A Sparse Marginal Epistasis Test

Julian Stamp
Center for Computational Molecular Biology
Brown University

CCMB

BROWN

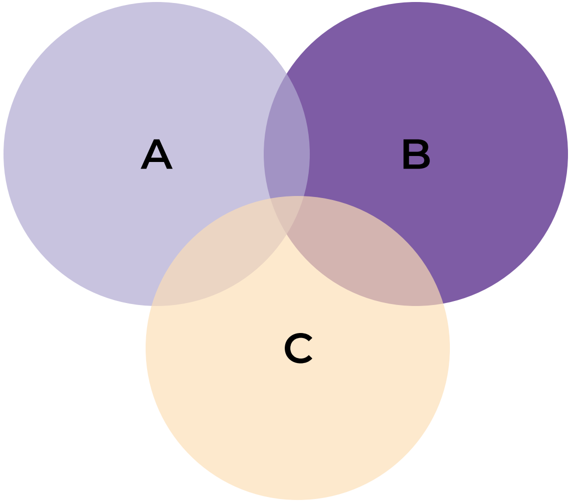
A Sparse Marginal Epistasis Test

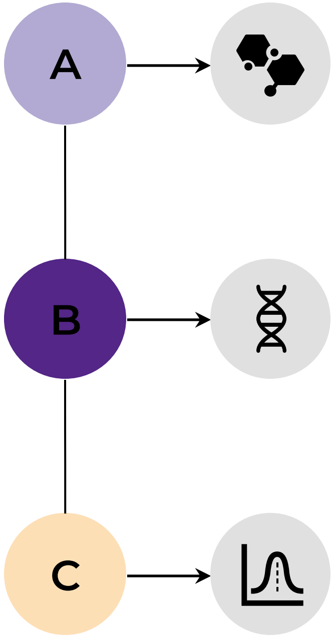
Julian Stamp
Center for Computational Molecular Biology
Brown University

Talk Outline

1. Background & Motivation
2. Method Overview
3. Application to Data
4. Takeaways

What is epistasis?





Functional Epistasis

Molecular interactions, e.g. protein interactions in the same pathway or complex

Compositional Epistasis

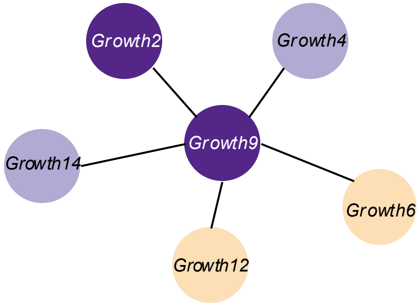
Allelic effect depends on the genetic context, e.g. genotype at another locus

Statistical Epistasis

Deviation from the additive allelic effect model

Phillips (2008) Nat Rev Genet.

Is there epistasis in complex traits?



Experiment

- Chicken selected for body size (40 gen.)
- Large phenotypic difference (6x between lines!)
- Expectation: large genetic effect

Epistasis can obscure gene mapping in complex traits

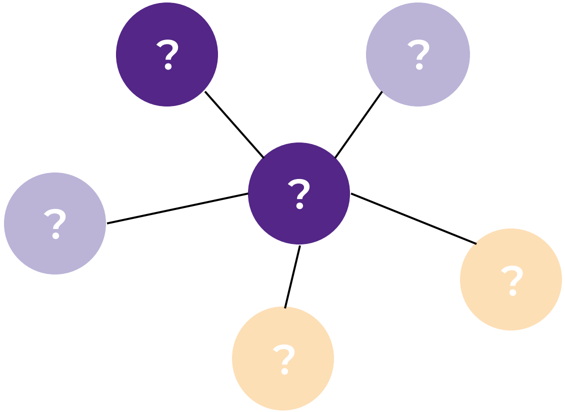
Gene Mapping

- Growth9 only minor additive effect
- Epistasis between 5 loci and Growth9 explained 45% of line difference
- Epistasis explained all of Growth9 effect

Carlborg et al. (2006) *Nat. Genet.*

Is there statistical epistasis in human traits?

