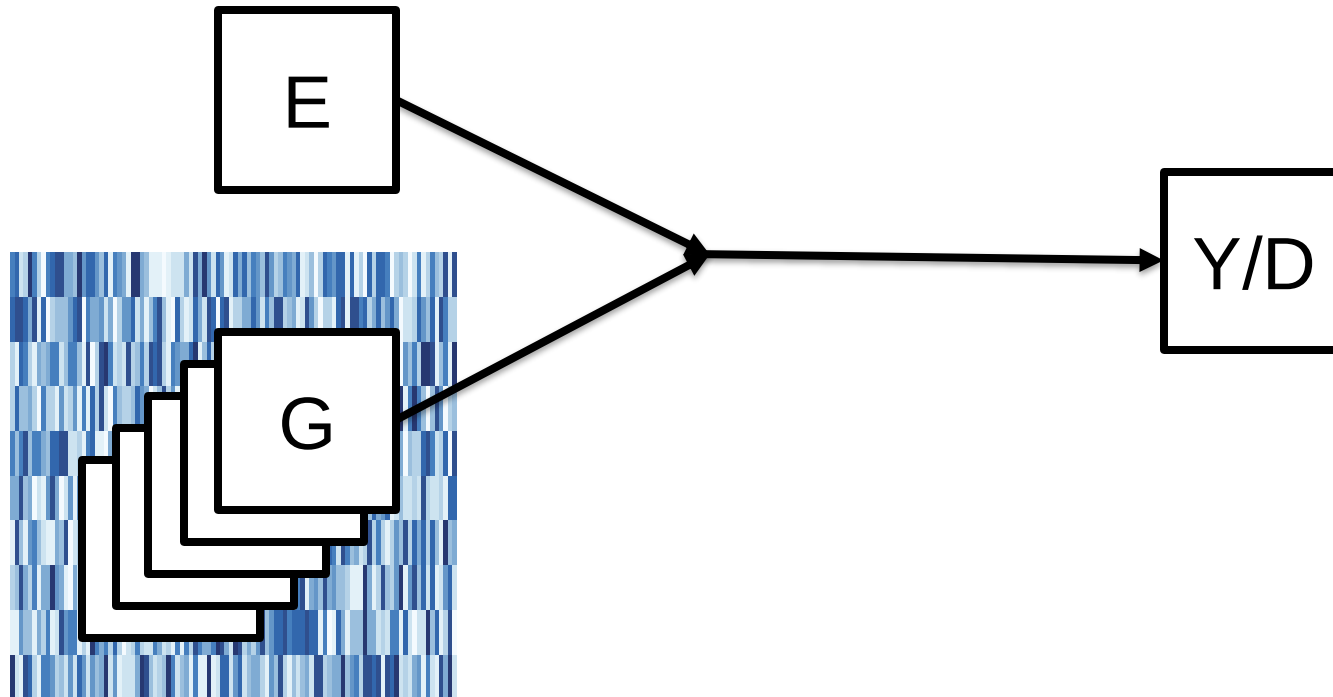
Many Single-Marker Interactions



Many Single-Marker Interactions



High Dimensional Interactions



Single-marker analysis vs. joint analysis

single-marker:
one-SNP-at-a-time $Y \sim \beta_0 + \beta_E E + \beta_{G_j} G_j + \beta_{G_j \times E} G_j \times E$, for each $j = 1, \dots, p$

joint:
all p SNPs together $Y \sim \beta_0 + \beta_E E + \sum_{j=1}^p \beta_{G_j} G_j + \sum_{j=1}^p \beta_{G_j \times E} G_j \times E$

- Polygenic traits
 - Nature of the signal is multi-marker/polygenic for complex traits
- Joint analysis considers the impact other markers on the outcome
 - A weak effect may be more apparent when other causal effects are already accounted for
 - A false signal may be weakened by inclusion in the model of a stronger signal from a true causal association

Single-marker analysis vs. joint analysis

single-marker:
one-SNP-at-a-time $Y \sim \beta_0 + \beta_E E + \beta_{G_j} G_j + \beta_{G_j \times E} G_j \times E$, for each $j = 1, \dots, p$

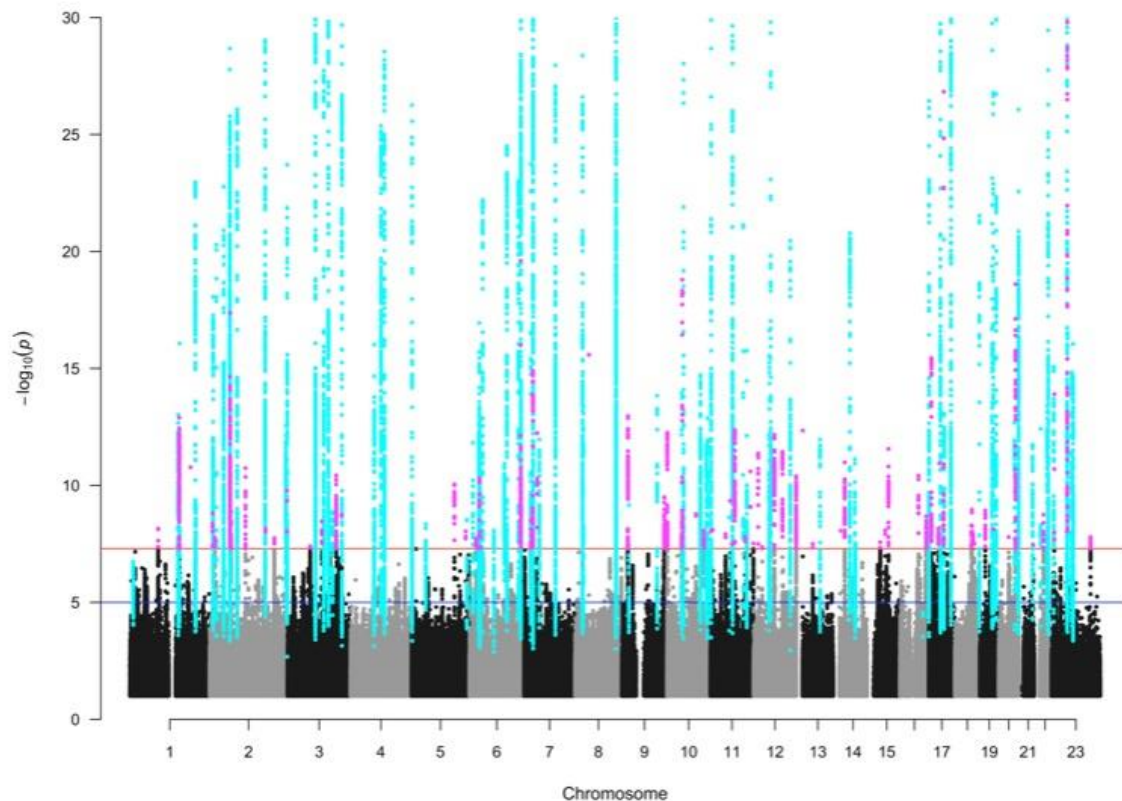
joint:
all p SNPs together $Y \sim \beta_0 + \beta_E E + \sum_{j=1}^p \beta_{G_j} G_j + \sum_{j=1}^p \beta_{G_j \times E} G_j \times E$

- gesso [G(by)E(la)sso] model

subject to $\left\{ \begin{array}{l} (1) \sum_{j=1}^p (|\beta_{G_j}| + |\beta_{G_j \times E}|) \leq t \\ (2) |\beta_{G_j \times E}| \leq |\beta_{G_j}| \end{array} \right.$ Hierarchical Constraints

$\beta_{G \times E} \neq 0 \Rightarrow \beta_G \neq 0$ or
 $\beta_G = 0 \Rightarrow \beta_{G \times E} = 0$

GWAS and Polygenic Risk



Polygenic Risk Score (PRS):

*Weighted sum of # risk alleles carried
by each participant*

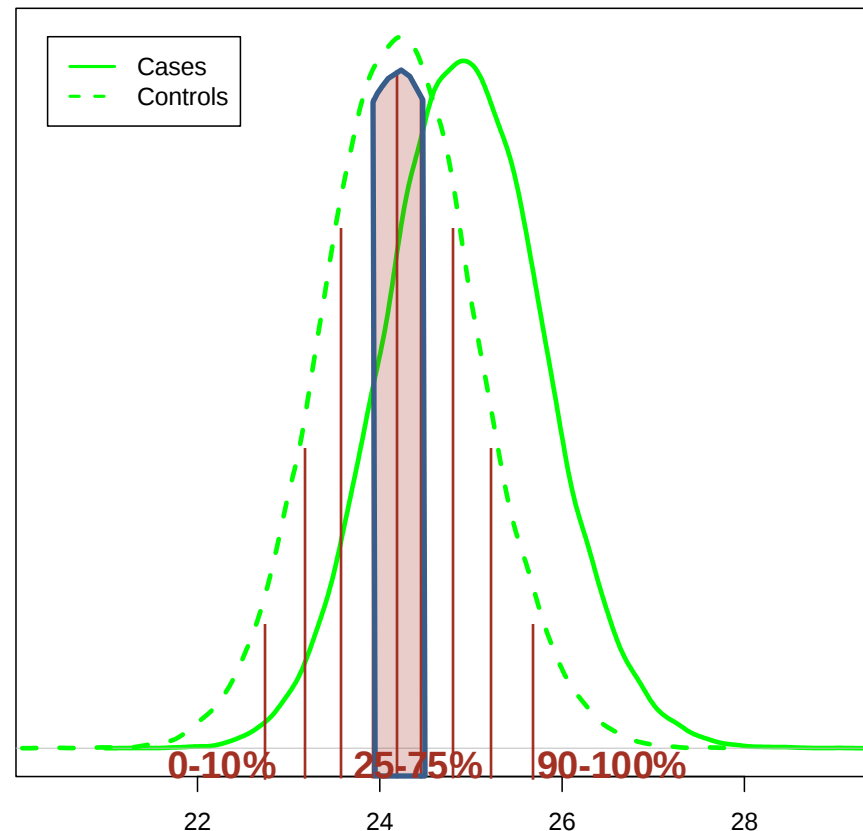
Count of risk alleles for
variant m for individual i

What SNPs?

