Testing for gene-environment (GxE) interaction using p-value aggregation identifies many GxE loci

Saurabh Mishra

Advisor: **Dr. Arunabha Majumdar**

Department of Mathematics, IIT Hyderabad

Outline

- Background
  - Genetics and inheritance models
  - Model misspecification
  - Gene–environment interactions
- Our approach
  - Objective
  - P-value aggregation framework
- Results
  - Simulation studies
  - UK Biobank applications
- Conclusions and Future Directions

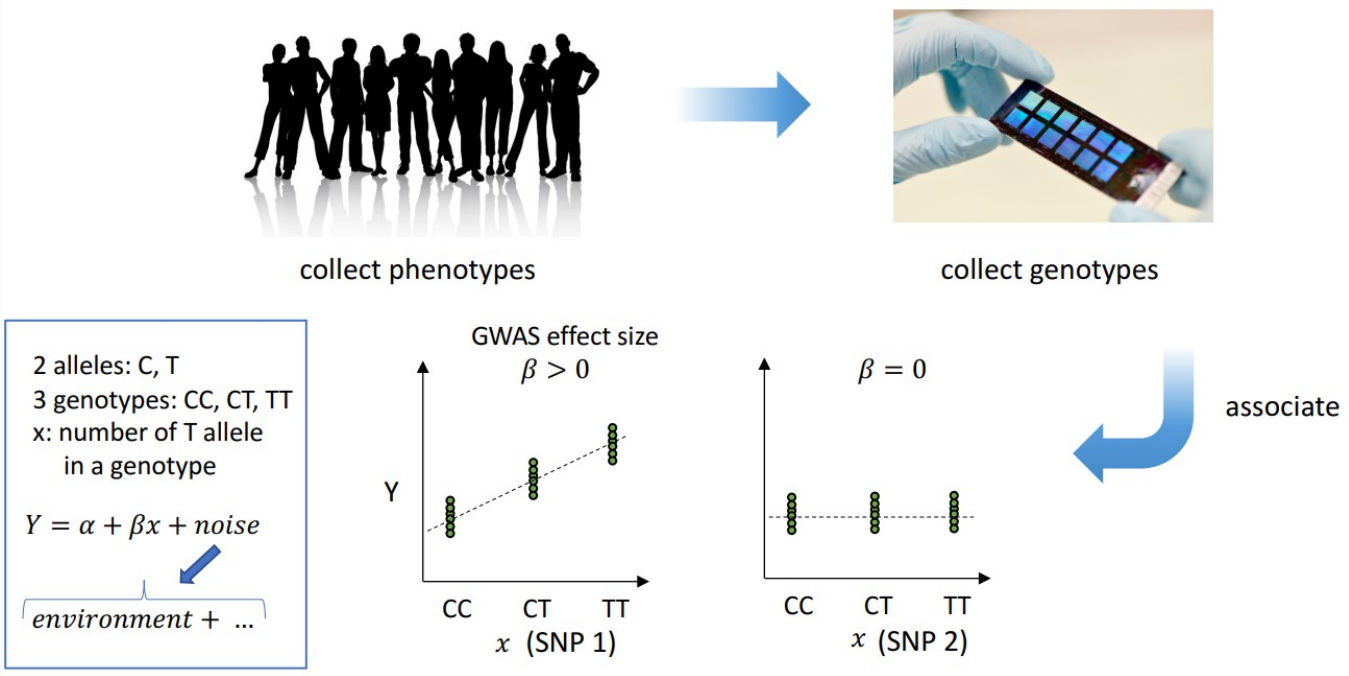
Basics of genetics

Phenotype = Genetics + Environment + · · ·

e.g., cholesterol level = genetics + diet + exercise + · · ·

Genome-wide association study (GWAS)

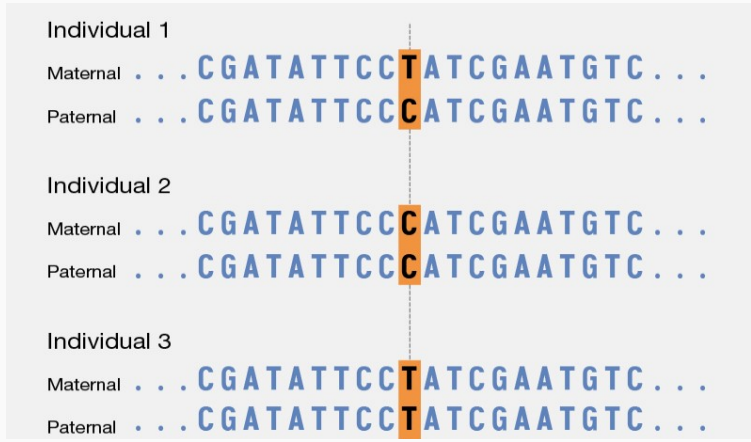
A statistical approach used to identify genomic variants that are statistically associated with a risk for a disease or a particular trait.

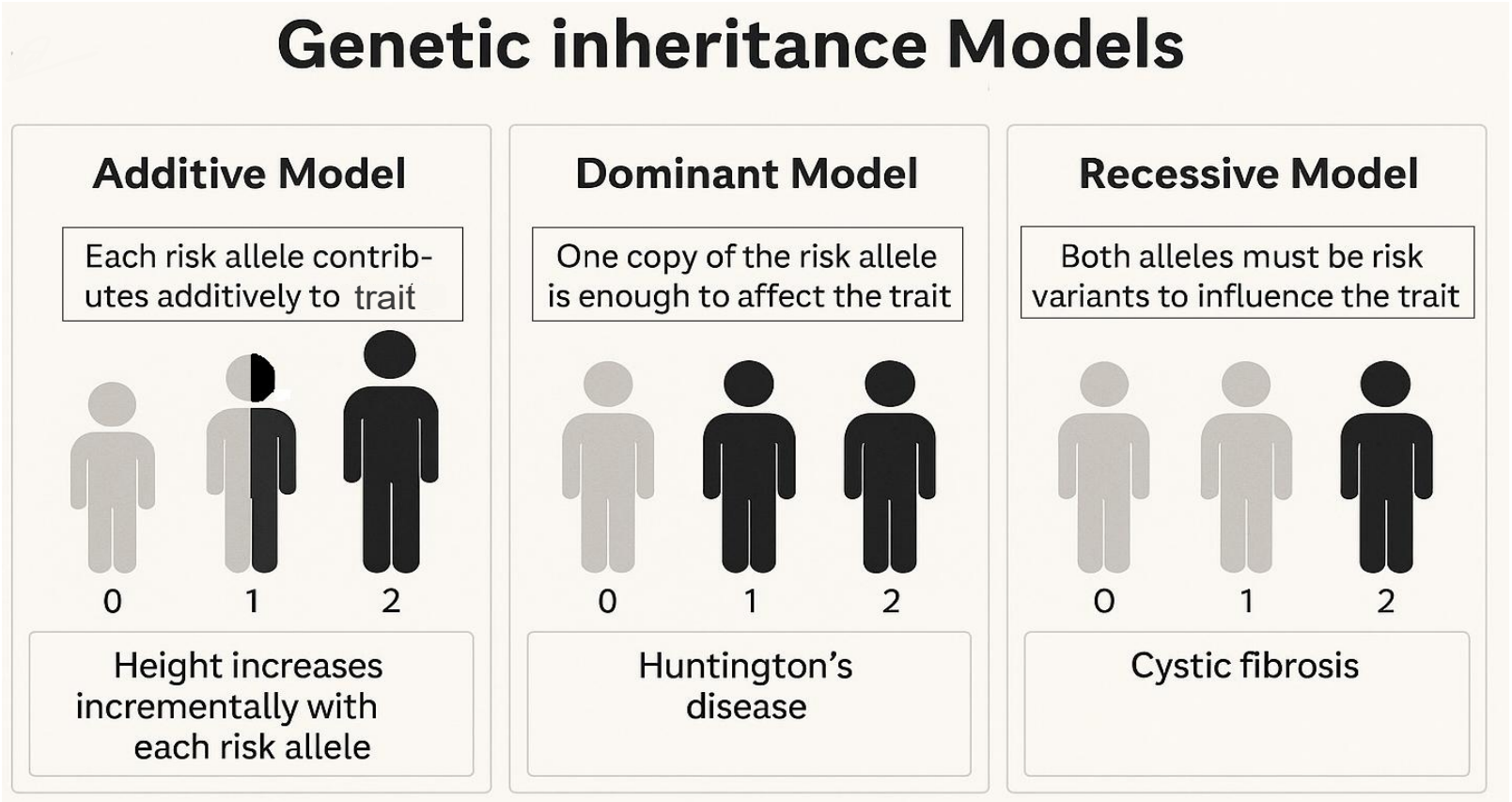


- T: risk allele, C: reference allele
- For **each SNP**, test:

$$H_0 : \beta = 0 \quad \text{vs} \quad H_1 : \beta \neq 0$$

- Apply multiple-testing correction





Model	CC	CT	TT
Additive (G_A)	0	1	2
Dominant (G_D)	0	1	1
Recessive (G_R)	0	0	1
Genotypic ($G_{\text{Het}}, G_{\text{Hom}}$)	(0,0)	(1,0)	(0,1)

Model	Model specification	Null hypothesis
Additive	$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{A,i} + \boldsymbol{\gamma}^\top \mathbf{C}_i$	$H_0 : \beta_1 = 0$
Dominant	$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{D,i} + \boldsymbol{\gamma}^\top \mathbf{C}_i$	$H_0 : \beta_1 = 0$
Recessive	$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{R,i} + \boldsymbol{\gamma}^\top \mathbf{C}_i$	$H_0 : \beta_1 = 0$
Genotypic (2df)*	$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{\text{Het},i} + \beta_2 G_{\text{Hom},i} + \boldsymbol{\gamma}^\top \mathbf{C}_i$	$H_0 : \beta_1 = \beta_2 = 0$

Y_i: phenotype; $g(\cdot)$ identity (continuous) or logit (binary); $\mathbf{C}_i$: covariates.