- Studied in natural populations – only **residual variation** among individuals observed[1]
- So far, **little evidence** for statistical epistasis in human traits[2]
- Computational methods to detect epistasis are **underpowered** or computationally **intractable**[3] for human data



Answering this question requires a method that scales to **biobank data**

1 Phillips (2008), Nat Rev Genet.
2 Pattillo Smith et al. (2024), *eLife*
3 Crawford et al. (2017), *PLOS Genetics*

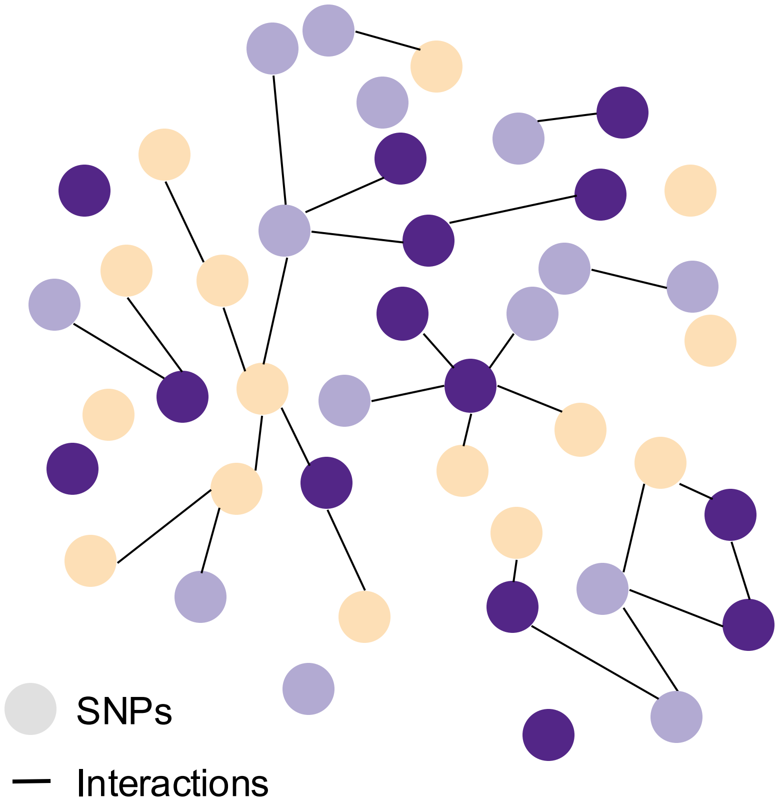
Computational challenges in mapping epistasis

- For p SNPs *$O(p^2)$ possible* pairwise interactions

- For 10^6 SNPs ➔ 10^{12} pairs

- Expectation:
  small residual phenotypic variance ➔ require large data

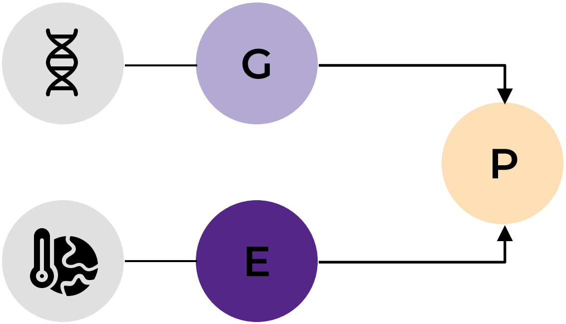
Exhaustive search is underpowered



Phenotypic Variance

Genetic Factors

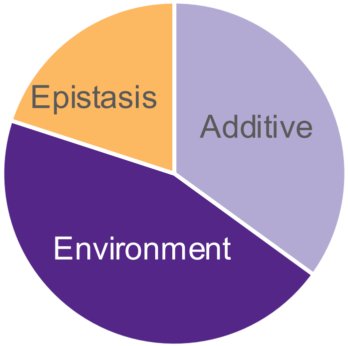
Environmental Factors



$$P = G + E$$

Heritability

$$H^2 = \frac{\mathrm{Var}[G]}{\mathrm{Var}[P]}$$



The Marginal Epistasis Test (MAPIT)

Trait

Geno.

Env.

$$\mathbf{y} = \mu + \beta_k \mathbf{x}_k + \sum_{l \neq k} \beta_l \mathbf{x}_l + \sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l) + \varepsilon$$

$$\varepsilon \sim \mathbf{MVN}(\mathbf{0}, \tau^2 \mathbf{I})$$

Mean

Linear Effects

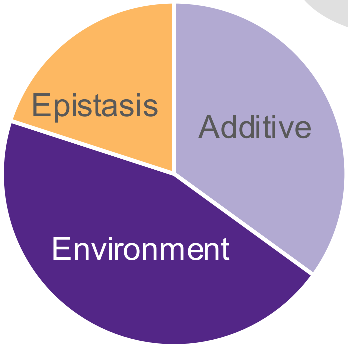
Interaction Effects

Env. Variance Component

$$= \underbrace{\mu + \beta_k \mathbf{x}_k + \mathbf{m}_k}_{\text{Additive Variance Component}} + \underbrace{\mathbf{g}_k}_{\text{Pairwise Epistatic Variance Component}} + \varepsilon$$