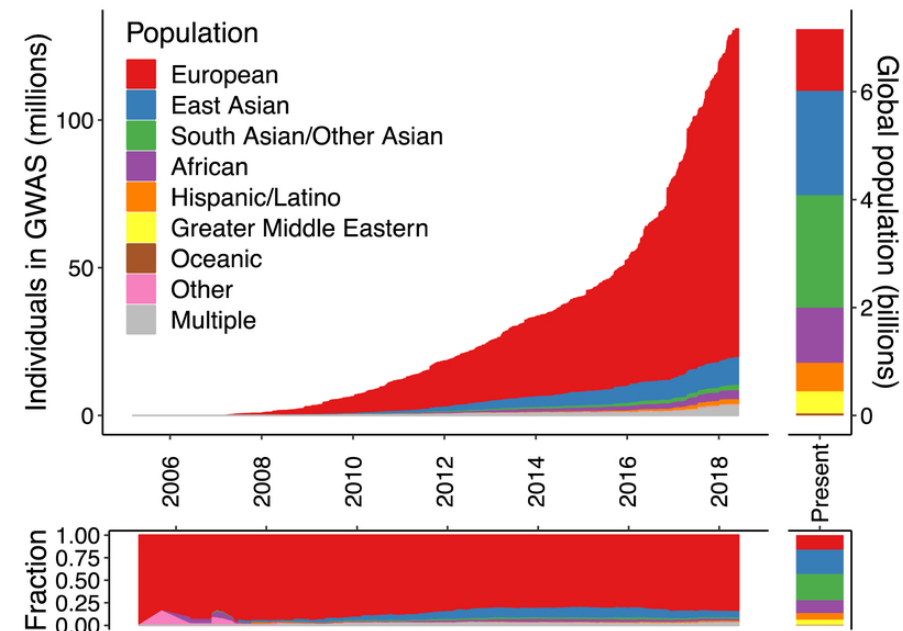
$$PRS_i = \sum_{m=1}^M w_m G_{im}$$

What weight?

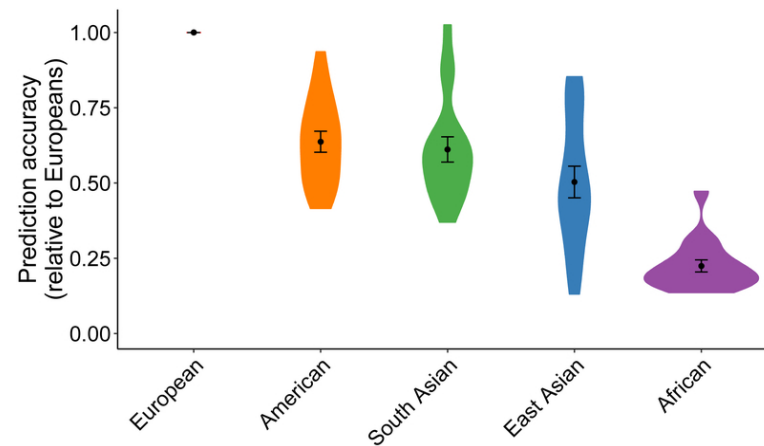
Density



Lack of Diversity Could Impact Health Disparities

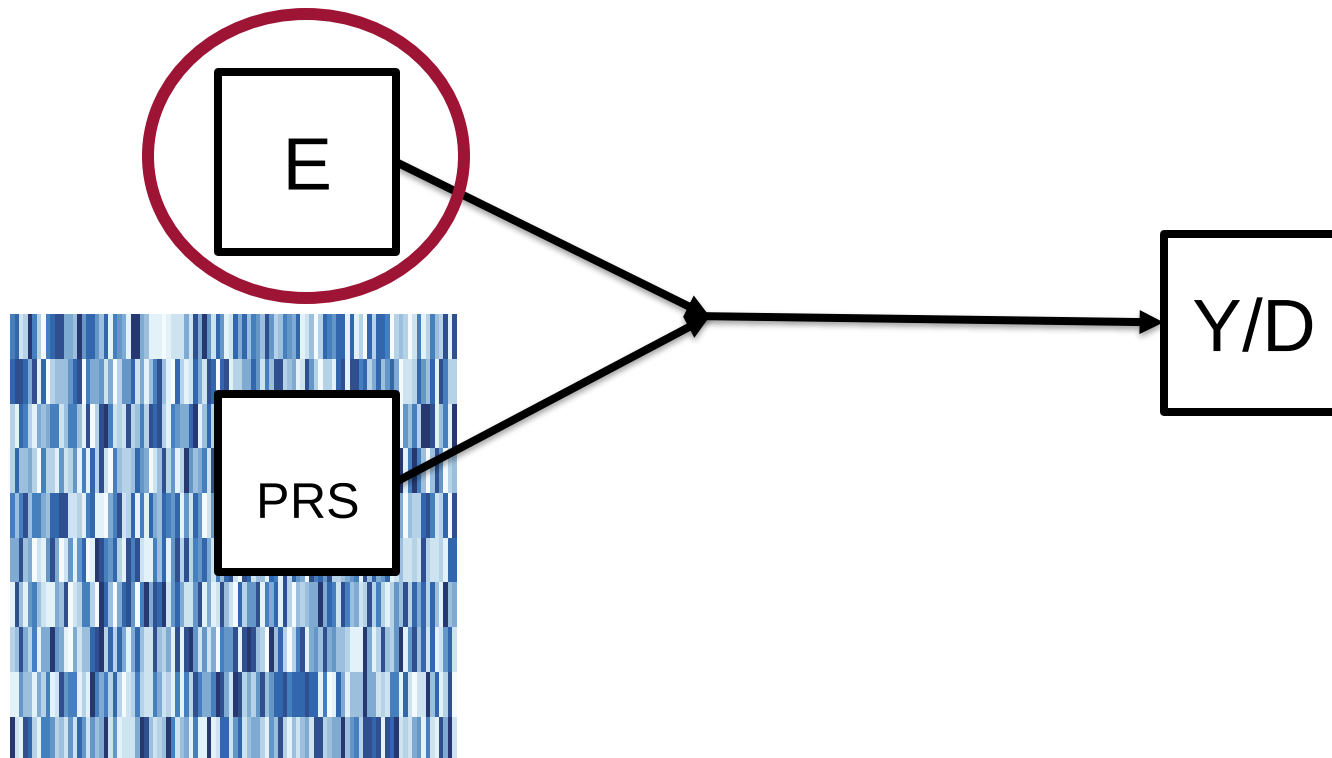


Ancestry of GWAS participants over time relative to the global population



Polygenic prediction accuracy relative to European ancestry individuals across 17 quantitative traits

Polygenic Risk Score and E Interactions

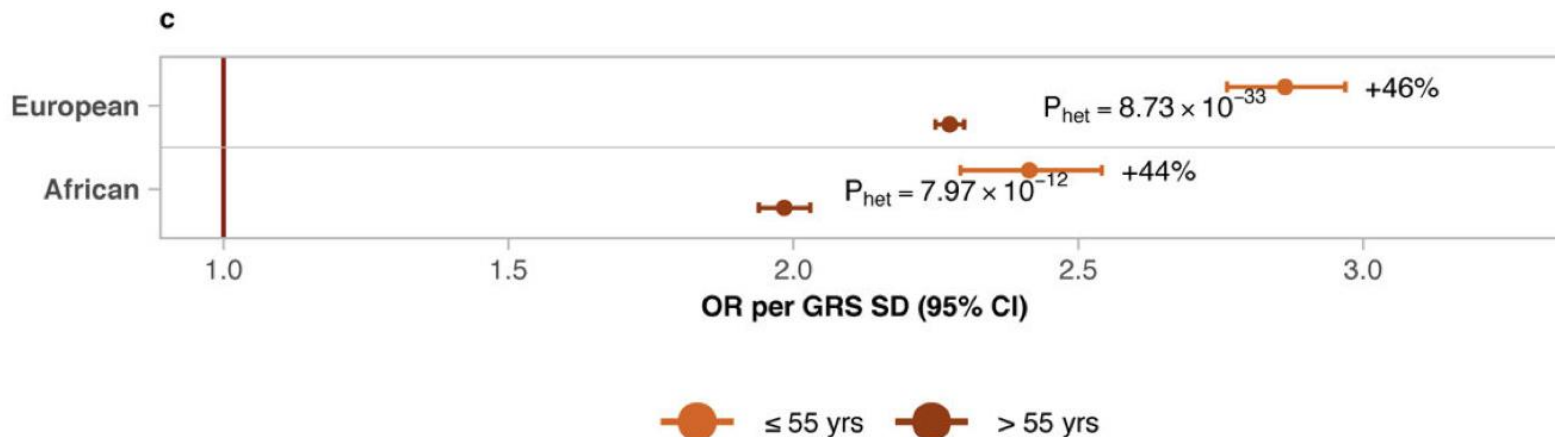


PRS Across Populations in Prostate Cancer

- 156,319 prostate cancer cases
- 788,443 controls
- European, African, Asian and Hispanic men
- A 57% increase in the number of non-European cases from previous GWAS.