**A model-free test that evaluates SNP–phenotype association without assuming a specific inheritance pattern.*

Genetic model misspecification

The assumed genetic model does not match the true mode of inheritance.

Why does it occur?

- True inheritance patterns are unknown *a priori*.
- It is common to assume a fixed model (e.g., additive) for all SNPs.
- A dominant or recessive SNP modeled additively may lead to poor fit.

Why does it matter?

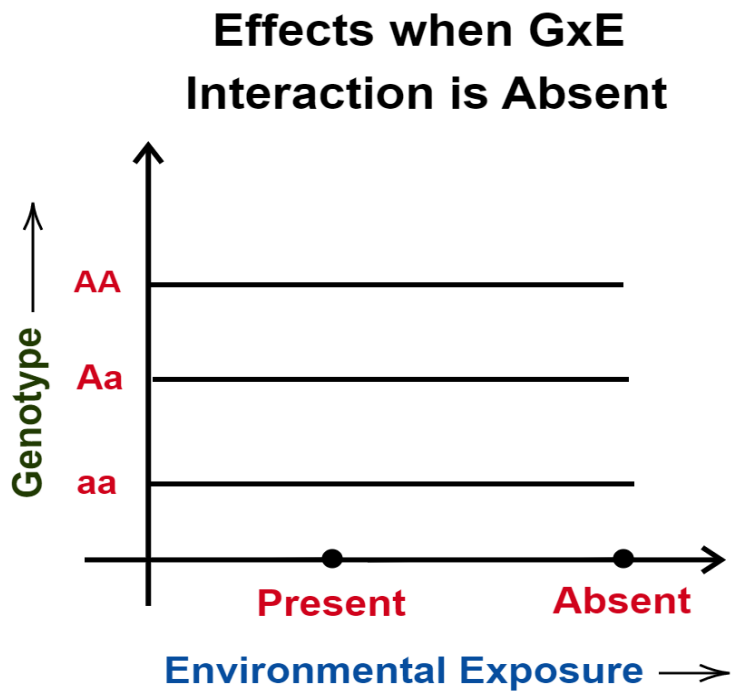
- Loss of statistical power.
- Biased effect estimates.

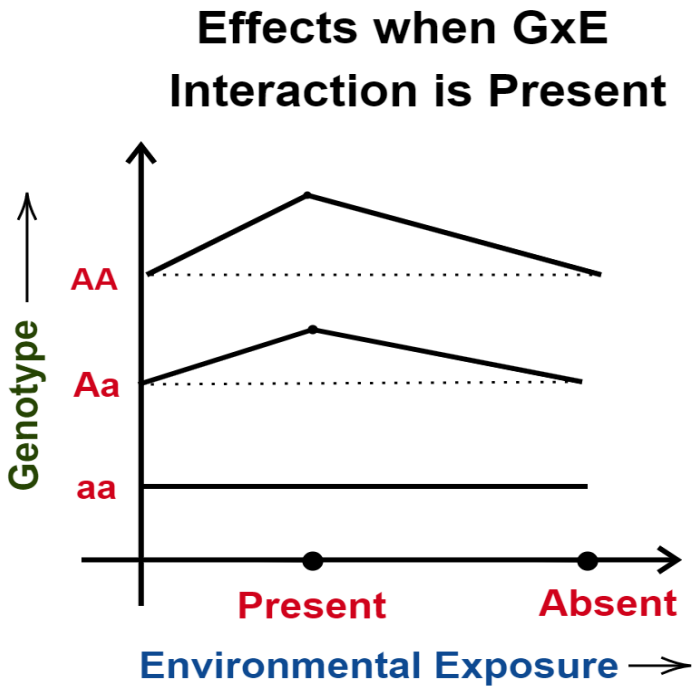
Gaye, Amadou, and Sharon K. Davis. *Genetic model misspecification in genetic association studies*. BMC Research Notes 10.1 (2017): 569.

Gene-Environment Interactions

Gene-Environment Interaction (Ottman,1996) [1]

"A different effect of an environmental exposure on disease risk in persons with different genotypes, and vice versa.





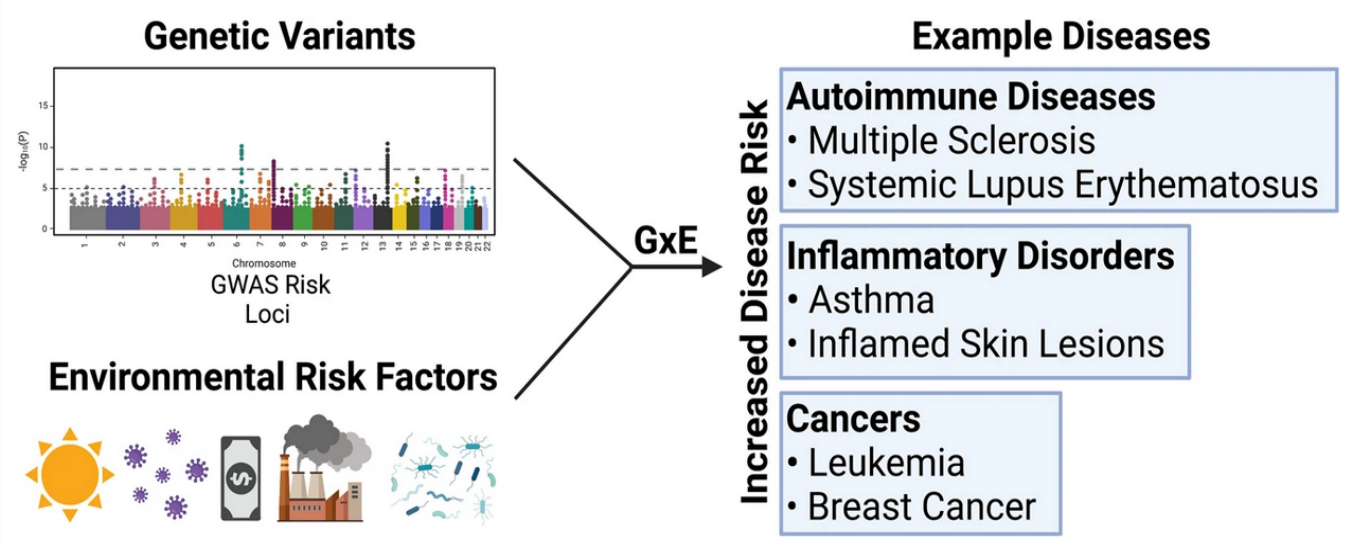
[1] Ottman, R. (1996). Gene–environment interaction: definitions and study design. Preventive medicine.

Real life examples

- Air pollutants (e.g., **$PM_{2.5}$**, NO_2, ozone) are known risk factors for asthma.

↑ Air pollution ⇒ Airway inflammation ⇒ Asthma risk

- The magnitude of pollution-induced asthma risk **differs by genotype**.



Standard GxE interaction models

GxE model (Additive / Dominant / Recessive):

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_i + \beta_2 E_i + \beta_3 (G_i \cdot E_i) + \boldsymbol{\gamma}^\top \mathbf{C}_i$$

Where $G_i \in \{G_{\text{add}}, G_{\text{dom}}, G_{\text{rec}}\}$

Hypothesis: G×E interaction (1 df)

$$H_0 : \beta_3 = 0 \quad \text{vs.} \quad H_A : \beta_3 \neq 0$$

Why Gene–Environment Interactions Are Difficult to Detect

- Interaction effect sizes are typically small.
- Statistical power is substantially lower than for main effects.
- **Beyond these challenges, misspecifying the genetic model further penalizes G×E tests, leading to additional loss of statistical power.**

2df GxE model[2]

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{\text{Het},i} + \beta_2 G_{\text{Hom},i} + \beta_E E_i + \beta_3 (G_{\text{Het},i} \cdot E_i) + \beta_4 (G_{\text{Hom},i} \cdot E_i) + \boldsymbol{\gamma}^\top \mathbf{C}_i$$

Hypothesis: G×E interaction (2 df)

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_A : \text{At least one of } \beta_3, \beta_4 \neq 0$$

[2] Moore, Camille M., Sean A. Jacobson, and Tasha E. Fingerlin. "Power and sample size calculations for genetic association studies in the presence of genetic model misspecification." Human heredity 84.6 (2020): 256-271.

2df GxE model[2]

$$g(\mathbb{E}[Y_i]) = \beta_0 + \beta_1 G_{\text{Het},i} + \beta_2 G_{\text{Hom},i} + \beta_E E_i + \beta_3 (G_{\text{Het},i} \cdot E_i) + \beta_4 (G_{\text{Hom},i} \cdot E_i) + \boldsymbol{\gamma}^\top \mathbf{C}_i$$

Hypothesis: G×E interaction (2 df)

$$H_0 : \beta_3 = \beta_4 = 0 \quad \text{vs.} \quad H_A : \text{At least one of } \beta_3, \beta_4 \neq 0$$

Strengths:

- Robust to model misspecification.
- Higher power when the true model is recessive, overdominant.
- Allows genotype-specific environmental effects.

Drawbacks:

- Lower power under true additive/ dominant inheritance.
- Requires larger sample sizes.

[2] Moore, Camille M., Sean A. Jacobson, and Tasha E. Fingerlin. "Power and sample size calculations for genetic association studies in the presence of genetic model misspecification." Human heredity 84.6 (2020): 256-271.