Additive Variance Component

Pairwise Epistatic Variance Component



Crawford et al. (2017), *PLOS Gen*

Marginal Epistasis

$$\mathbf{y} = \mu + \beta_k \mathbf{x}_k + \sum_{l \neq k} \beta_l \mathbf{x}_l + \sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l) + \varepsilon$$

Approach

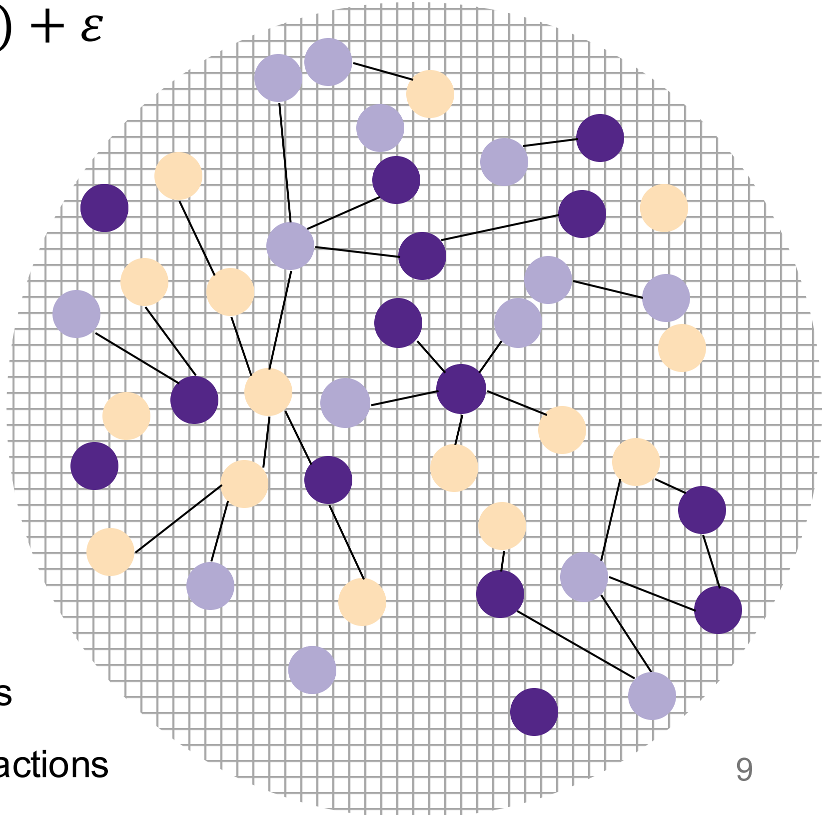
1. Test every SNP for interactions with **any** other
2. Accumulate small effects into **larger** marginal effect
3. Search space O(p) for p SNPs

$$\sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l)$$

$$\sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l)$$

$$\sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l)$$

⋮



SNPs

Interactions

Estimating Variance Components

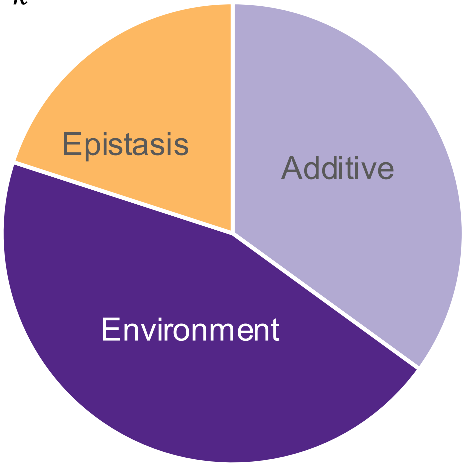
1. Genetic relatedness matrix $\mathbf{K} = \mathbf{X}_{-k}\mathbf{X}_{-k}^T$
2. Interactions of SNP k with its background $\mathbf{G} = \mathbf{D}_k\mathbf{K}\mathbf{D}_k$ with $\mathbf{D}_k = \mathrm{diag}(\mathbf{x}_k)$
3. Estimate variance parameters jointly[1]