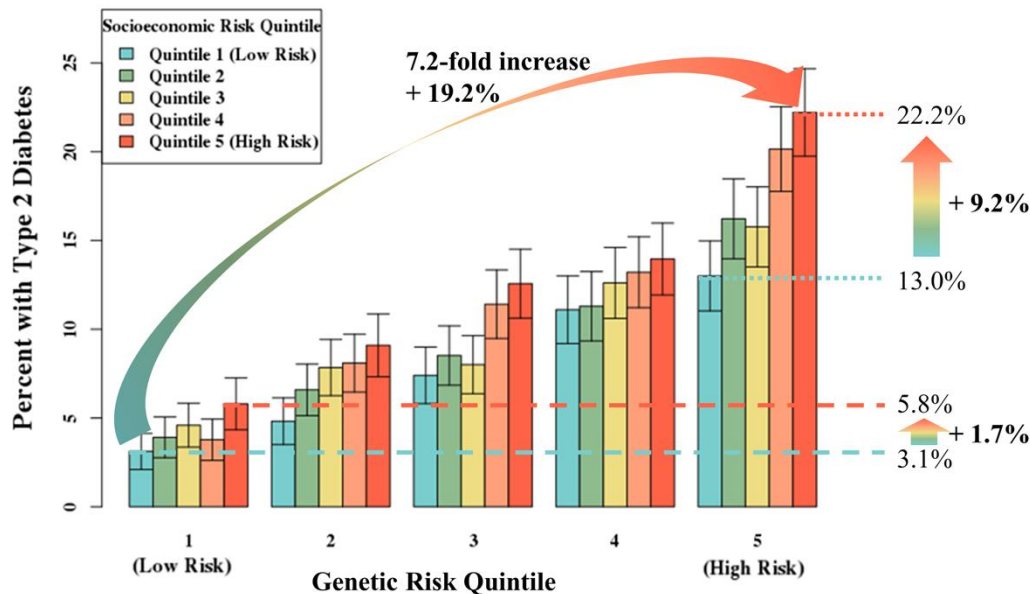
Population	1 SD OR
European	2.32 [95%CI: 2.30-2.35]
African	2.04 [95%CI: 2.00-2.08]
Asian	2.15 [95%CI: 1.99-2.32]
Hispanics	2.12 [95%CI: 2.03-2.23]

PRS x Age in Prostate Cancer



PRS x SDoH

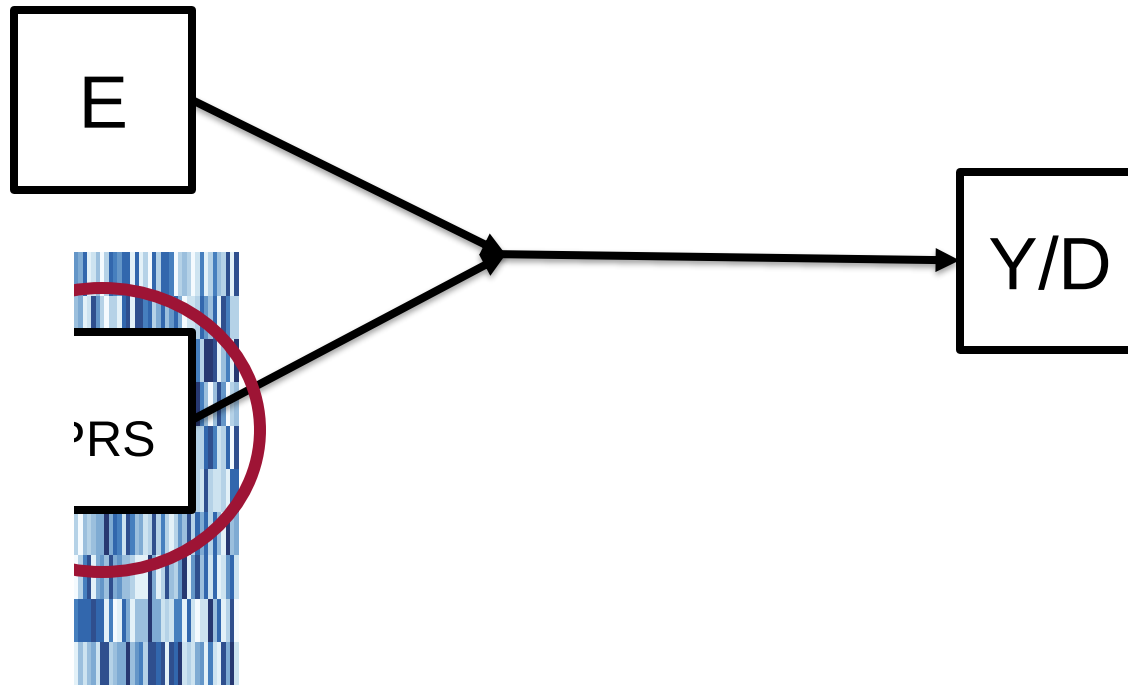
What is the combined effect of genetic and socioeconomic risk on the prevalence of type 2 diabetes (T2D) and obesity?



Combined high genetic and socioeconomic risk, compared to combined low risk, was associated with a 7-fold and 3-fold increase, respectively, in T2D and obesity prevalence.

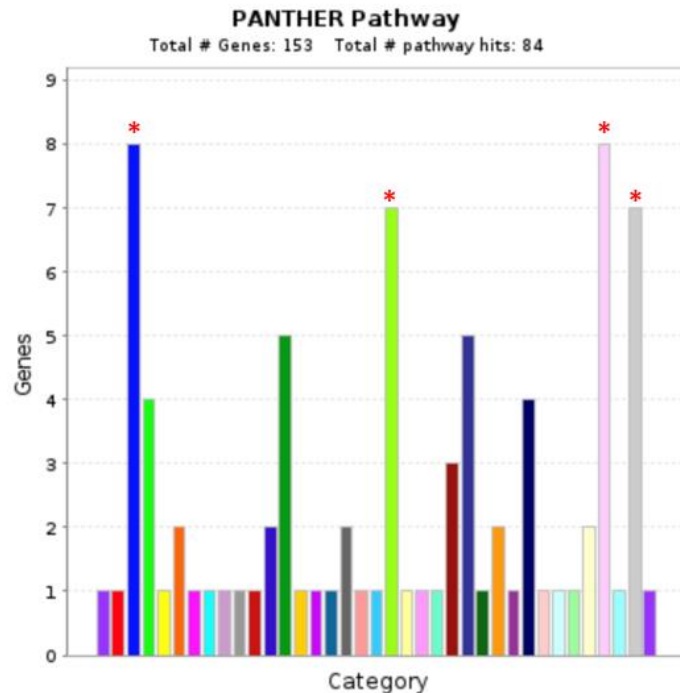
Increasing socioeconomic risk is associated with a greater absolute increase in T2D and obesity prevalence among those at high genetic risk compared to those at low genetic risk.

Polygenic Risk Score and E Interactions



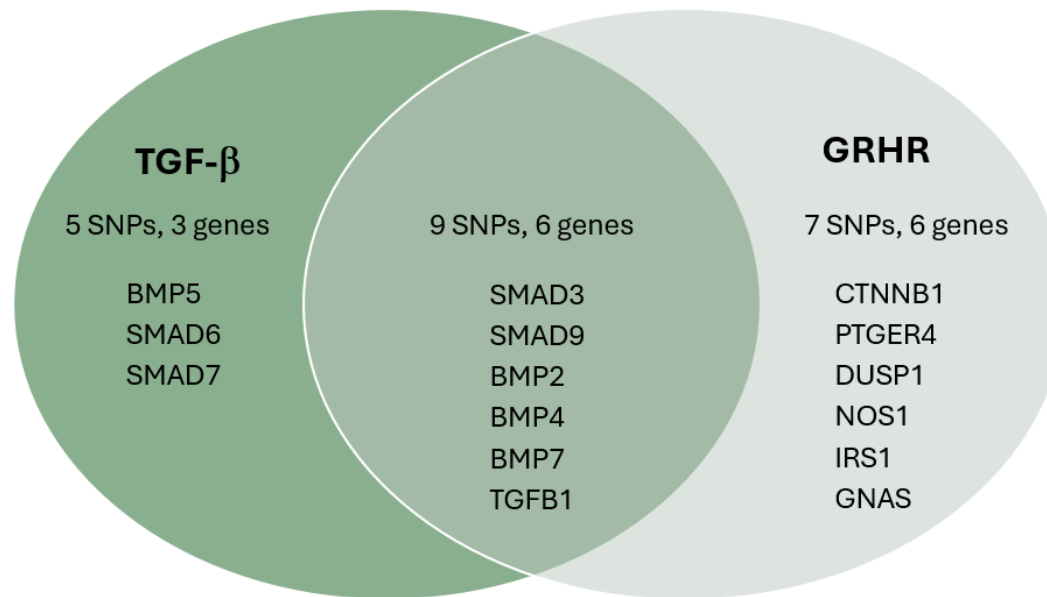
Colorectal Cancer PRS: Incorporating Functional Annotations

- ANNOQ (Liu et al., 2022) used to annotate GWAS SNPs to genes (153/205 SNPs annotated)
- PANTHER (Mi et al., 2017) used for pathway analysis of those genes



- 5HT4 type receptor mediated signaling pathway (P04376)
- Adrenaline and noradrenaline biosynthesis (P00001)
- * Alzheimer disease-presenilin pathway (P00004)
- Angiogenesis (P00005)
- Apoptosis signaling pathway (P00006)
- Axon guidance mediated by Slit/Robo (P00008)
- Axon guidance mediated by netrin (P00009)
- B cell activation (P00010)
- Beta1 adrenergic receptor signalling pathway (P04377)
- Beta2 adrenergic receptor signalling pathway (P04378)
- Beta3 adrenergic receptor signalling pathway (P04379)
- CCKR signaling map (P06959)
- Cadherin signalling pathway (P00012)
- Cell cycle (P00013)
- Corticotropin releasing factor receptor signaling pathway (P00014)
- Cytoskeletal regulation by Rho GTPase (P00016)
- Endothelin signaling pathway (P00019)
- Enkephalin release (P05913)
- Glycolysis (P00024)
- * Gonadotropin-releasing hormone receptor pathway (P06664)
- Heterotrimeric G-protein signaling pathway-Gi alpha and Gs (P00025)
- Histamine H2 receptor mediated signaling pathway (P04386)
- Huntington disease (P00029)
- Inflammation mediated by chemokine and cytokine signaling (P00030)
- Integrin signalling pathway (P00034)
- Metabotropic glutamate receptor group III pathway (P00035)
- Nicotinic acetylcholine receptor signaling pathway (P00040)
- Notch signaling pathway (P00045)
- PDGF signaling pathway (P00047)
- PI3 kinase pathway (P00048)
- Purine metabolism (P02769)
- Pyruvate metabolism (P02772)
- T cell activation (P00053)
- * TGF-beta signaling pathway (P00052)
- Vasopressin synthesis (P04395)
- * Wnt signaling pathway (P00057)
- p38 MAPK pathway (P05918)