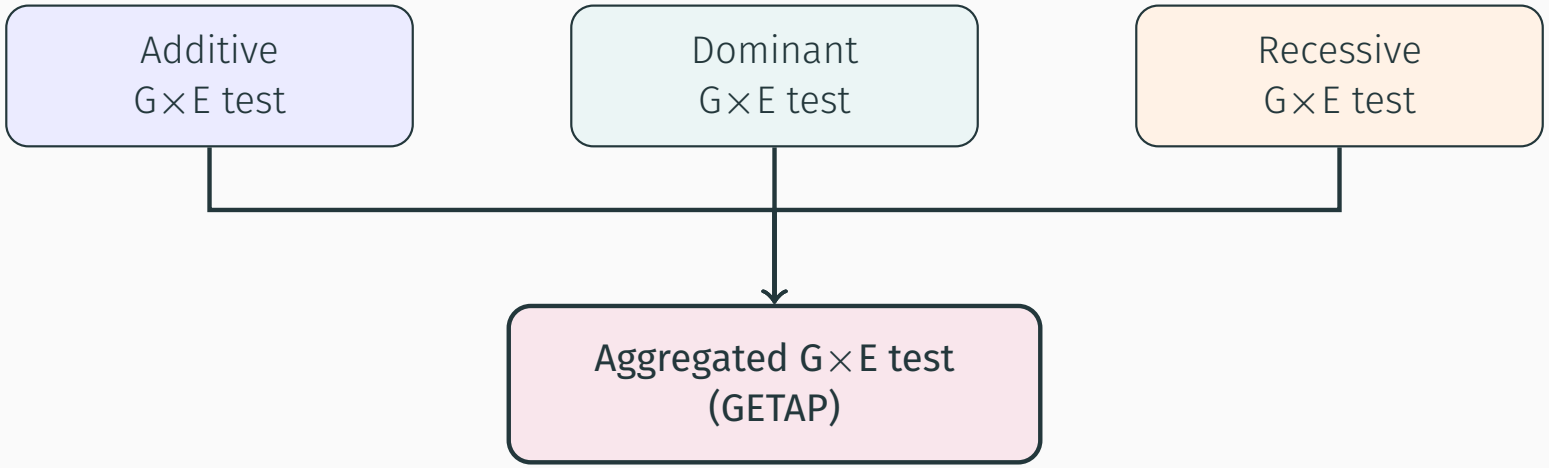
The problem we address

- G×E analyses typically assume a **single genetic model**, and **misspecification leads to substantial power loss**.
- The 2df test is **robust**, but **inefficient** under true additive or dominant models.
- Testing multiple single models separately introduces **multiple-testing burden**.

Our solution: aggregate evidence across genetic models

- Instead of selecting a single genetic model, we test G×E under multiple models
- Evidence across models is combined into a single omnibus test

This avoids committing to a possibly misspecified genetic model.



Aggregation performed using ACAT or HMP

P-value aggregation

P-value aggregation

Combining multiple p-values into a single test of a **global null hypothesis**.

Why is aggregation useful?

- **Power**: Detect weak but consistent signals across tests.
- **Robustness**: Protect against model misspecification.
- **Multiplicity**: Avoid repeated multiple-testing corrections.

Why is this important in genomics?

- Tests are often **dependent** (e.g., LD, correlated models).
- Standard combination methods may fail under dependence.

The Aggregated Cauchy Association Test (ACAT)[3]

Consider k hypothesis tests with p-values $p_1, p_2, \ldots, p_k$, where p_i is from the ith test.

Step 1: Transform individual P-values

$$C_i = \tan\left[(0.5 - p_i)\pi\right]$$

Step 2: ACAT test statistic

$$T_{\text{ACAT}} = \sum_{i=1}^{k} w_i C_i = \sum_{i=1}^{k} w_i \tan\left[(0.5 - p_i)\pi\right]$$

Here, $w_i \geq 0$ are user-specified weights (uniform weights $w_i = 1/k$ used in our analysis).

Step 3: Combined P-value (Cauchy tail approximation)

$$p_{\text{ACAT}} \approx 1 - \frac{1}{\pi} \arctan\left(\frac{T_{\text{ACAT}}}{\bar{w}}\right) \quad \text{where} \quad \bar{w} = \sum_{i=1}^{k} w_i$$

- *Very fast and analytically tractable, while remaining robust to* ***arbitrary dependency****.*

[3] Liu, Yaowu, et al. "ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies." The American Journal of Human Genetics 104.3 (2019): 410-421.

The Harmonic Mean P-value (HMP)[4]

Step 1: HMP Statistic

$$\overset{\circ}{p} = \frac{\sum_{i=1}^{k} w_i}{\sum_{i=1}^{k} \frac{w_i}{p_i}}, \qquad (\sum_{i=1}^{k} w_i = 1)$$

For equal weights: $\overset{\circ}{p} = \frac{k}{\sum_{i=1}^{k} \frac{1}{p_i}}$

Step 2: Final P-value (Two Methods)

- **Method A:** Compare $\overset{\circ}{p}$ to critical value α_k (approximate).
- **Method B:** Asymptotically exact p-value:

$$p_{\overset{\circ}{p}} = \int_{1/\overset{\circ}{p}}^{\infty} f_{\text{Landau}}(x \mid \mu, \sigma)\, dx$$

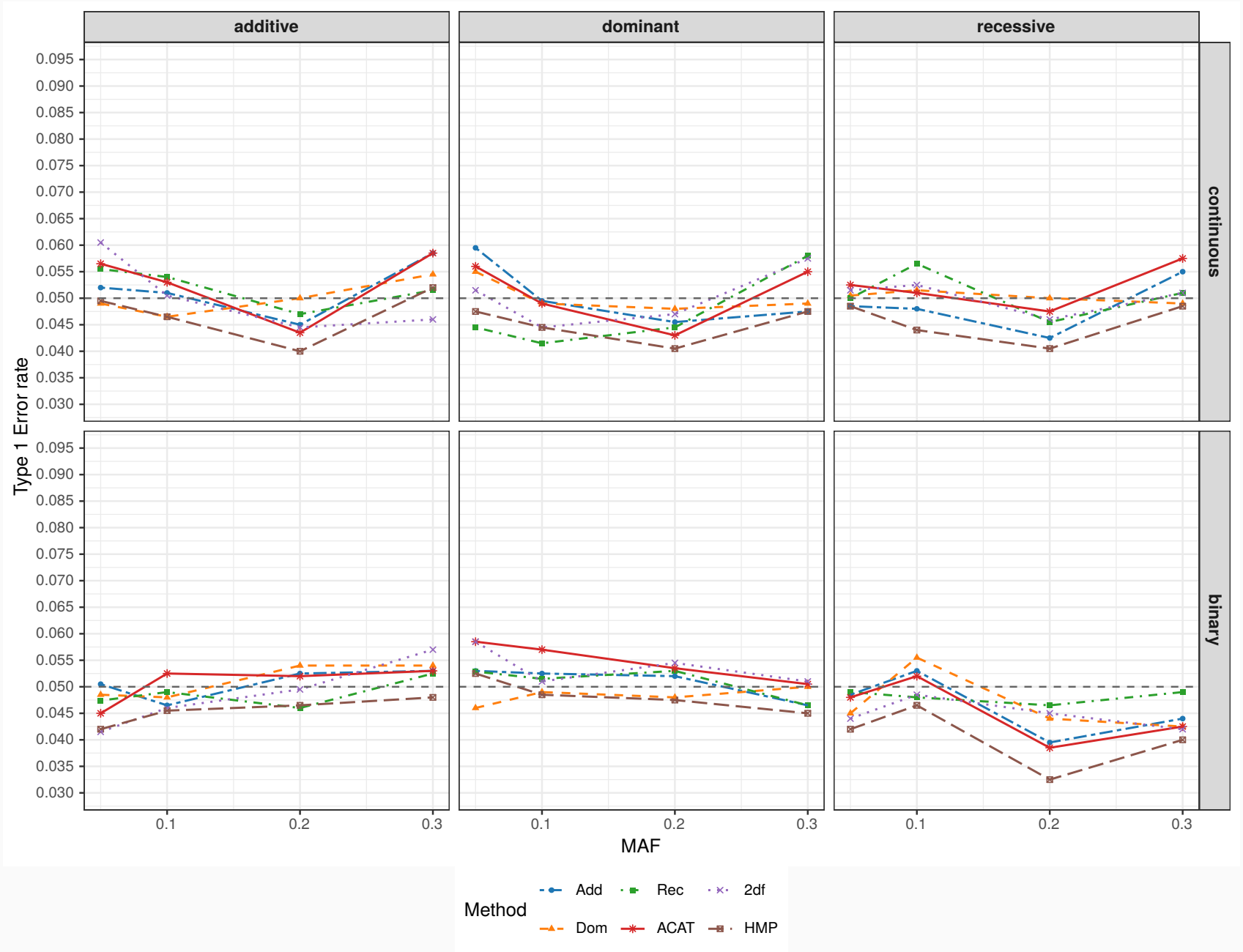
  where $\mu \approx \log(k) + 0.874$, $\sigma = \pi/2$