Need to compute traces of matrix products

$$\mathrm{tr}(\mathbf{KK}),\ \mathrm{tr}(\mathbf{KG}),\ \mathrm{tr}(\mathbf{GG})$$

!

$$\mathbf{y} = \mathbf{m}_k + \mathbf{g}_k + \varepsilon$$



$$\mathbf{m}_k \sim \mathbf{MVN}(\mathbf{0}, \omega^2\mathbf{K})$$

$$\mathbf{g}_k \sim \mathbf{MVN}(\mathbf{0}, \sigma^2\mathbf{G})$$

$$\varepsilon \sim \mathbf{MVN}(\mathbf{0}, \tau^2\mathbf{I})$$

1 Zhu (2017), Ann. Appl. Stat.

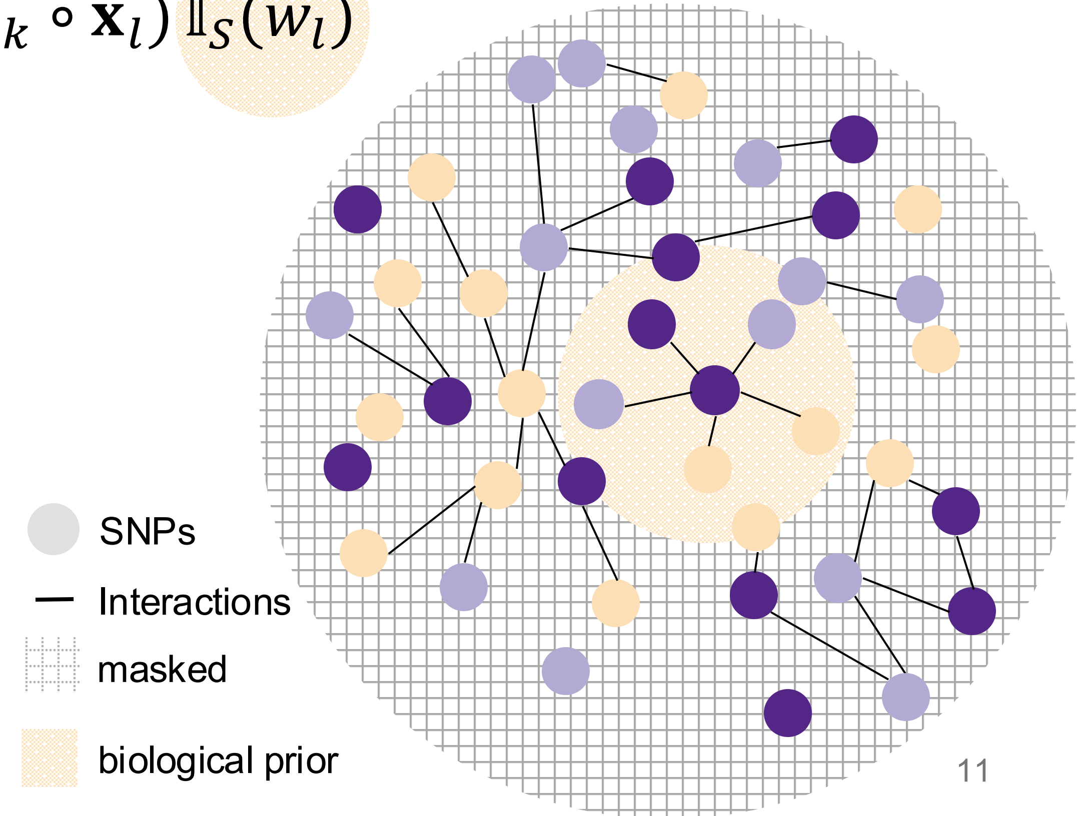
The Sparse Marginal Epistasis Test (SME)

$$\sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l)\, \mathbb{I}_S(w_l)$$

Approach

External data source S as biological prior via indicator function $\mathbb{I}_S(w_l)$

1. Unlocks scalability of method
2. Improves signal of marginal epistasis



Computational Improvements

Stochastic Trace[1]

→ $\mathrm{tr}(\mathbf{KK}) \approx \frac{1}{B}\sum_{i=1}^{B} \boldsymbol{z}_i^T \boldsymbol{X}^T \boldsymbol{X}\boldsymbol{X}^T \boldsymbol{X}\boldsymbol{z}$ with $\boldsymbol{z} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I})$ **random vector**

→

- **Faster** vector by matrix multiplication
- Enables **column wise processing** of large data[2] ➔ low memory need