SNPs In Multiple Pathways



PRS* x NSAIDs

Table S2: Analysis of PGS Catalog derived polygenic risk score x NSAIDs interaction for Colorectal Cancer

PRS Type	PRS		E (NSAIDs use)		PRS x E		
	OR ^a	(95% CI)	OR	(95% CI)	OR	(95% CI)	p-value ^b
PRS: All SNPs*	1.59	(1.56, 1.61)	0.77	(0.74, 0.79)	0.98	(0.95, 1.01)	0.240
<u>Pathways^{&}</u>							
pPRS: TGF- β	1.18	(1.16, 1.20)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.017
pPRS: Gonadotropin-receptor	1.17	(1.15, 1.18)	0.76	(0.74, 0.79)	0.96	(0.94, 1.00)	0.021
pPRS: Cadherin-signaling	1.10	(1.08, 1.11)	0.76	(0.74, 0.79)	1.00	(0.97, 1.03)	0.840
pPRS: Alzheimer's presenillin	1.08	(1.07, 1.10)	0.76	(0.74, 0.79)	0.99	(0.96, 1.02)	0.640
PRS Other [#]	1.51	(1.48, 1.53)	0.77	(0.74, 0.79)	0.998	(0.97, 1.03)	0.900

* PRS formed based on 204 GWAS significant SNPs with weights extracted from the PGS Catalog

& pPRS based on subsets of the 204 SNPs within the indicated pathway

PRS based on the subset of 174 of the 204 SNPs that are not within any of the indicated pathways

a Odds ratios (OR) are scaled to a 1 s.d. increase for the indicated PRS and compare users to non-users for NSAIDs

b p-value tests the null hypothesis of no PRS x E interaction. For PRS and E main effects, all $p < 10^{-10}$.

PRS* x NSAIDs

Table S3: Analysis of PGS catalog-derived pPRS x NSAIDs for SNPs in the TGF- β and GRHR pathways

PRS Type	PRS		E (NSAIDS use)		PRS x E		
	OR ^a	(95% CI)	OR	(95% CI)	OR	(95% CI)	p-value ^b
<u>Pathways^{&}</u>							
TGF-Beta (14 SNPs)	1.18	(1.16, 1.20)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.017
Gonadotropin-receptor (16 SNPs)	1.17	(1.15, 1.18)	0.76	(0.74, 0.79)	0.96	(0.94, 1.00)	0.021

TGF-Beta or Gonadotropin (21 SNPs)	1.21	(1.19, 1.23)	0.76	(0.74, 0.79)	0.95	(0.92, 0.98)	0.0009
TGF-Beta Unique (5 SNPs)	1.12	(1.10, 1.14)	0.76	(0.74, 0.79)	0.96	(0.93, 1.00)	0.019
Gonadotropin Unique (7 SNPs)	1.10	(1.08, 1.11)	0.76	(0.74, 0.79)	0.96	(0.93, 0.99)	0.010
TGF-GNR shared (9 SNPs)	1.13	(1.11, 1.15)	0.76	(0.74, 0.79)	0.99	(0.96, 1.02)	0.351

* PRS formed based on 204 GWAS significant SNPs with weights extracted from the PGS Catalog

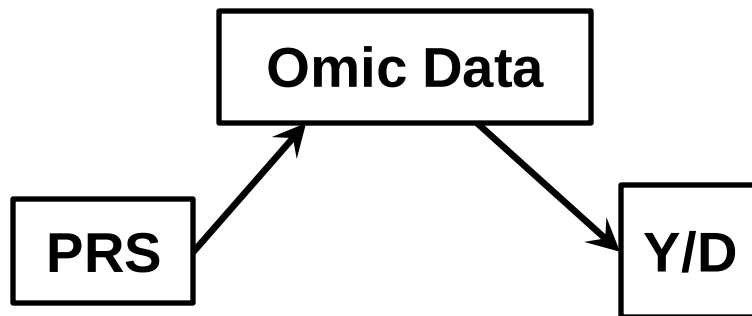
& pPRS based on subsets of the 204 SNPs within the indicated pathway

PRS based on the subset of 174 of the 204 SNPs that are not within any of the indicated pathways

a Odds ratios (OR) are scaled to a 1 s.d. increase for the indicated PRS and compare users to non-users for NSAIDs

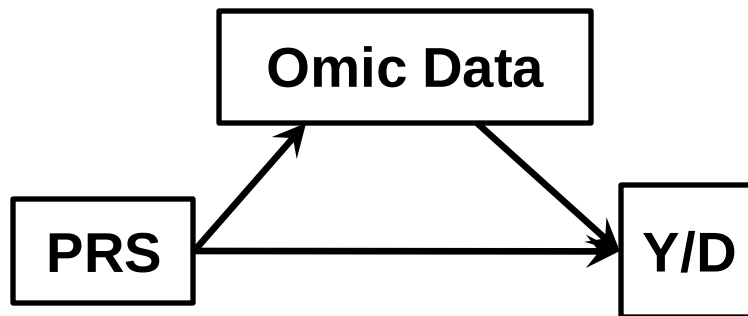
b p-value tests the null hypothesis of no PRS x E interaction. For PRS and E main effects, all $p < 10^{-10}$.

Omic Data



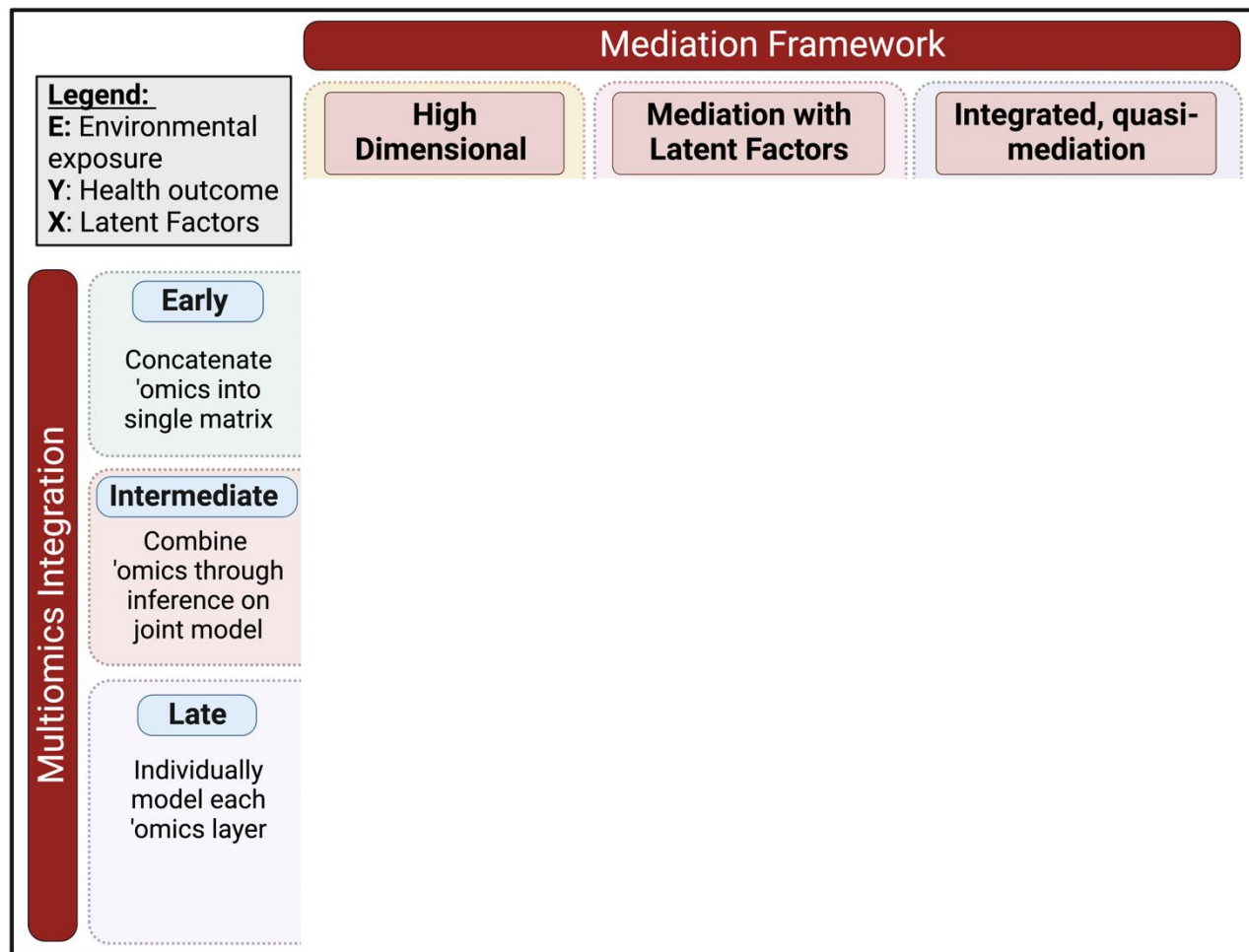
Can we clarify the impact of each SNP within a PRS with ***measured*** omic data that captures the underlying biology?

