Inferring environmental factors to increase genetic discovery power in genome-wide association studies

Saurav Mathur[1,2], Tiffany Phan[1,3], Robert Brown[4], Sriram Sankararaman[4,5]

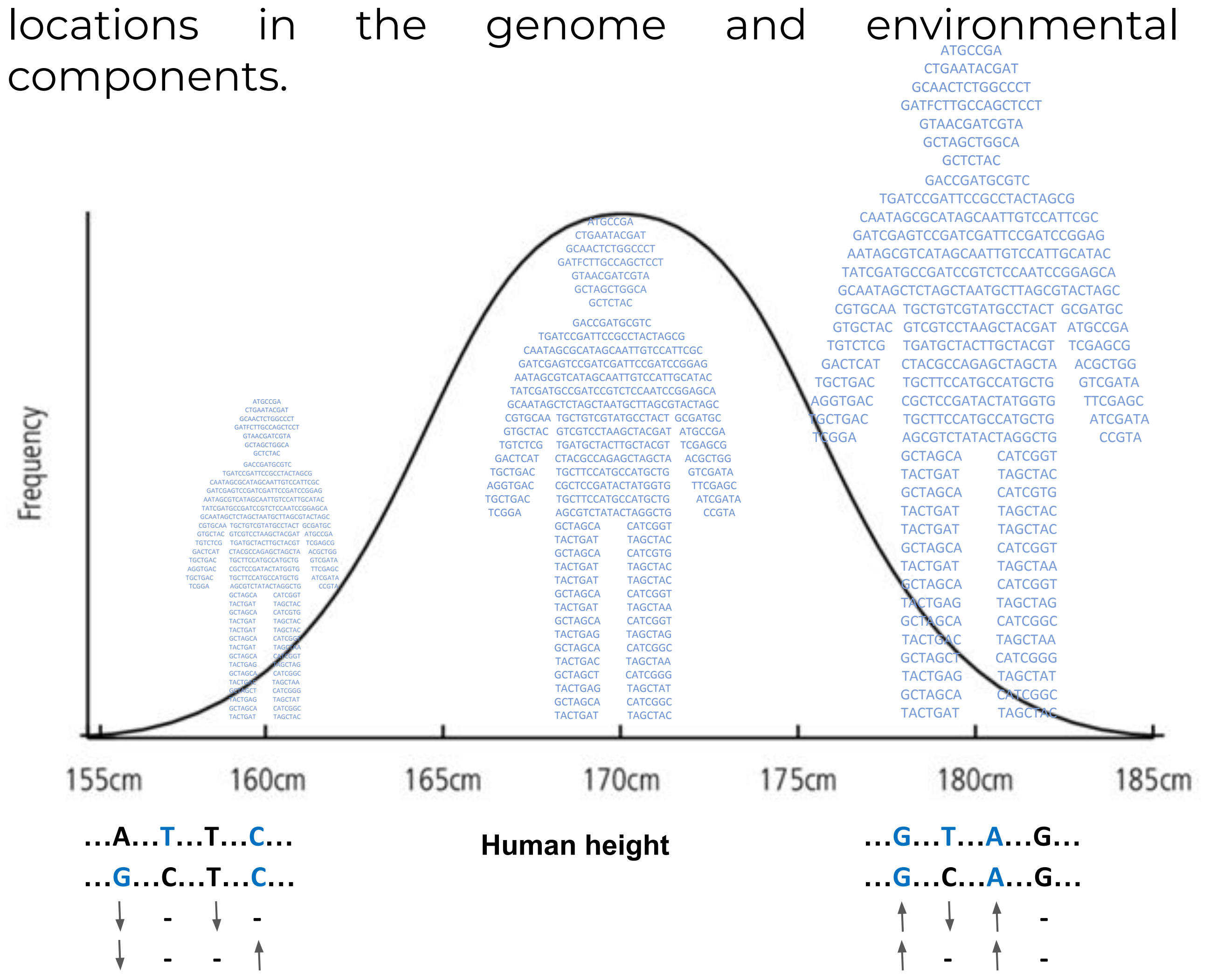
[1]BIG Summer Program, Institute for Quantitative and Computational Biosciences, University of California, Los Angeles, Los Angeles, 90095, USA, [2]Department of Electrical and Computer Engineering, University of Wisconsin-Madison, 53706, USA, [3]Department of Computer Science, University of Colorado Boulder, Boulder, Colorado, 80309, USA, [4]Department of Computer Science, University of California, Los Angeles, Los Angeles, 90095, USA, [5]Department of Human Genetics, University of California, Los Angeles, Los Angeles, 90095, USA