"Recycle" random vectors

→ GRM $\mathbf{K}$ is the same for all tests ($\mathbf{G}$ changes)

Compute $\mathbf{K}\boldsymbol{z}_i$ **once for multiple tests**

→

- Avoid repeated computation of GRM derived quantities
- **Speed up** computation

Sparse Modeling of Epistasis

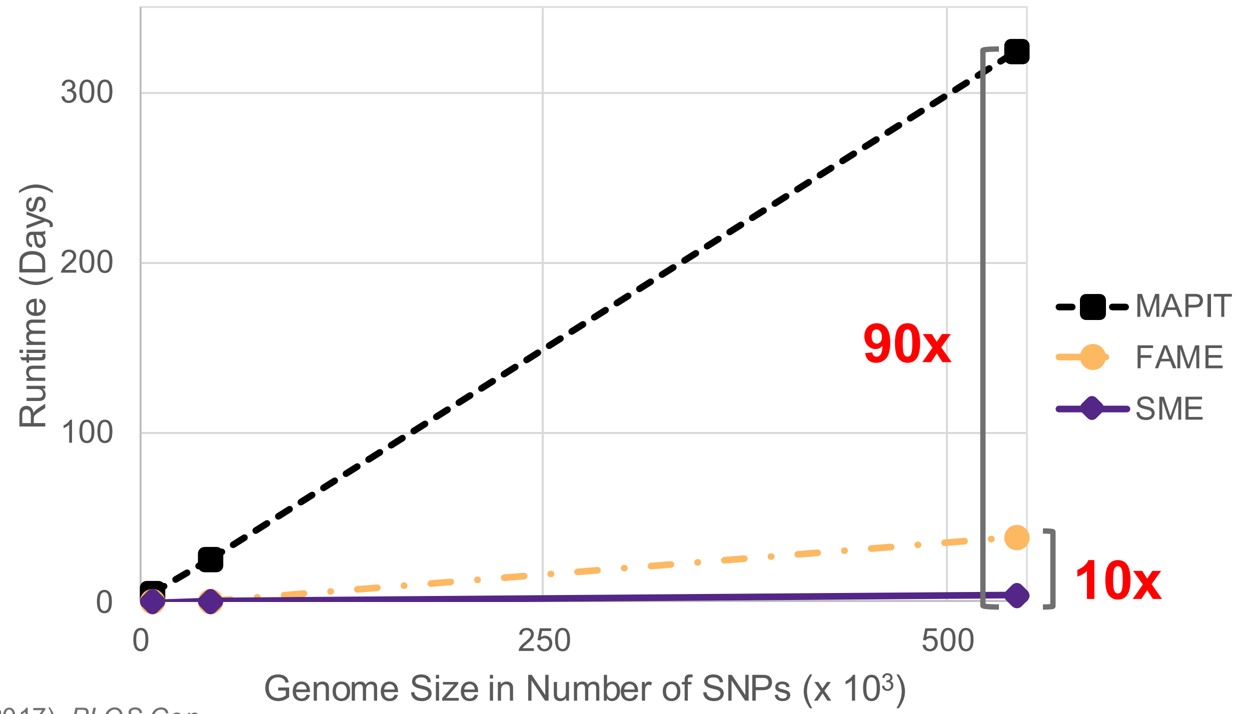
→ Use additional reference data as **biological prior**