Omic Mediation

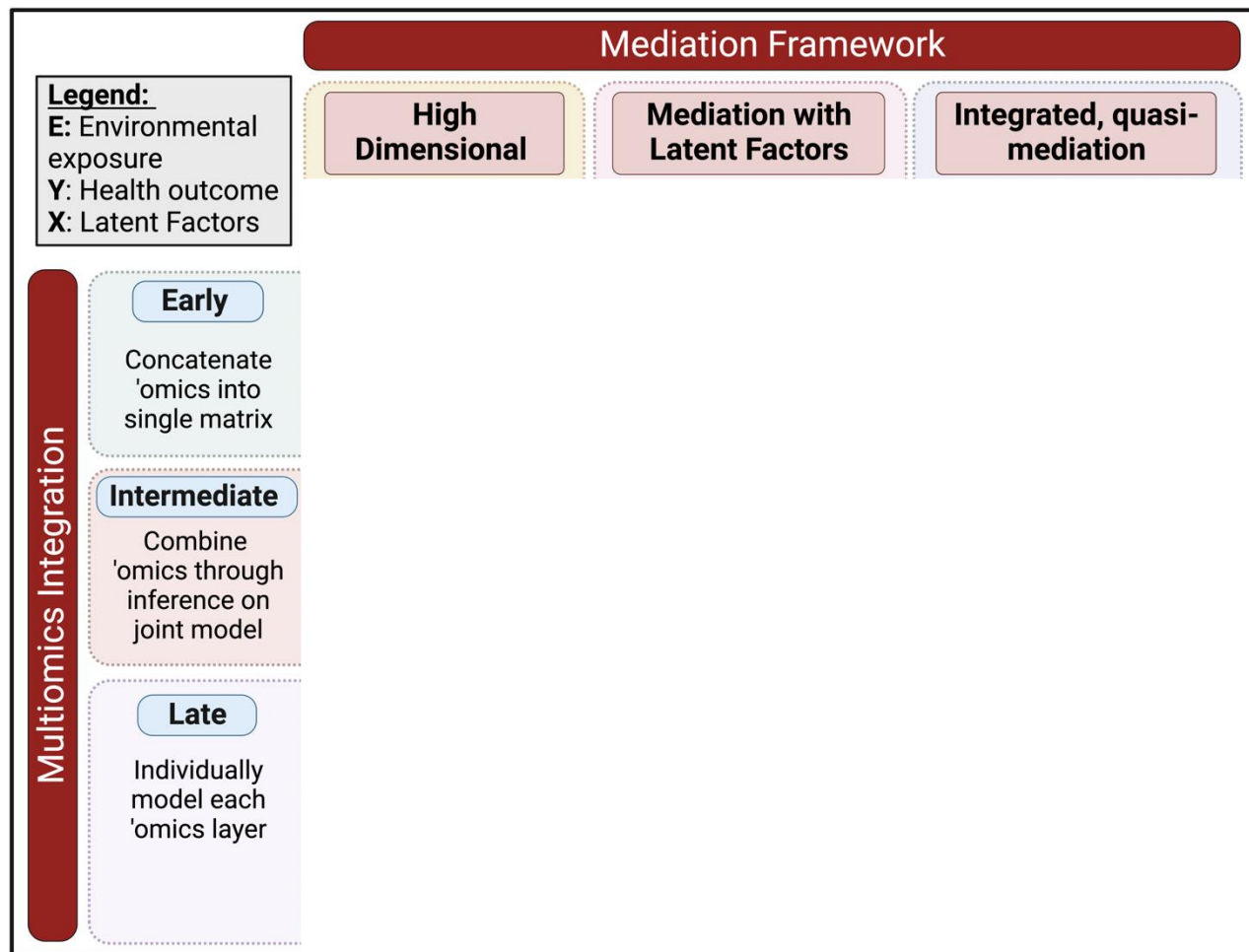


Can we clarify the impact of each SNP within a PRS with ***measured*** omic data that captures the underlying biology?

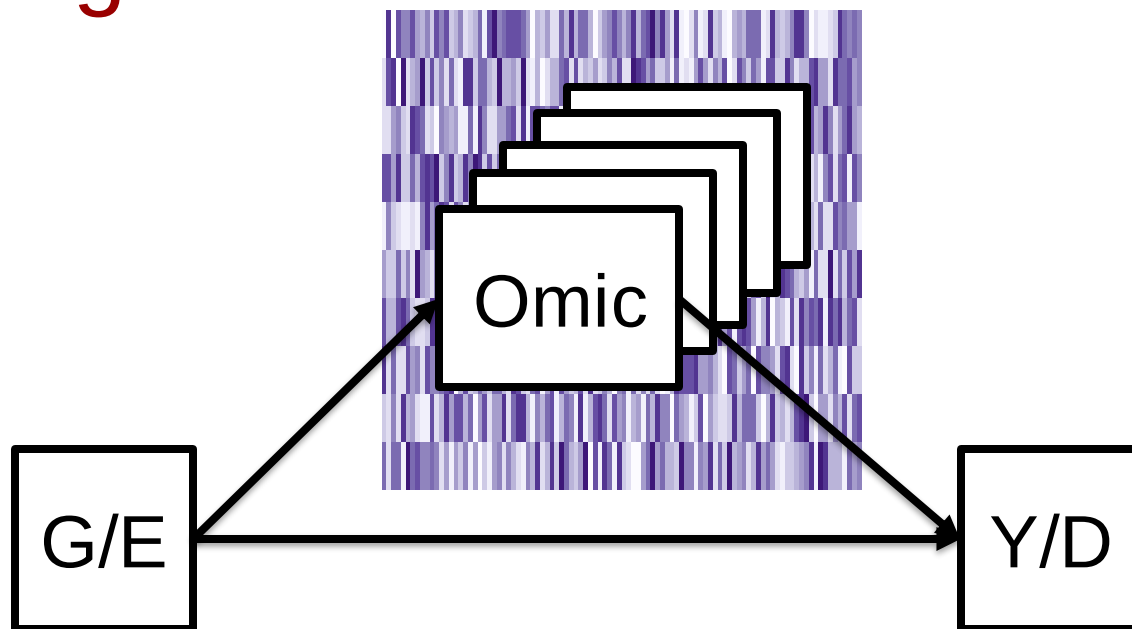
Multiomic Mediation Framework For Precision Environmental Health



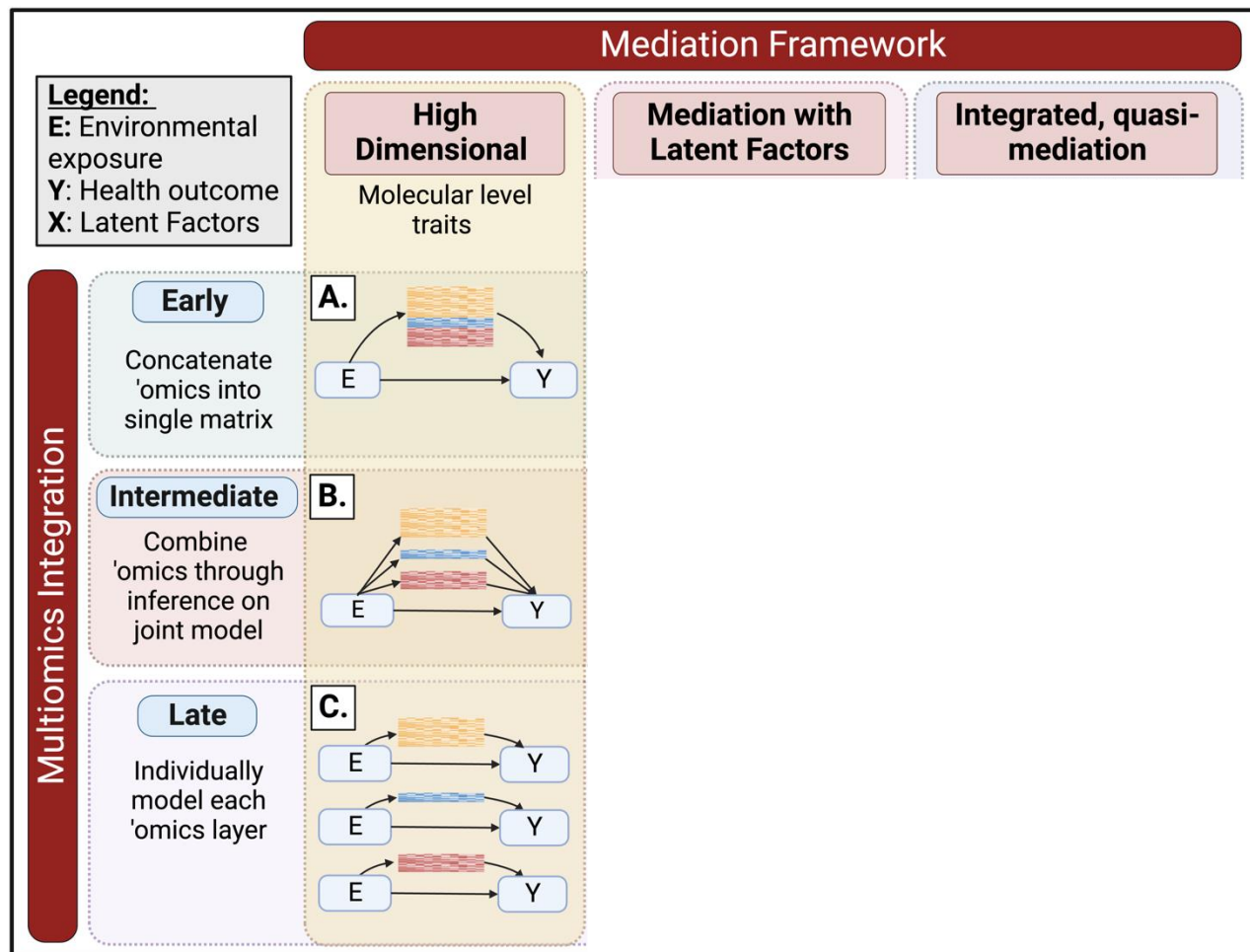
Multiomic Mediation Framework For Precision Environmental Health