Objectives

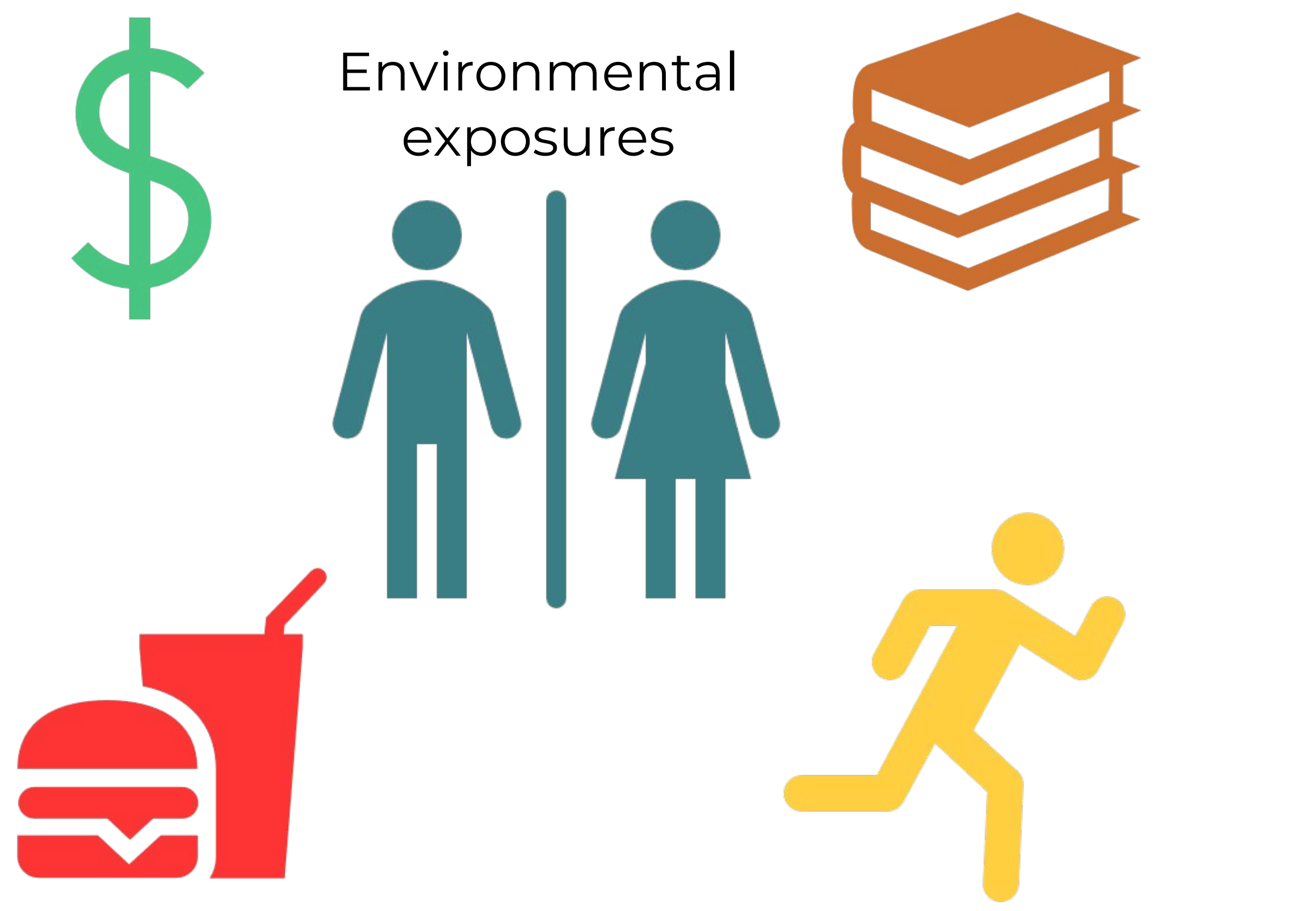
Environmental factors affecting disease risks are often difficult to measure. Our objective is to infer independent environmental components from easy to ascertain phenotype data.