→

- More efficient estimator ➔ **improved power**
- **Speed up** computation

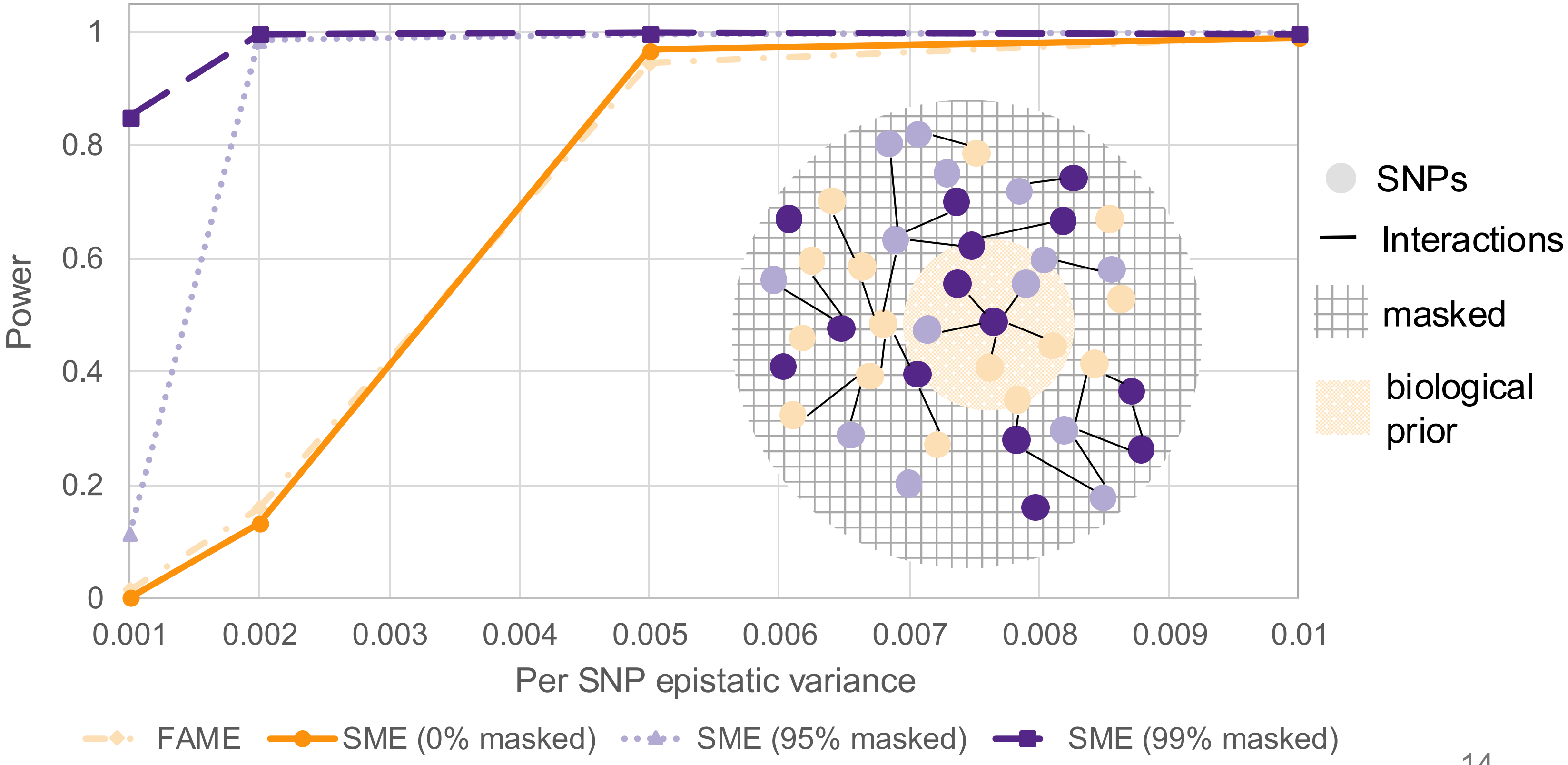
1 Hutchinson (1979). *Commun. Stat. Simul. Comput.*   2 Fu et al. (2023). *bioRxiv.*