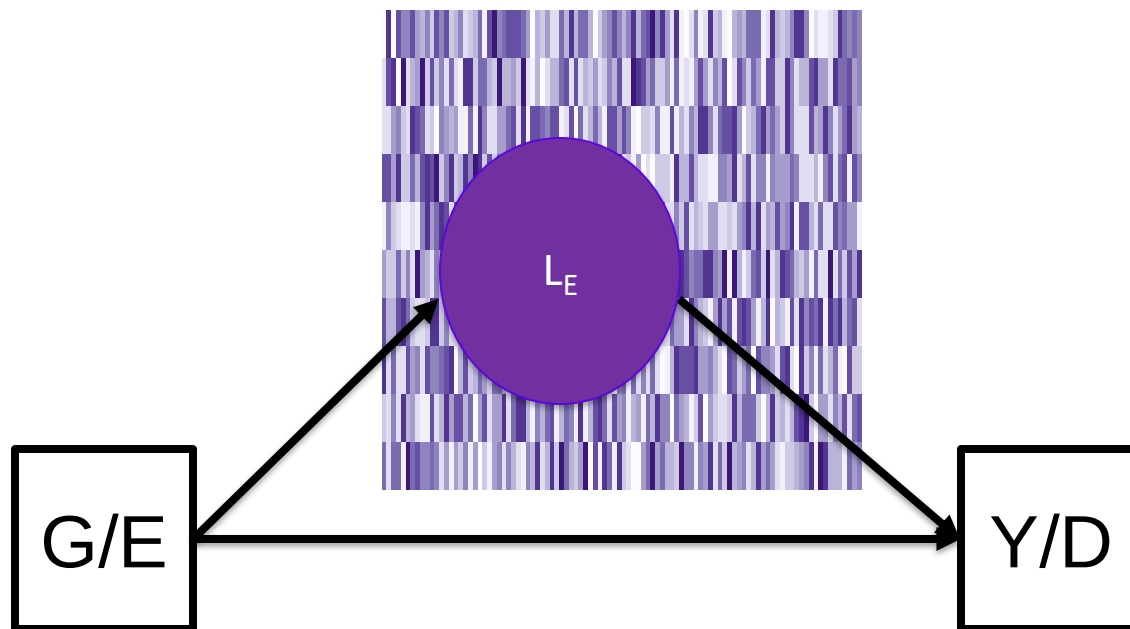
High Dimensional Mediation



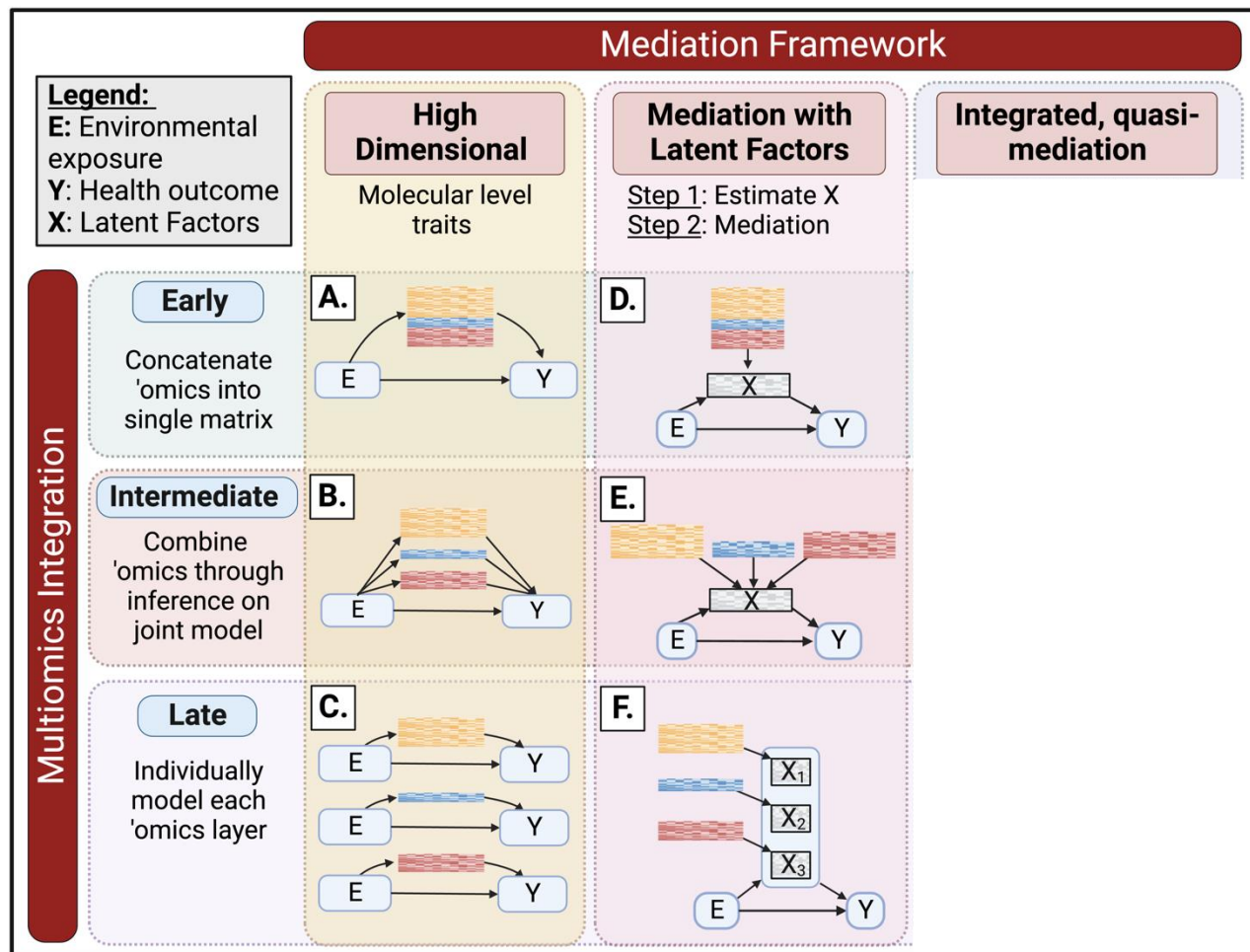
Multiomic Mediation Framework For Precision Environmental Health