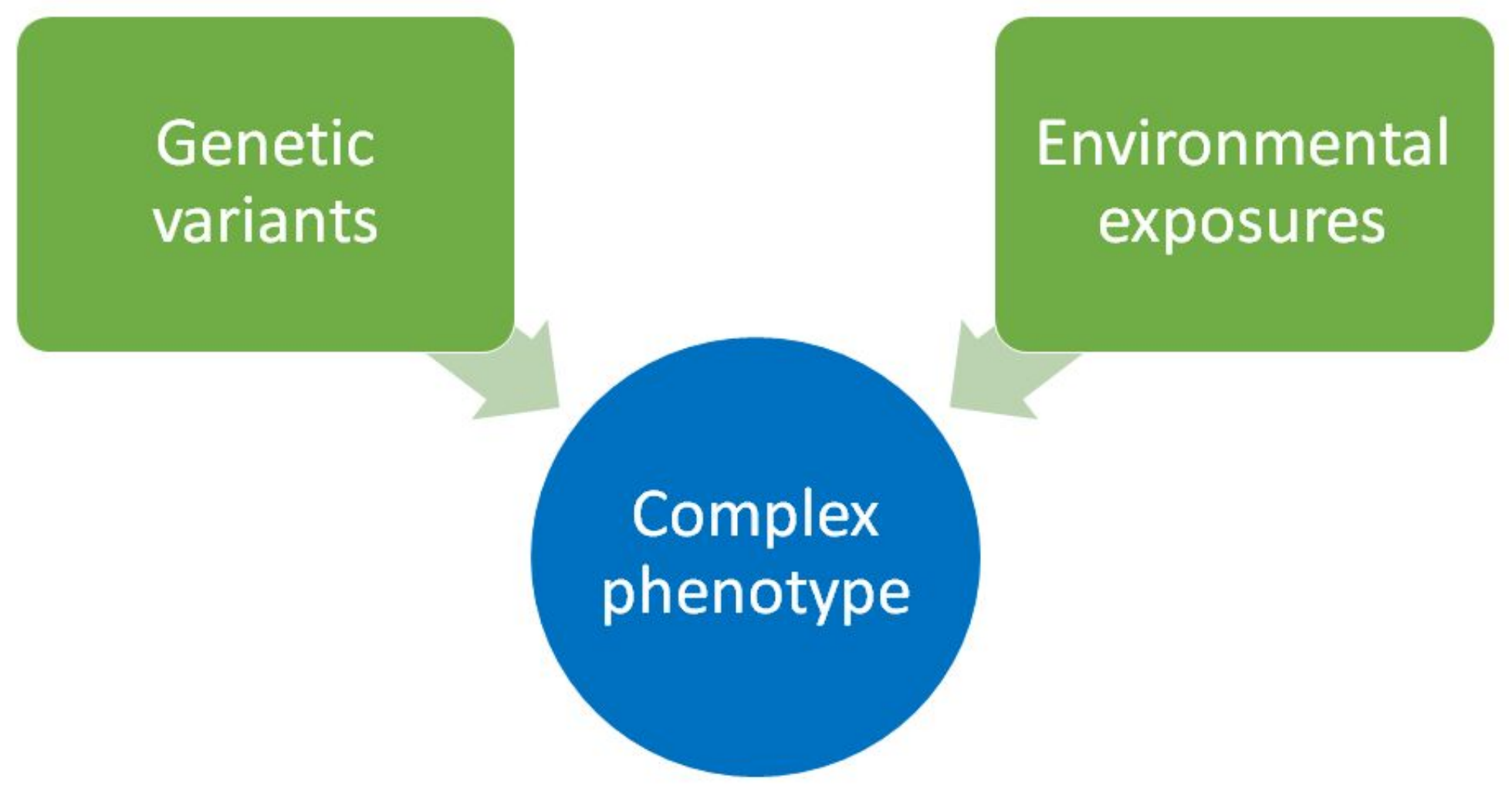
Background

- Complex phenotypes are traits or diseases, such as height or diabetes, that are impacted by multiple locations in the genome and environmental components.