SME scales genome-wide in biobank data

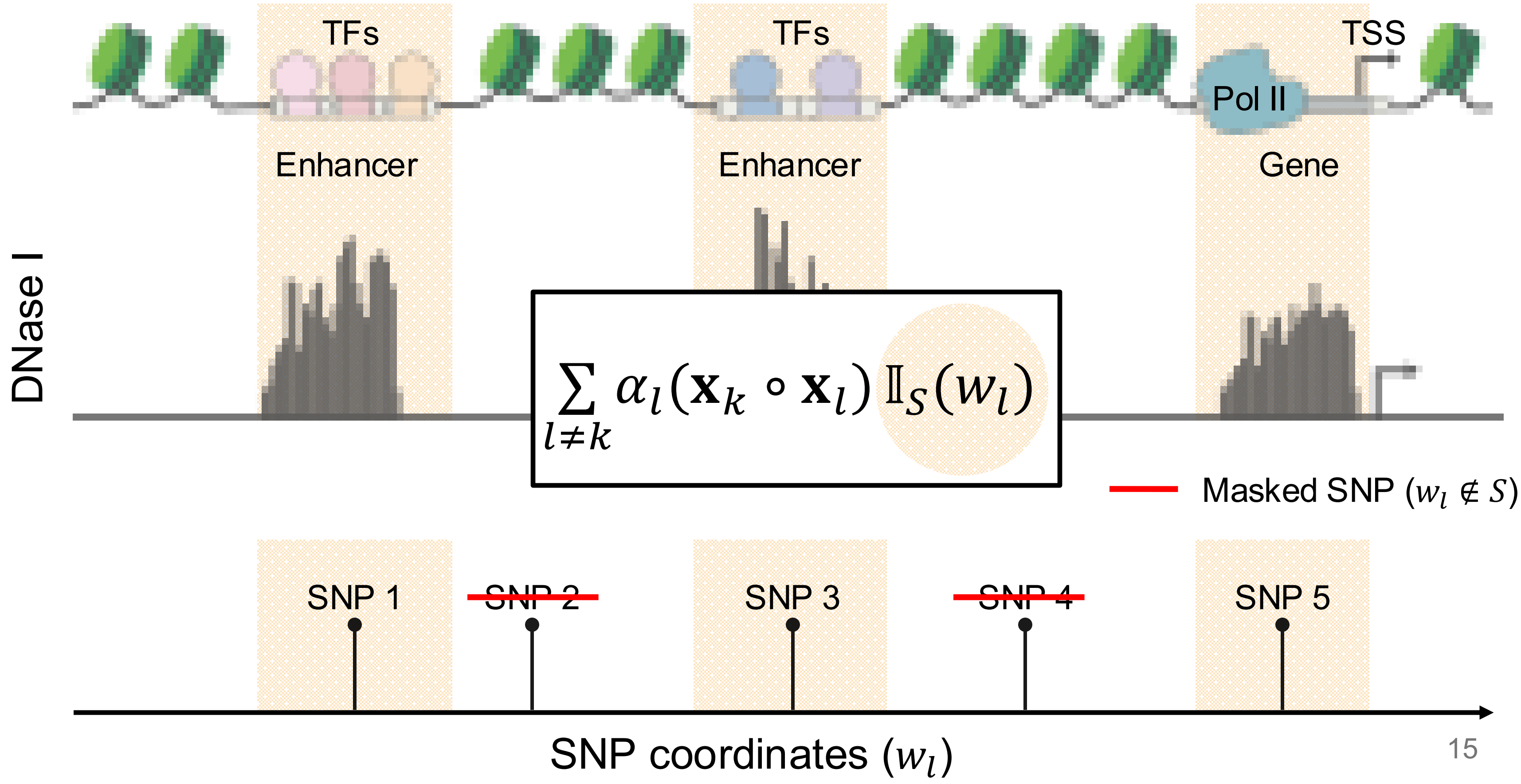


MAPIT - Crawford et al. (2017), *PLOS Gen.*
FAME - Fu et al. (2023), *bioRxiv*

Sparsity improves power



Example Biological Prior: Open Chromatin



$$\sum_{l \neq k} \alpha_l (\mathbf{x}_k \circ \mathbf{x}_l)\, \mathbb{I}_S(w_l)$$