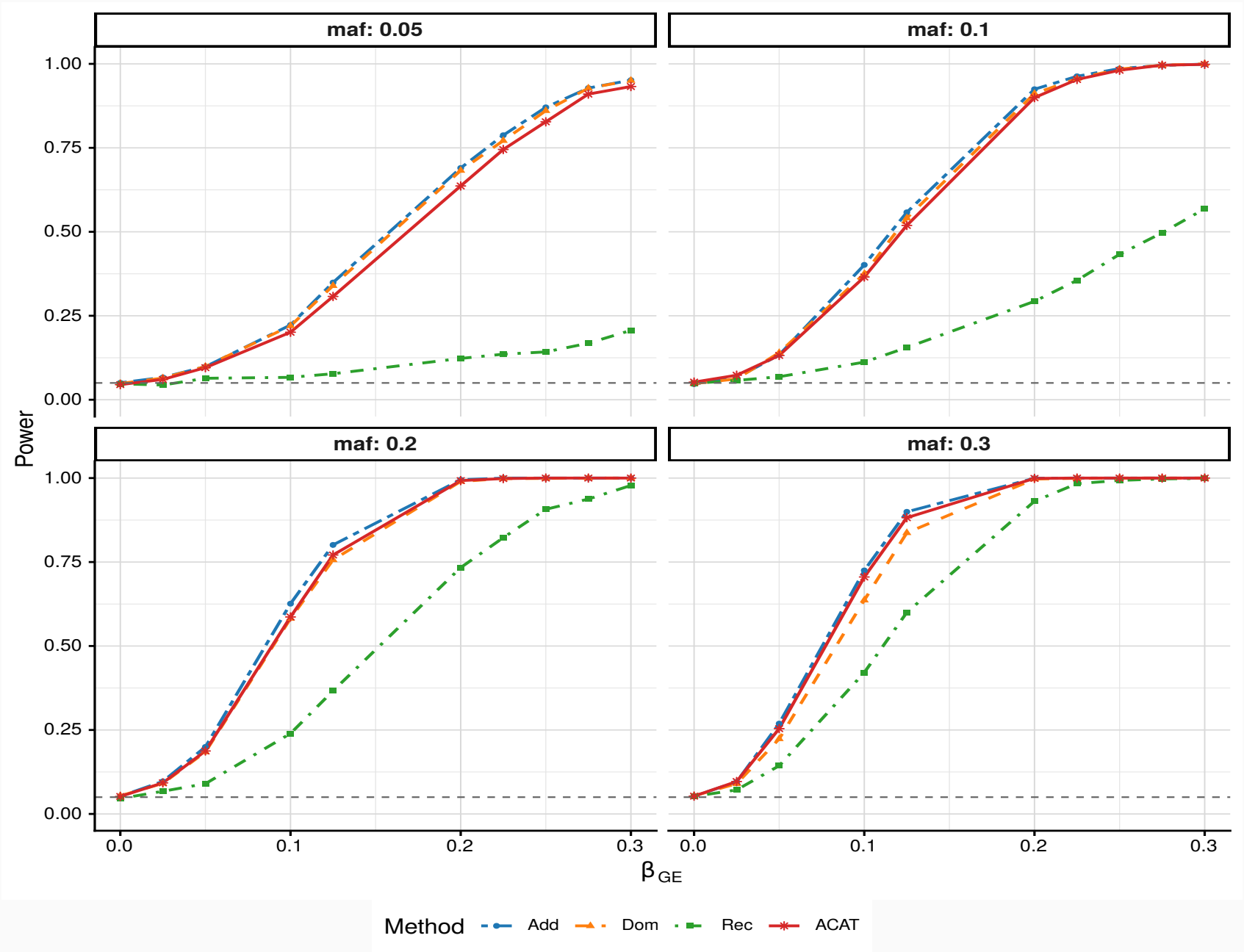
[4] Wilson, Daniel J. "The harmonic mean p-value for combining dependent tests." Proceedings of the National Academy of Sciences 116.4 (2019): 1195-1200.