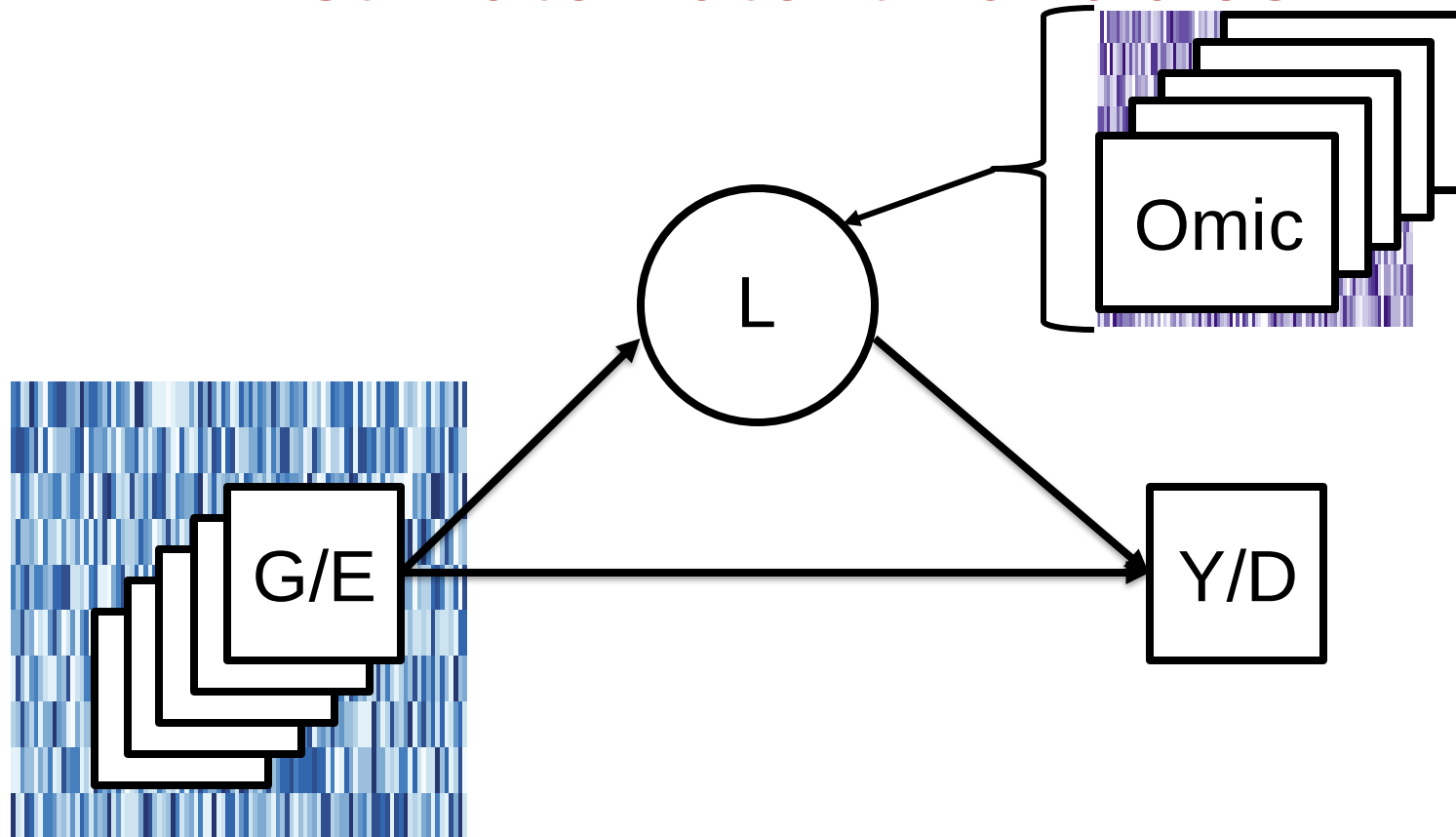
Latent Mediation



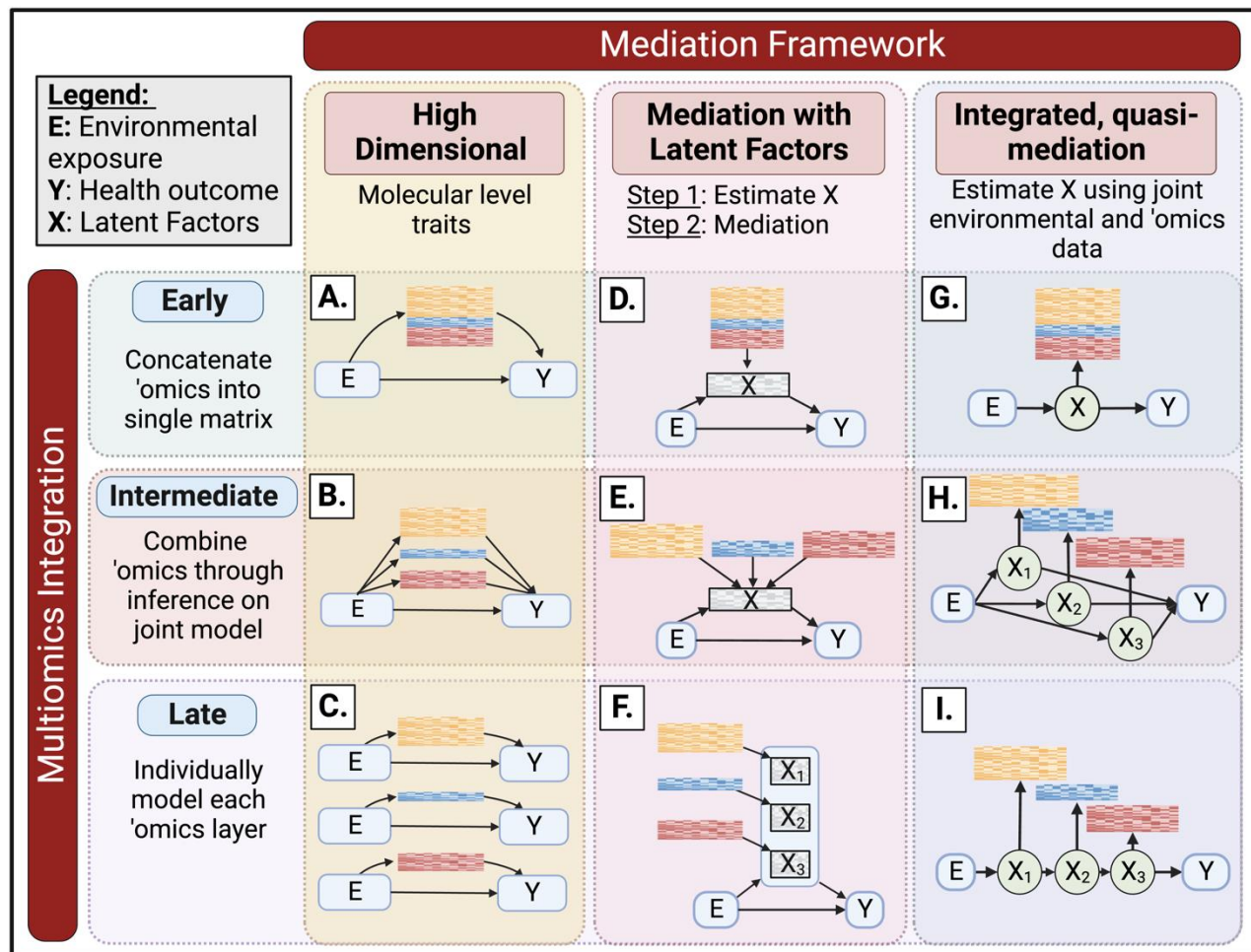
Multiomic Mediation Framework For Precision Environmental Health



Estimate Latent Variables

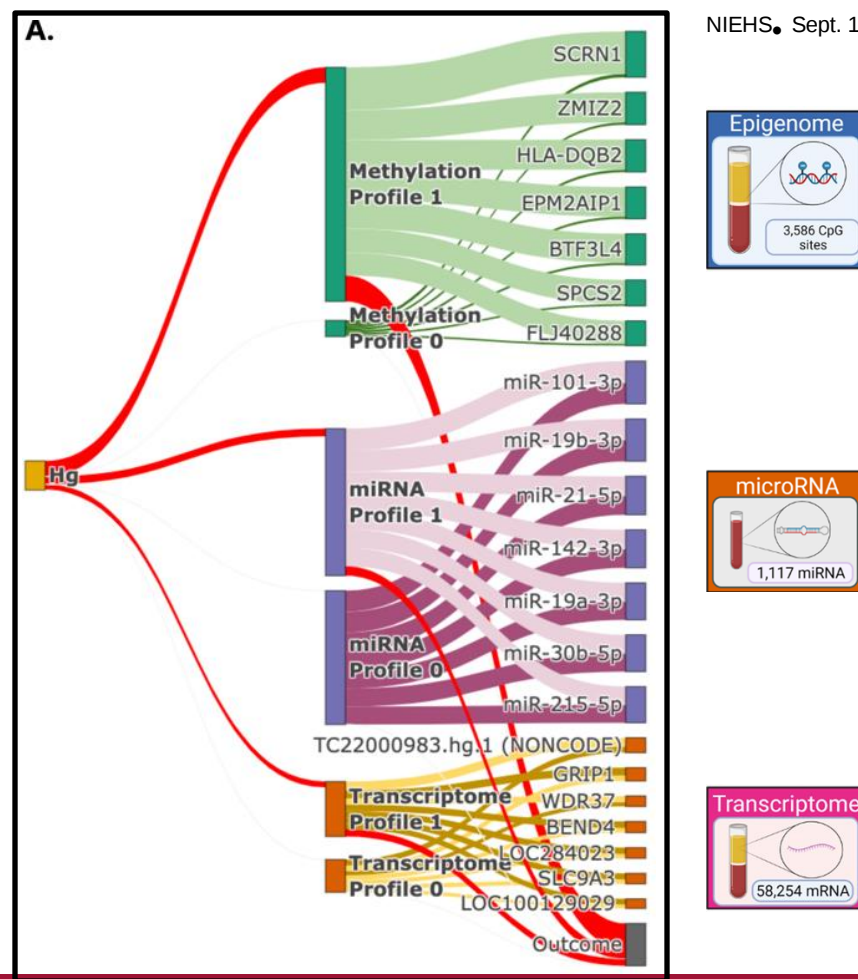
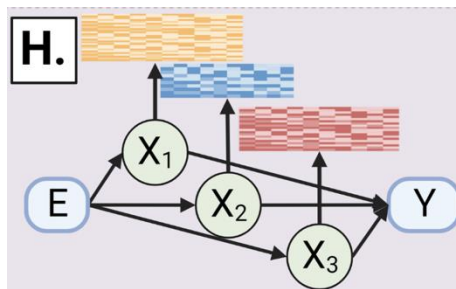


Multiomic Mediation Framework For Precision Environmental Health

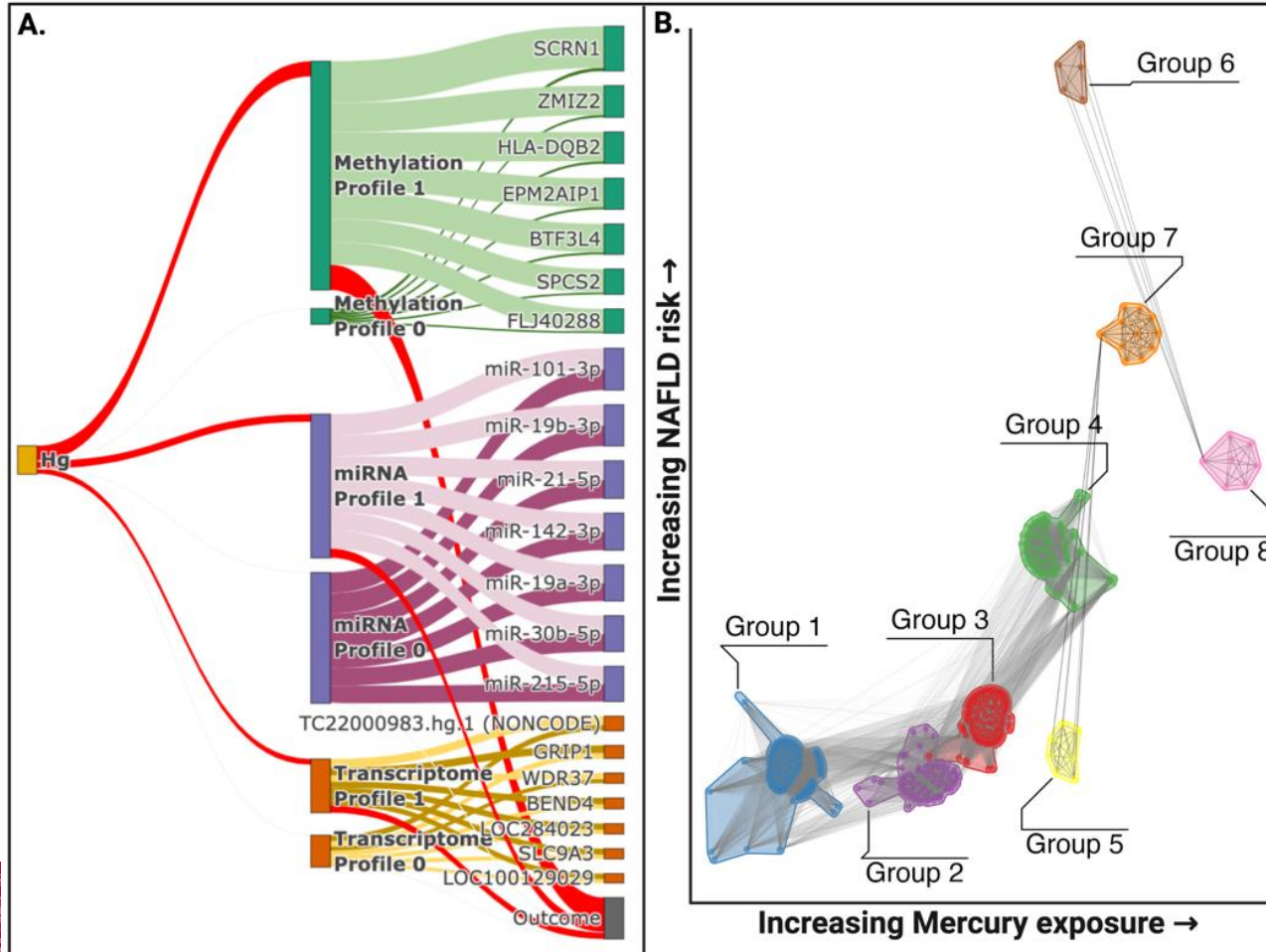


Joint Analysis, Intermediate Integration

Integrated information on environmental exposures, DNA methylation, miRNA levels, and transcripts can identify groups of children at elevated **risk of liver injury**