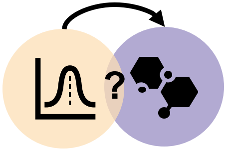
Application to the UK Biobank

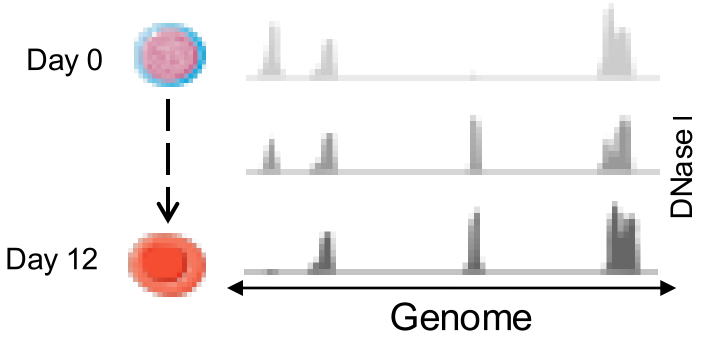
SME with Open Chromatin

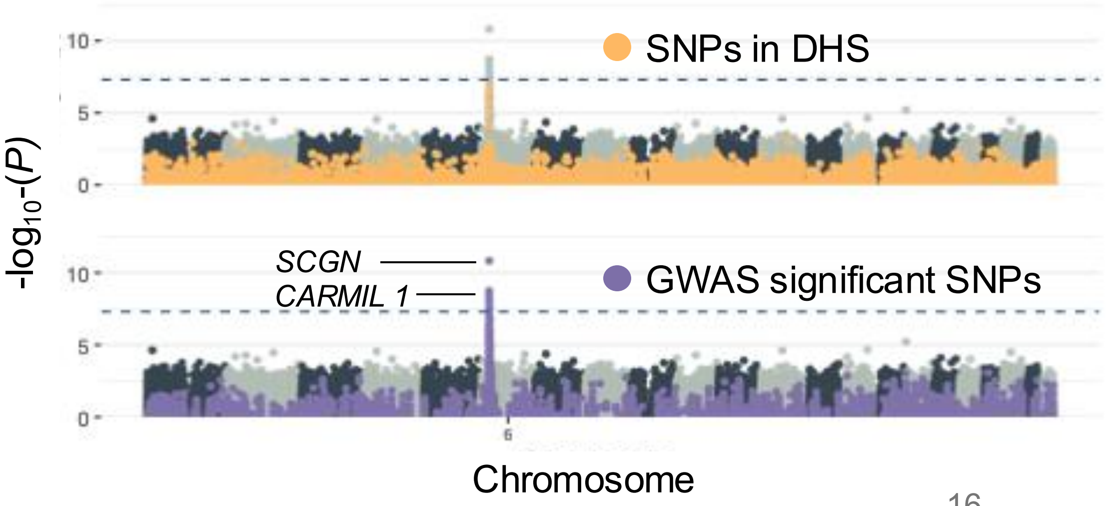
1. Mean Corpuscular Hemoglobin in 350k individual genotyped at 544k SNPs[1]

2. ~5k SNPs in DNase I hypersensitive sites (DHS) during *ex-vivo* erythroid differentiation[2]

3. Identifies plausible **statistical** signal for **functional epistasis:** *SCGN* and *CARMIL1*







1 Bycroft et al. (2018). *Nature*
2 Georgolopoulos et al. (2021). *Nat. Commun.*

Takeaways

SME is a well powered epistasis test that scales genome-wide in biobank data

01 Marginal epistasis overcomes the search space and small effect problem

02 Sparse biological priors focus search and improve efficiency of estimators

03 Stochastic trace estimators allow resource efficient processing

The manuscript is accepted at the American Journal of Human Genetics (publication on July 29 2025)

SME software is open source available on CRAN (R Package)

SME

```
> install.packages('smer')
```

Acknowledgements

Committee

Lorin Crawford
Daniel Weinreich
Roberta DeVito
Bogdan Pasaniuc (UPenn)

Collaborator

Samuel Pattillo Smith
(Genomics Ltd.)

Friends and Colleagues

Alan DenAdel
Alex Wong
Ashok Ragavendran
Boyang Fu (UCLA)
Cecile Meier-Scherling
Chibuikem Nwizu
Dave Peede
Ria Vinod
Shevaughn Holness
Sriram Sankararaman (UCLA)
Xavier Roca-Rada
Whitney Sloneker

CCMB
BROWN
Crawford Lab.

References

Variance Component Estimation

- X. Zhou. *A unified framework for variance component estimation with summary statistics in genome-wide association studies.* Ann. Appl. Stat. 11 (4) 2017 - 2051, December 2017. https://doi.org/10.1214/17-AOAS1052

Marginal Epistasis Detection

- L. Crawford, P. Zeng, S. Mukherjee, & X. Zhou, (2017). *Detecting epistasis with the marginal epistasis test in genetic mapping studies of quantitative traits*. PLOS Genetics, 13(7), e1006869. https://doi.org/10.1371/journal.pgen.1006869
- **J. Stamp**, A. DenAdel, D. Weinreich, & L. Crawford, (2023). *Leveraging the Genetic Correlation between Traits Improves the Detection of Epistasis in Genome-wide Association Studies*. G3 Genes|Genomes|Genetics, jkad118. https://doi.org/10.1093/g3journal/jkad118
- **J. Stamp,** S. Pattillo Smith, D. Weinreich, L. Crawford (2025). *Sparse modeling of interactions enables fast detection of genome-wide epistasis in biobank-scale studies*. American Journal of Human Genetics (publication on July 29 2025)

Interaction-LD Score Regression:

- S. Pattillo Smith, G. Darnell, D. Udwin, **J. Stamp**, A. Harpak, S. Ramachandran, L. Crawford (2024) *Discovering non-additive heritability using additive GWAS summary statistics*. eLife 13:e90459

Related Software:

- mvMAPIT: https://lcrawlab.github.io/mvMAPIT/
- SME: https://lcrawlab.github.io/sme/