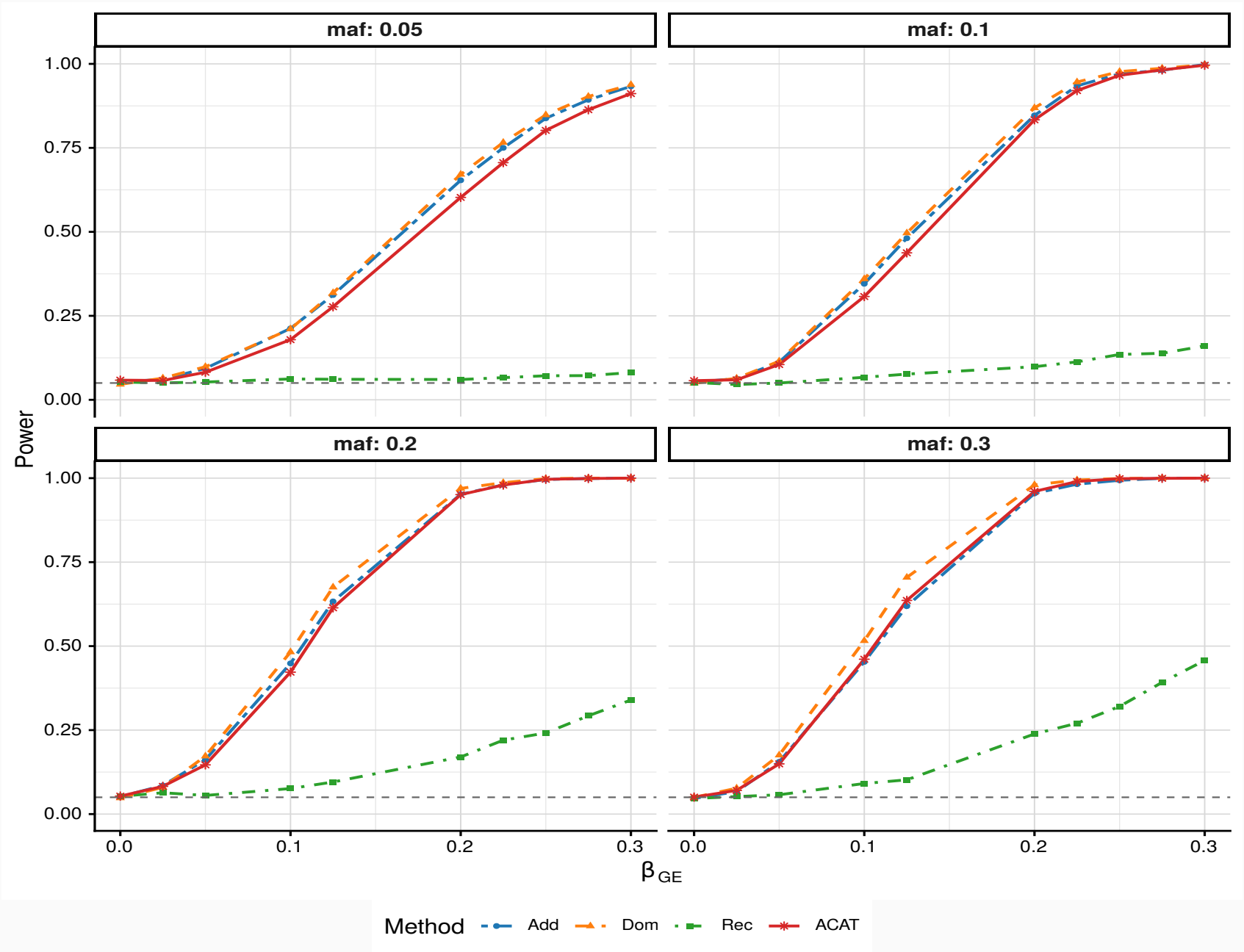
Simulation results

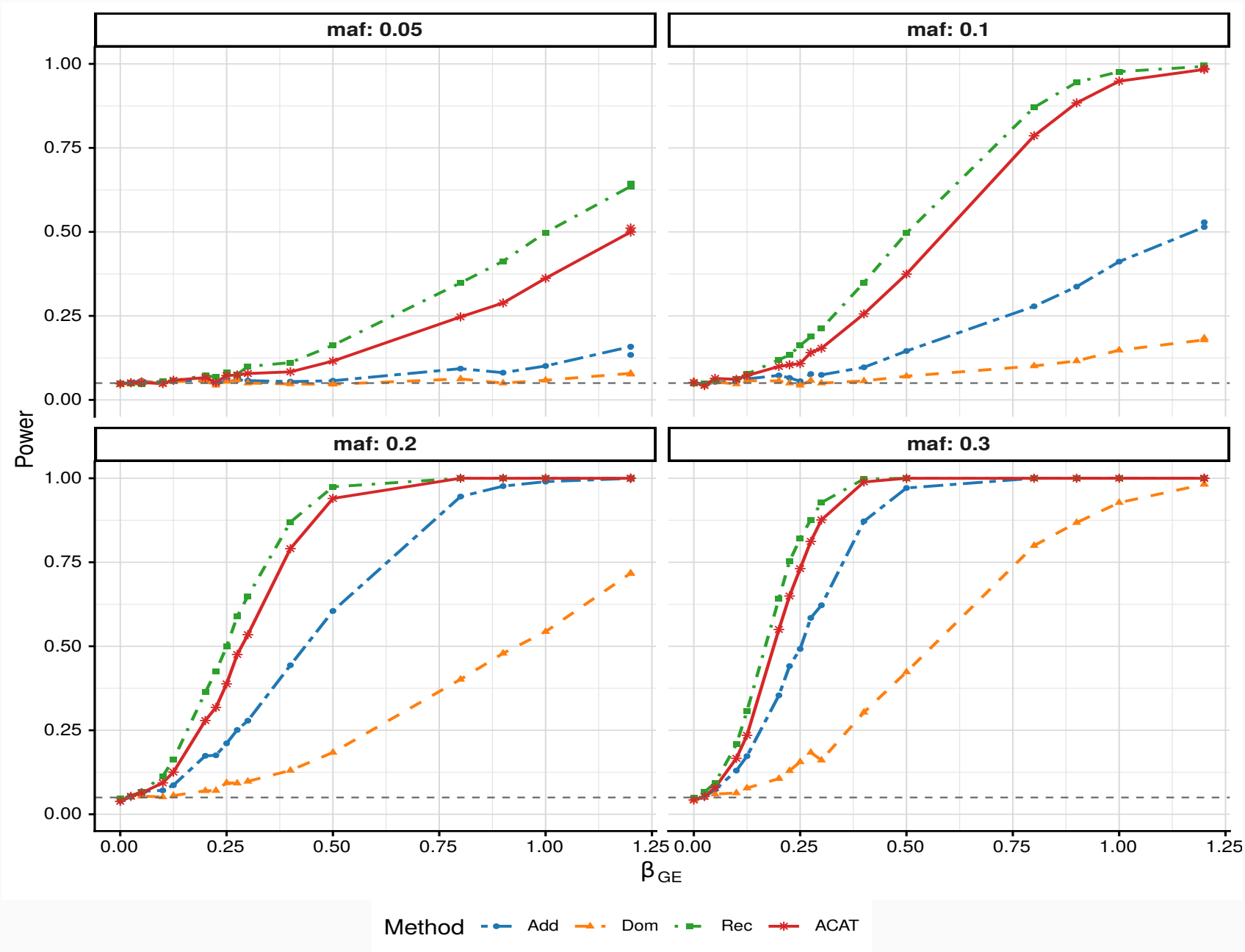
Simulation design

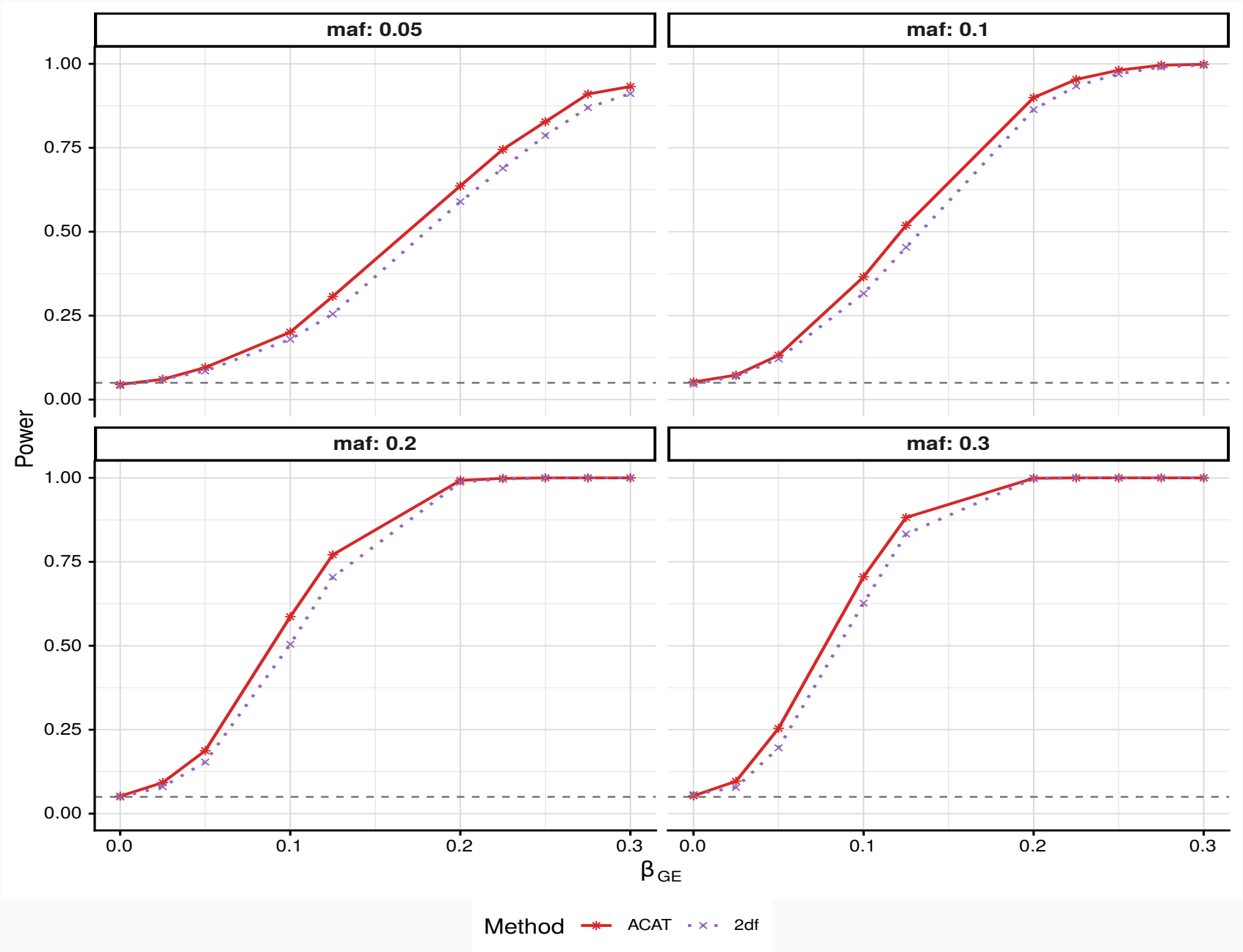
- **True genetic model:** Additive, dominant, recessive
- **Trait type:** Continuous and binary
- **Environmental exposure:** Continuous and binary
- **Sample size:** $n = 10{,}000$
- **Replicates:** 2,000 simulations
- **Minor allele frequency (MAF):** 0.05, 0.10, 0.20, 0.30
- **Significance level:** $\alpha = 0.05$

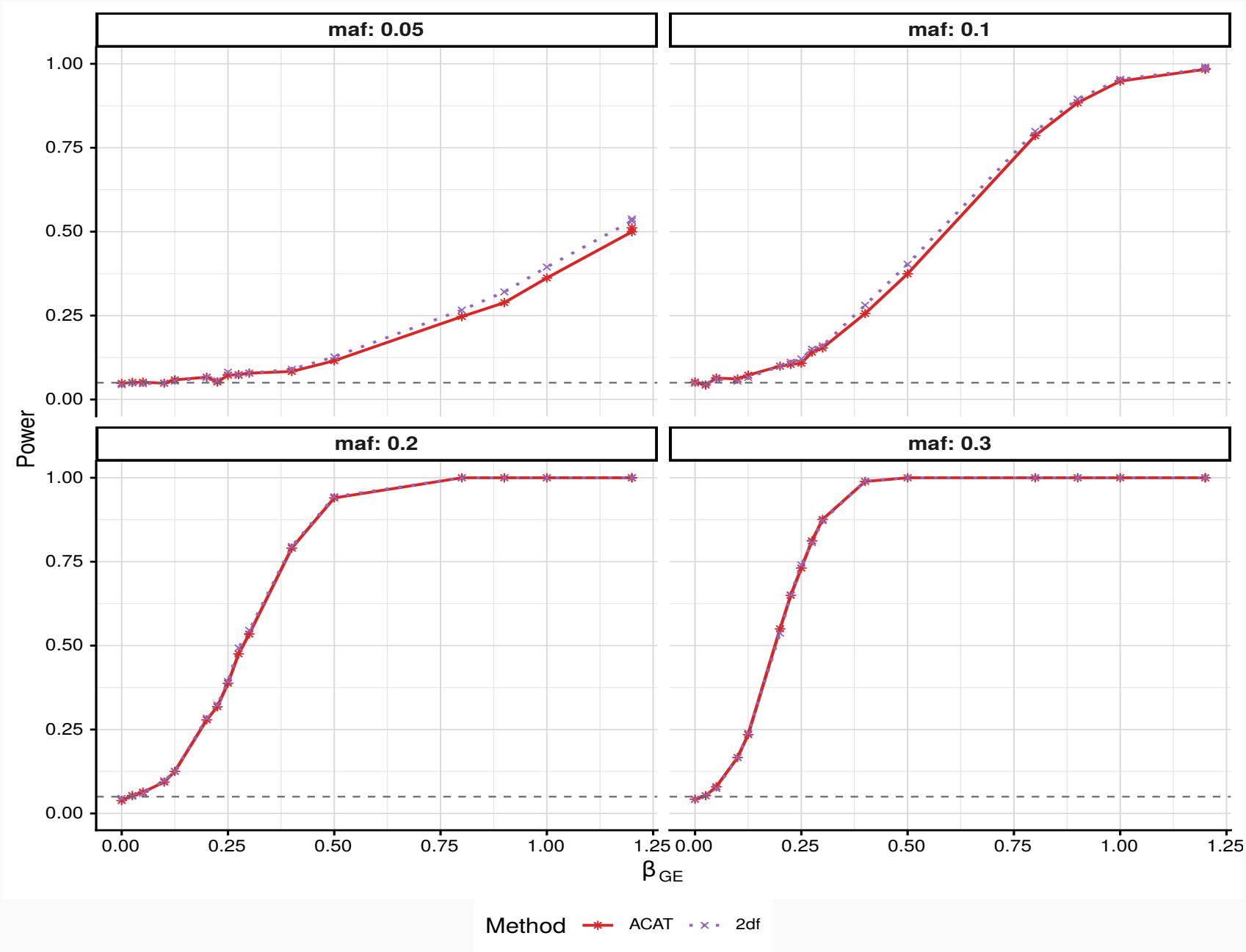

Estimated T1ER for continuous phenotype (n = 10000)
additive
dominant
recessive
continuous
binary
Type 1 Error rate
0.095
0.090
0.085
0.080
0.075
0.070
0.065
0.060
0.055
0.050
0.045
0.040
0.035
0.030
0.1
0.2
0.3
MAF
Method
Add
Rec
2df
Dom
ACAT
HMP