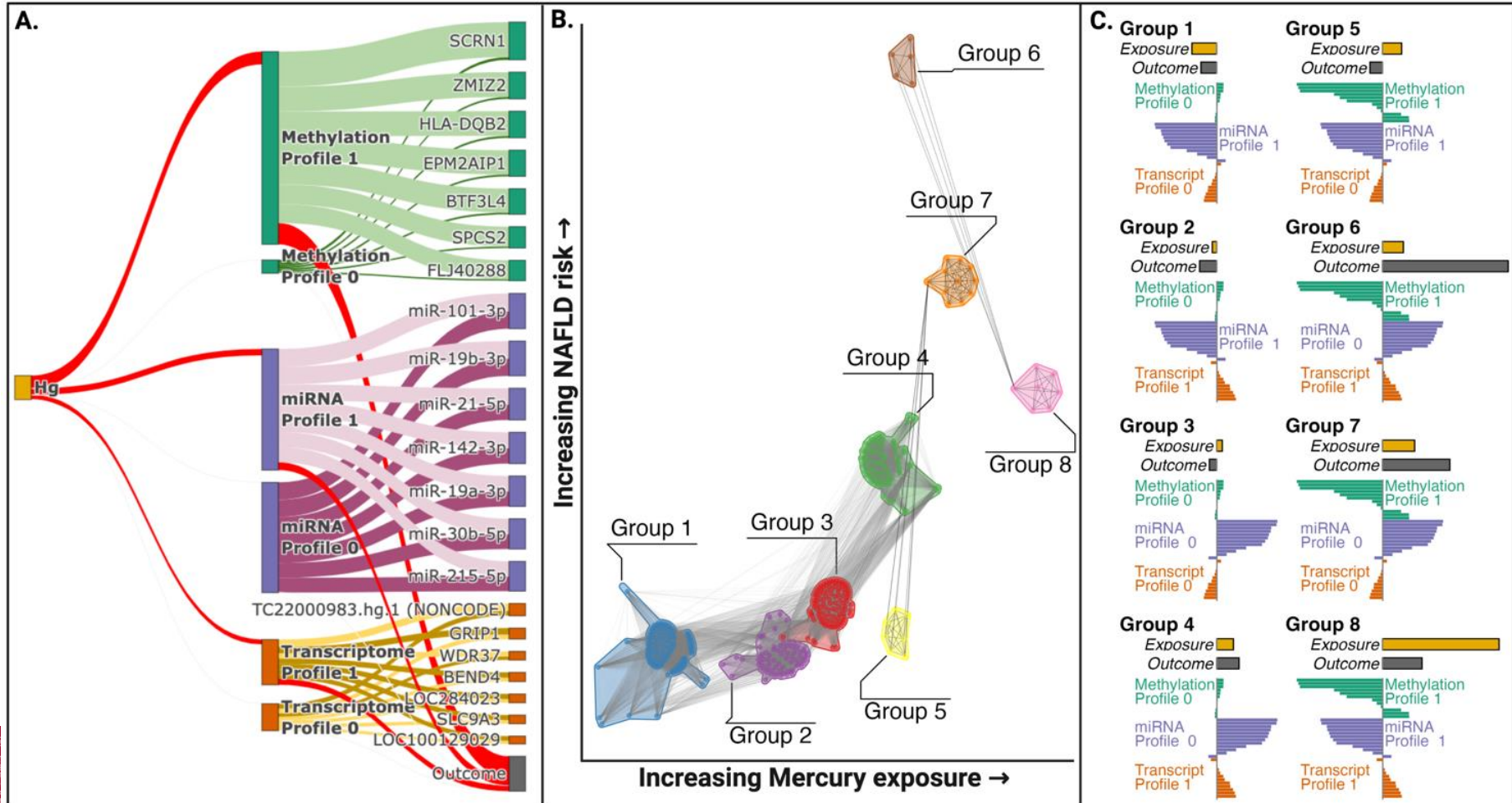


Joint Analysis, Interactions Between Omic



- Eight groups: defined by their exposure and outcome levels
- Here, each point represents an individual from our data. Lines connect individuals with similar omics profiles

Joint Analysis, Interactions Between Omic



Multimic Interaction

Legend:
G/E: Genetic or environmental exposure
D: Disease or health outcome
X: Latent Factors

Multimics Integration

Interaction Framework

High Dimensional

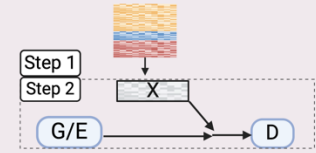
Molecular level traits

Interaction with Latent Factors

Step 1: Estimate X
 Step 2: Interaction

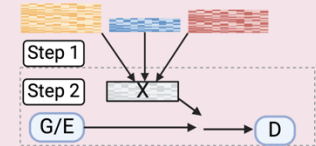
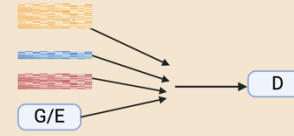
Early

Concatenate 'omics into single matrix



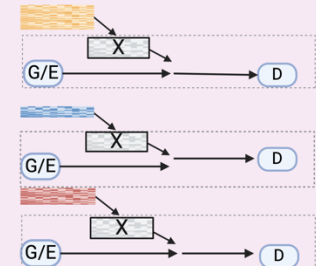
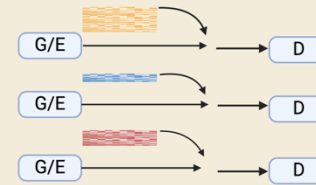
Intermediate

Combine 'omics through inference on joint model



Late

Individually model each 'omics layer



Scalable analytic framework for performing analysis with multiple 'omics datasets as effect modifiers of the relationship between genetics/environmental factors (G/E) and disease or other health outcomes (D).

Characterize Associations:

Human Studies



Discovery:

**Large Biobanks
Cohorts
Databases**



- AI and Analytic Methods that Leverage Prior Information
 - Pathways
 - Exposure Profiles and Mixtures
 - Omic Features that Reflect the Exposome

Characterize Associations:
Human Studies



Understand Biology:
Experimental Studies

- Pathways
- Exposure Profiles and Mixtures
- Omic Features that Reflect the Exposome
- Interventional Impact

Multi-Omics for Health and Disease (MOHD*)

Contributing NIH Institutes:

National Human Genome Research Institute (NHGRI)

National Cancer Institute (NCI)

National Institute of Environmental Health Sciences (NIEHS)

*MOHD: pronounced "mode"



Jesse Goodrich, PhD



Rob McConnell, PhD



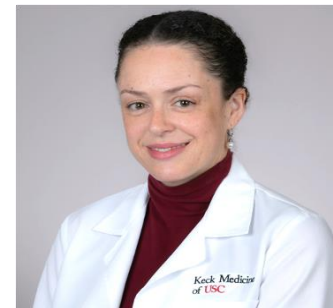
Lida Chatzi, PhD



Max Aung, PhD



Lucy Golden, PhD



Ana Maretti-Garcia, PhD



Matthew Salomon, PhD

Southern California Superfund Research and Training Program for PFAS Assessment, Remediation, and Prevention (ShARP)

PRIMED-Cancer

USC

- David Conti,
Chris Haiman,
Dan Stram

Stanford

- John Witte

Kaiser

- Lori Sakoda

Harvard:

- Mingyang Song

Fred Hutch

- Riki Peters, Charles Kooperberg

PRIMED Consortium

Polygenic Risk Methods in Diverse Populations

[Read more about us!](#)

Hawaii Cancer Center

- Loic Le Marchand, Lynne Wilkens

NCI

- Stephen Chanock, Sonja Berndts,
Pete Kraft



USCIMAGE

Integrative Methods of Analysis
for Genetic Epidemiology

U01CA261339



USCIMAGE

Integrative Methods of Analysis for Genetic Epidemiology



- MPIs: Jim Gauderman and Kim Siegmund
- PROJECT 1: INTEGRATION OF OMIC DATA TO ESTIMATE MEDIATION OR LATENT STRUCTURES:
 - David Conti, Josh Millstein, Nick Mancuso
- PROJECT 2: INTEGRATION OF OMIC DATA IN THE ANALYSIS OF GENE x ENVIRONMENT INTERACTION:
 - Jim Gauderman, Juan Pablo Lewinger, Eric Kawaguchi, Lu Zhang
- PROJECT 3: STATISTICAL METHODS FOR GENOME CHARACTERIZATION:
 - Paul Marjoram, Huaiyu Mi, Kim Siegmund, Kelly Street, Paul Thomas

Special Thanks to All the Students and Post-Docs

Environmental Genomics (T32 ES013678 NIEHS)

Thank You!