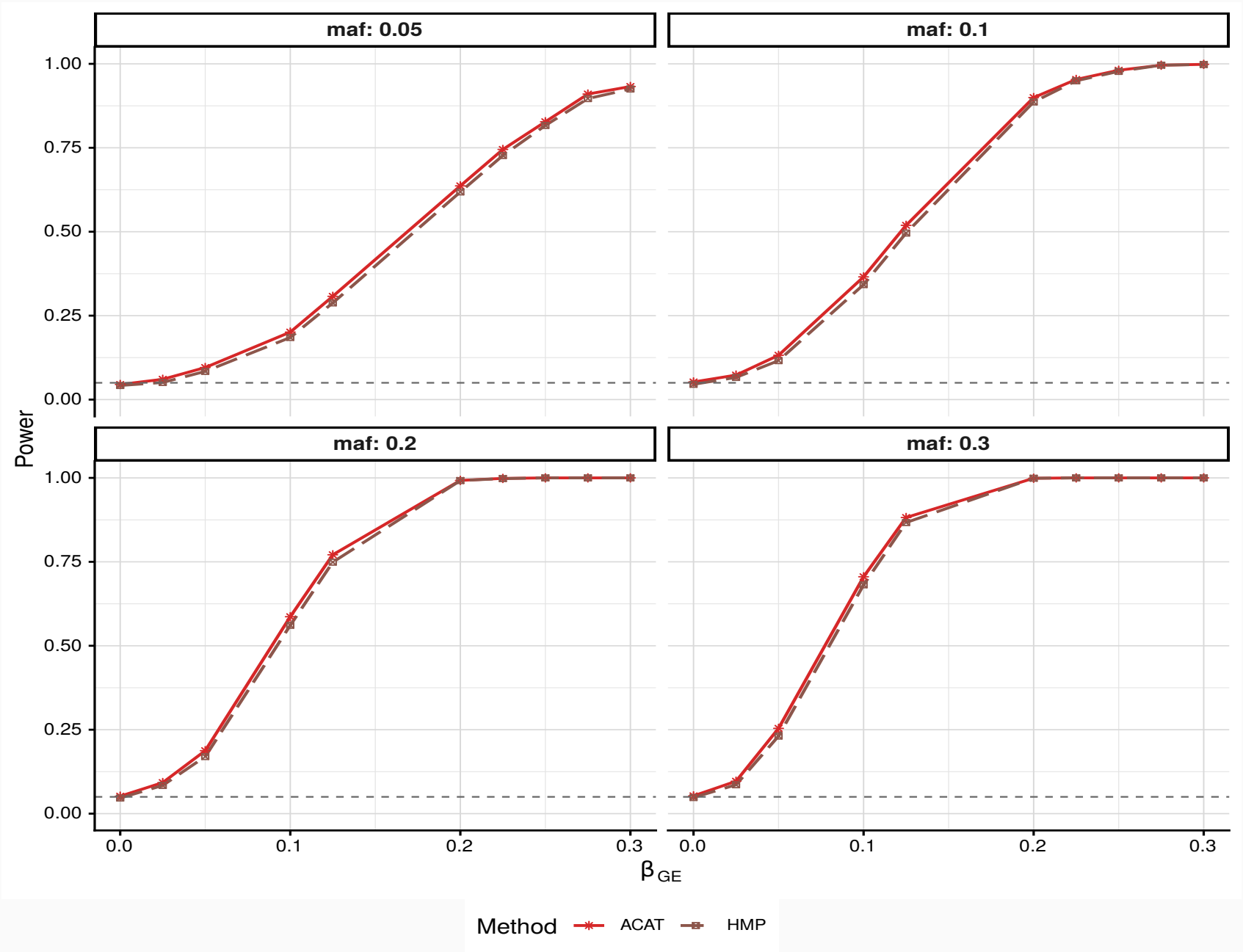

Power comparison (cont) | True model=additive | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
β GE
Method
Add
Dom
Rec
ACAT


Power comparison (cont) | True model=dominant | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
β GE
Method
Add
Dom
Rec
ACAT


Power comparison (cont) | True model=recessive | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
β GE
Method
Add
Dom
Rec
ACAT


Power comparison (cont) | True model=additive | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
β GE
Method
ACAT
2df


Power comparison (cont) | True model=recessive | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
1.25
β GE
Method
ACAT
2df


Power comparison (cont) | True model=additive | Env. type=binary | n=10000
maf: 0.05
maf: 0.1
maf: 0.2
maf: 0.3
Power
1.00
0.75
0.50
0.25
0.00
0.0
0.1
0.2
0.3
β GE
Method
ACAT
HMP

Real data application

Real data application: UK Biobank

UK Biobank

- ∼500,000 participants
- Unrelated White British individuals
- Autosomal SNPs after QC (∼600K)

Phenotypes and environments

- Continuous and binary traits
- Lifestyle and behavioral exposures

Phenotype + Genotype + Environment data

↓

Additive / Dominant / Recessive
G×E tests

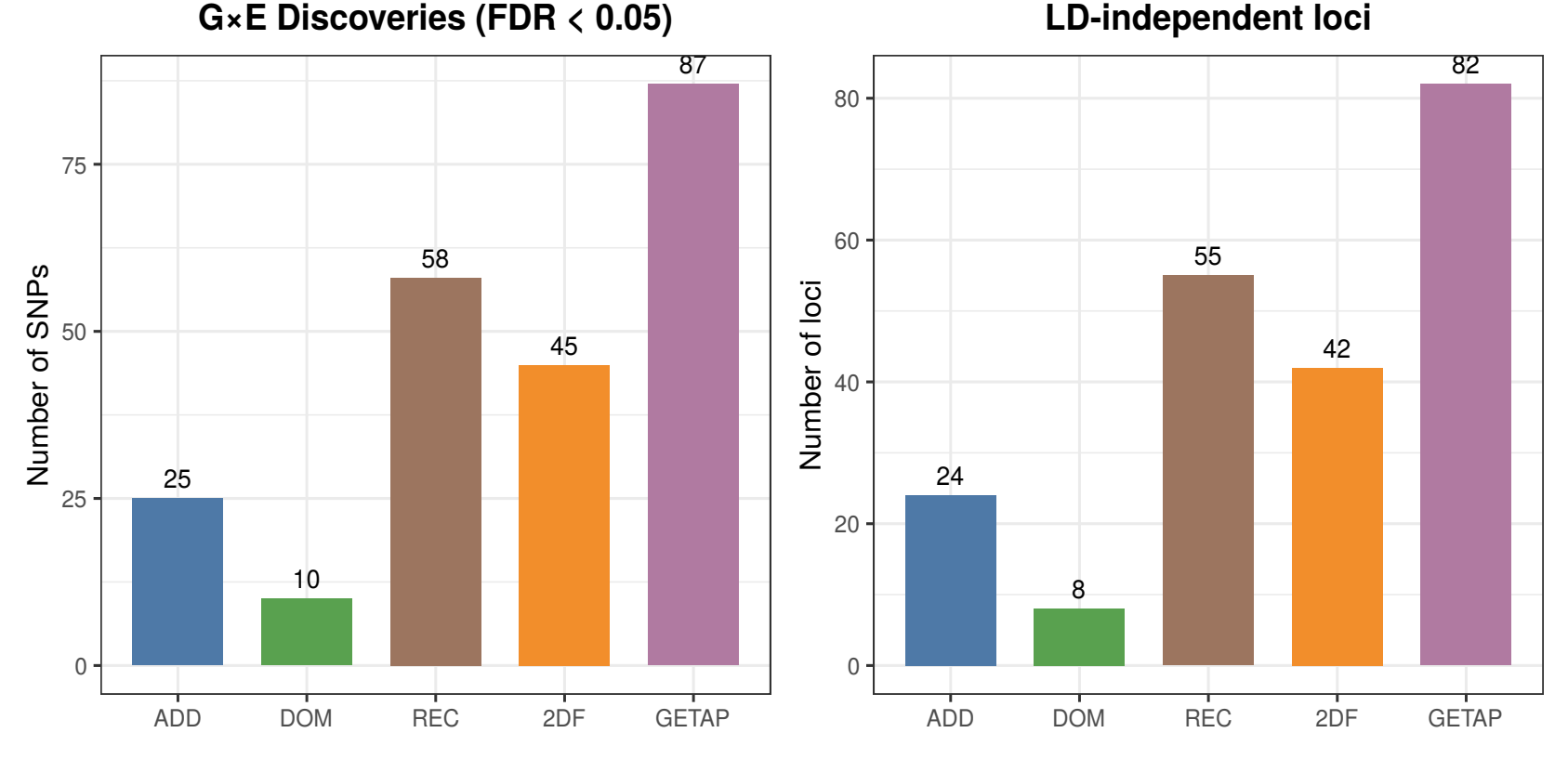
↓

GETAP (ACAT)
P-value aggregation

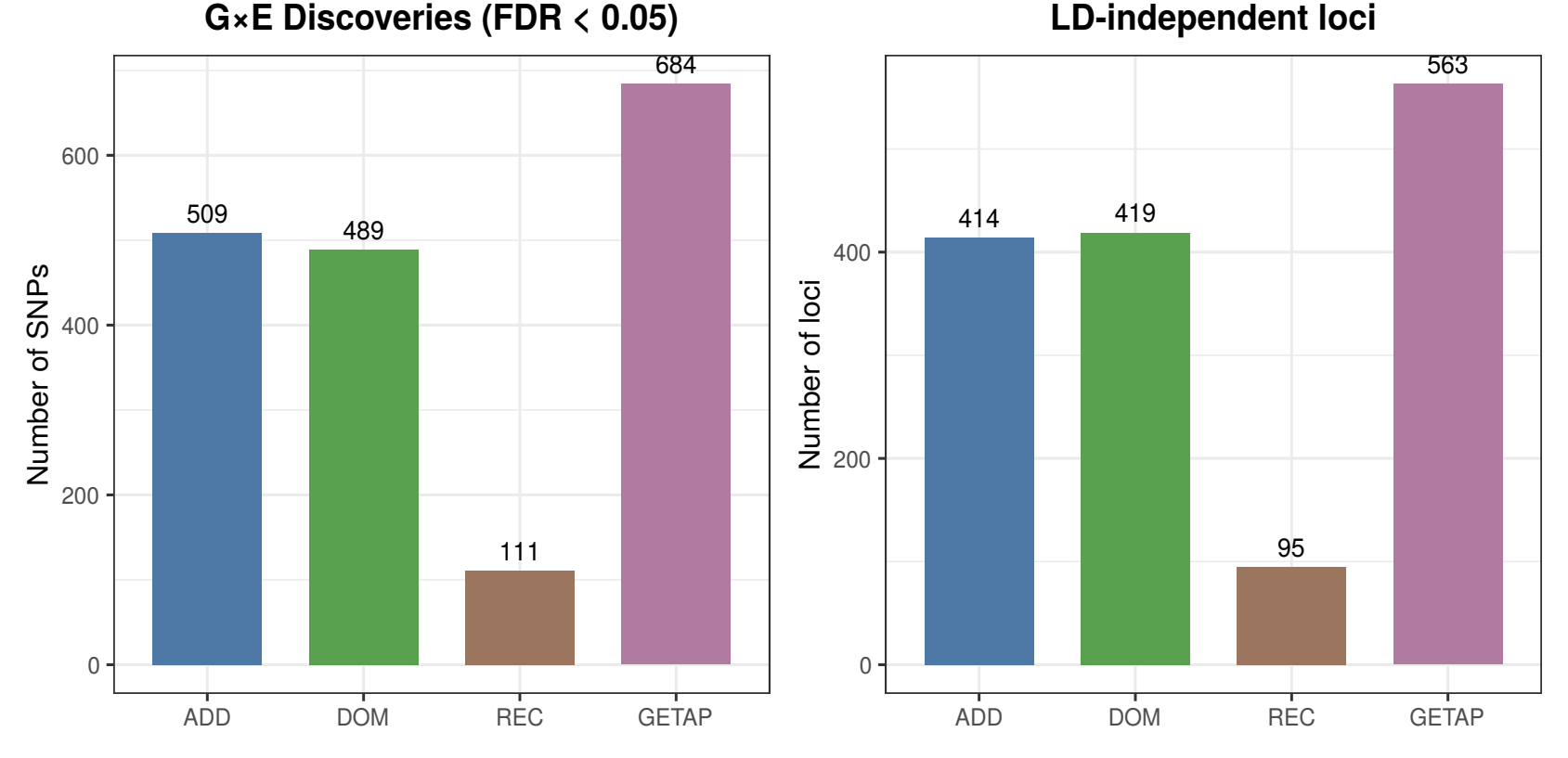
↓

Significant G×E loci

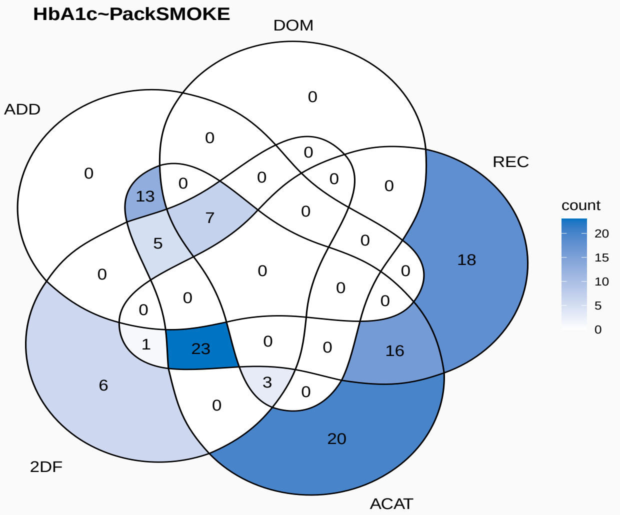
Phenotype: HbA1c, Env.: Pack years of smoking



Phenotype: Type 2 Diabetes, Env.: Sleep duration



Overlap structure

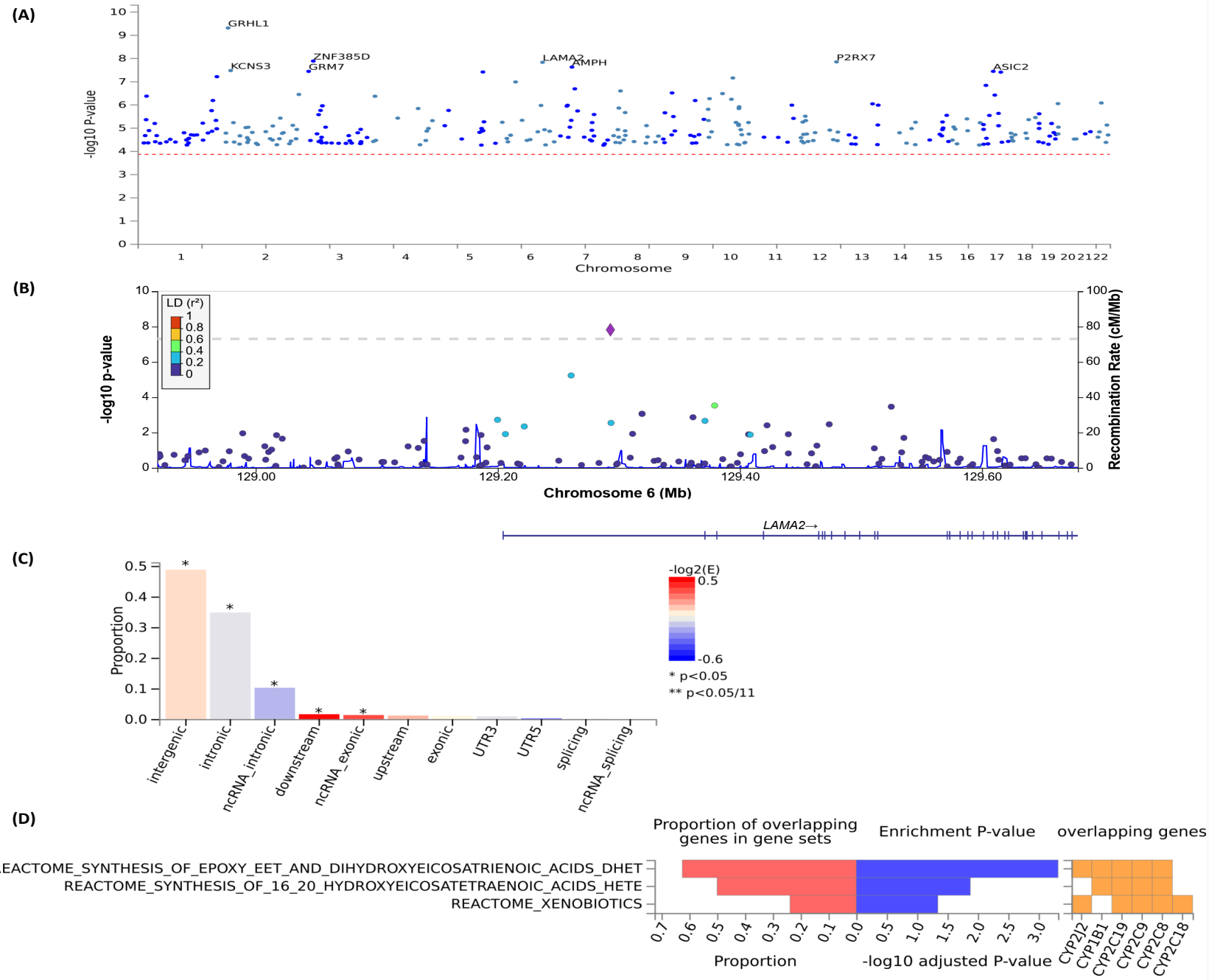


Genome wide significant SNPs

Table 1: Selected SNPs showing genome-wide significant G×E signals.

Phenotype	Environment	SNP	Chr	P_{ADD}	P_{DOM}	P_{REC}	P_{GETAP}	P_{2DF}
HbA1c	Pack-years (Smoking)	rs407423	8	2.7×10^{-2}	1.7×10^{-1}	5.1×10^{-10}	1.5×10^{-9}	3.7×10^{-9}
FEV_1/FVC	Pack-years (Smoking)	rs13180	15	1.0×10^{-8}	2.9×10^{-6}	2.8×10^{-6}	3.0×10^{-8}	5.8×10^{-8}
T2D	Sleep duration	rs2801198	1	1.4×10^{-3}	6.7×10^{-2}	3.0×10^{-8}	8.9×10^{-8}	–

Biological significance: T2D


(A)
GRHL1
KCNS3
ZNF385D
GRM7
LAMA2
AMPH
P2RX7
ASIC2
-log10 P-value
Chromosome
(B)
LD (r²)
1
0.8
0.6
0.4
0.2
0
-log10 p-value
Recombination Rate (cM/Mb)
129.00
129.20
129.40
129.60
Chromosome 6 (Mb)
LAMA2→
(C)
Proportion
-log2(E)
0.5
-0.6
* p<0.05
** p<0.05/11
intergenic
intronic
ncRNA_intronic
downstream
ncRNA_exonic
upstream
exonic
UTR3
UTR5
splicing
ncRNA_splicing
(D)
Proportion of overlapping genes in gene sets
Enrichment P-value
overlapping genes
REACTOME_SYNTHESIS_OF_EPOXY_EET_AND_DIHYDROXYEICOSATRIENOIC_ACIDS_DHET
REACTOME_SYNTHESIS_OF_16_20_HYDROXYEICOSATETRAENOIC_ACIDS_HETE
REACTOME_XENOBIOTICS
Proportion
-log10 adjusted P-value
CYP2J2
CYP1B1
CYP2C19
CYP2C9
CYP2C8
CYP2C18

Summary

- Across scenarios, the aggregated test behaves close to the best-performing model without knowing that model in advance.
- GETAP provides robust and scalable inference under genetic model uncertainty.
- P-value aggregation recovers G×E loci missed by single-model tests.

Previous Work

- Mishra, S. and Majumdar, A., 2025. A Multi-Phenotype Approach to Joint Testing of Main Genetic and Gene-Environment Interaction Effects. *Statistics in Medicine*, 44(20-22), p.e70253.

- **R package:** *MvGGE* (implements the above method).

  Available open source on GitHub: *`https://github.com/SauMStats/MvGGE`*

Way forward

- **Ancestry-specific G×E analysis:**
  Apply our multivariate G×E framework (MvGGE) to UK Biobank South Asian ancestry data to study population-specific interaction effects.

- **Multiple-environment G×E methods:**
  Extend current models to jointly incorporate multiple environmental exposures, moving beyond single-environment interaction analyses.

- **Software dissemination:**
  Publish the existing *MvGGE* R package on Bioconductor and expand it into a scalable, user-friendly tool for large-scale multivariate G×E analysis.

References I

- Zeng, Ping, et al. "Aggregating multiple expression prediction models improves the power of transcriptome-wide association studies." Human Molecular Genetics 30.10 (2021): 939-951.
- Gaye, Amadou, and Sharon K. Davis. "Genetic model misspecification in genetic association studies." BMC research notes 10.1 (2017): 569.
- Moore, Camille M., Sean A. Jacobson, and Tasha E. Fingerlin. "Power and sample size calculations for genetic association studies in the presence of genetic model misspecification." Human heredity 84.6 (2020): 256-271.
- Haas, Cameron B., et al. "Interactions between folate intake and genetic predictors of gene expression levels associated with colorectal cancer risk." Scientific Reports 12.1 (2022): 18852.
- Evans, Luke M., et al. "Transcriptome-wide gene-gene interaction associations elucidate pathways and functional enrichment of complex traits." PLoS Genetics 19.5 (2023): e1010693.
- Gao, Guimin, et al. "A joint transcriptome-wide association study across multiple tissues identifies candidate breast cancer susceptibility genes." The American Journal of Human Genetics 110.6 (2023): 950-962.

THANK YOU

Backup Slides

Wald and Likelihood Ratio Tests (2 df)

1. **Wald Test:**

$$W = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix}^\top \left[\widehat{\text{Var}} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} \right]^{-1} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} \sim \chi^2_2$$

- Uses estimated coefficients and their variance-covariance matrix.

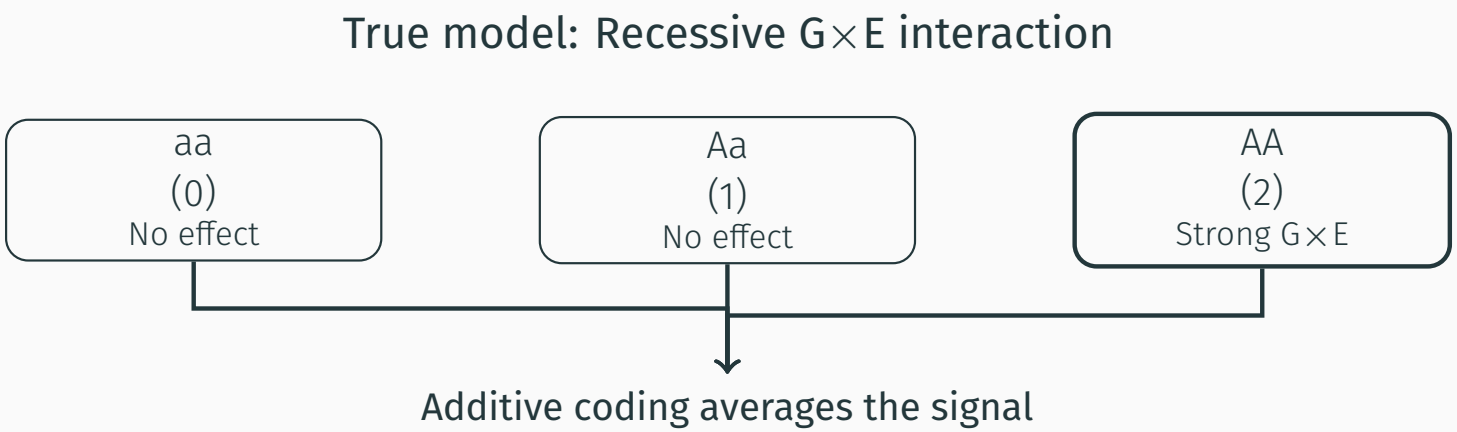
2. **Likelihood Ratio Test (LRT):**

$$\Lambda = -2 \left[\ell(\hat{\theta}_0) - \ell(\hat{\theta}_1) \right] \sim \chi^2_2$$

- $\ell(\hat{\theta}_0)$: log-likelihood under null model (no genotype effect)
- $\ell(\hat{\theta}_1)$: log-likelihood under full model (with β_1, β_2)
- Tests improvement in model fit when including genotype indicators

Genetic model misspecification causes power loss

- Genetic effects may be additive, dominant, or recessive
- Incorrect genotype coding dilutes interaction signals
- This problem is amplified for rare variants and G×E effects



Why BH FDR in GxE Analysis?

- Balances power and error control in high-dimensional genomic data
- Controls expected proportion of false positives ($q < 0.05$), unlike conservative FWER methods
- Maintains robust control under moderate LD dependence typical in post-QC SNPs
- Standard practice in large biobank GxE studies (e.g., UK Biobank)
- Avoids Bonferroni's type II error inflation and BY's unnecessary power loss

Comparison of Multiple Testing Corrections

Multiple testing is an analysis in which multiple independent hypotheses are tested. The overall combined probability of making a type I error will increase.

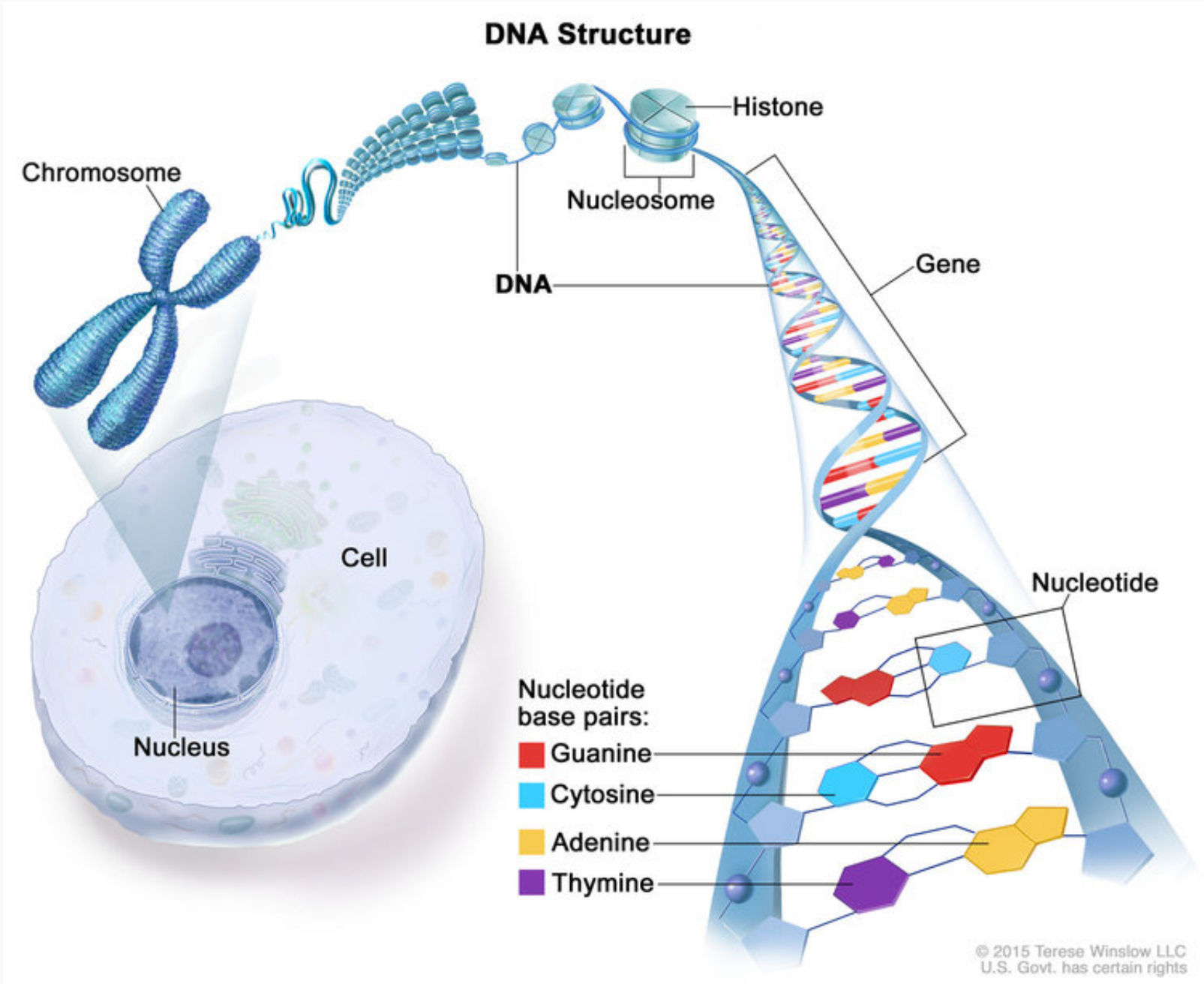
Method	Controls	Suitable for	Stringency	Ideal Usage Scenario
Bonferroni	FWER	Small/independent tests	Very conservative	High-stakes studies where false positives are critical
BH	FDR	Large/high-dimensional data	Less conservative	Exploratory studies with large test numbers (e.g., genomics)
BY	FDR under dependence	Correlated/multiple related tests	Moderate	Studies with known dependencies among tests (e.g., pathway analyses)

The Benjamini-Hochberg (BH) procedure ranks the p-values $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ in ascending order and finds the largest k such that:

$$p_{(k)} \leq \frac{k}{m} \cdot \alpha$$

where m is the total number of tests and α is the desired false discovery rate. All p-values $p_{(1)}, \ldots, p_{(k)}$ are considered significant.

Cell → Nucleus → Chromosome → DNA



- **SNP:** Single nucleotide polymorphism (SNP, pronounced ”snip”) is a genomic variant at a single base position in the DNA.

- **Allele:** One of two versions of DNA sequence at a given genomic location. Example: C, T.

- **Genotype:** The overall genetic makeup of an individual.
  - The 3 genotypes for alleles C and T will be CC, CT, and TT.
  - Genotype values vary from one person to another.

- **Phenotype:** A characteristic of an individual which can be observed/ measured.