- Measuring diverse environmental factors is difficult and it is unknown a priori which environmental factors affect the trait. Because of this, many studies only account for the most basic environmental factors, such as age and gender.



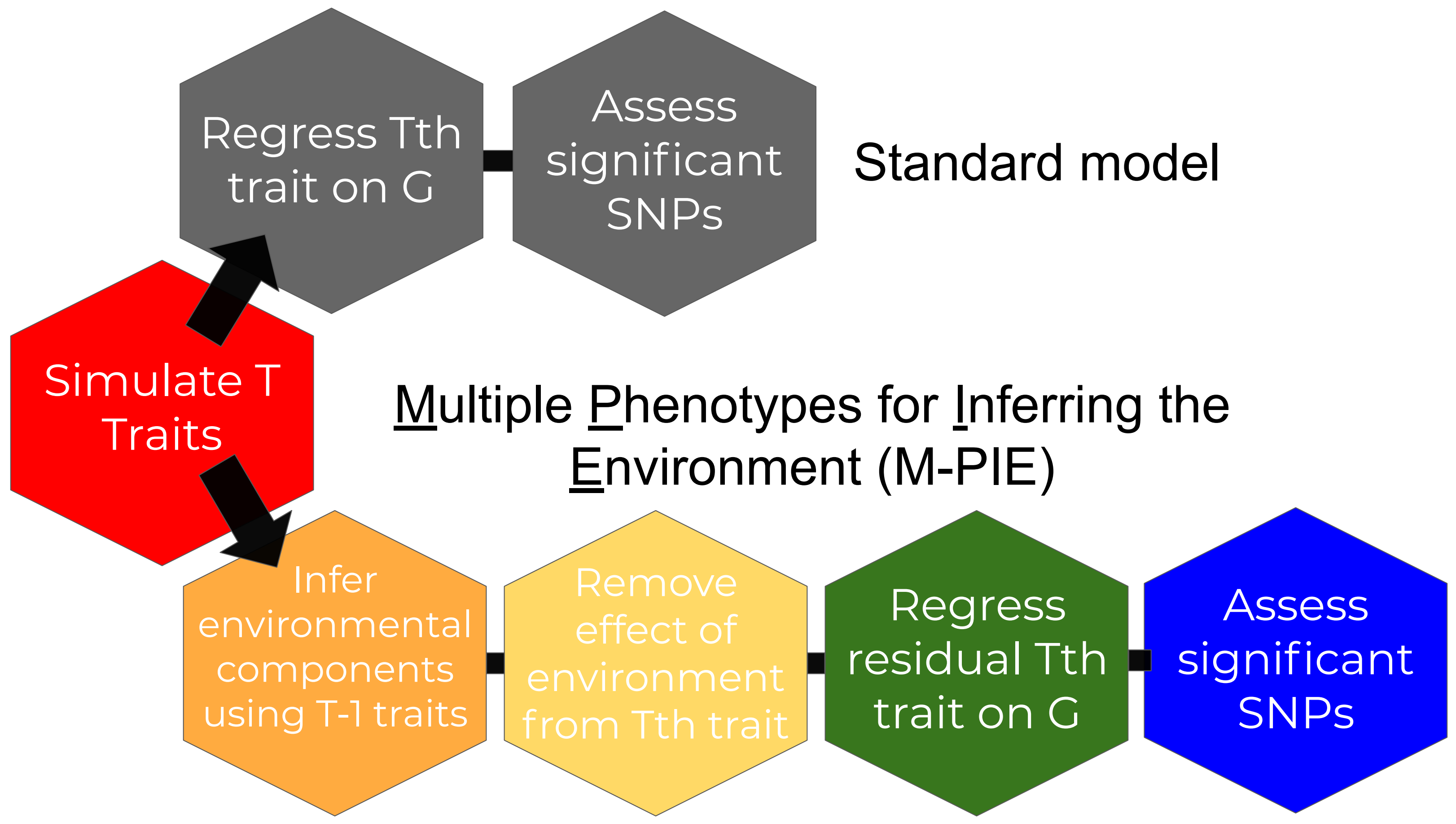


Methods

Simulation Model

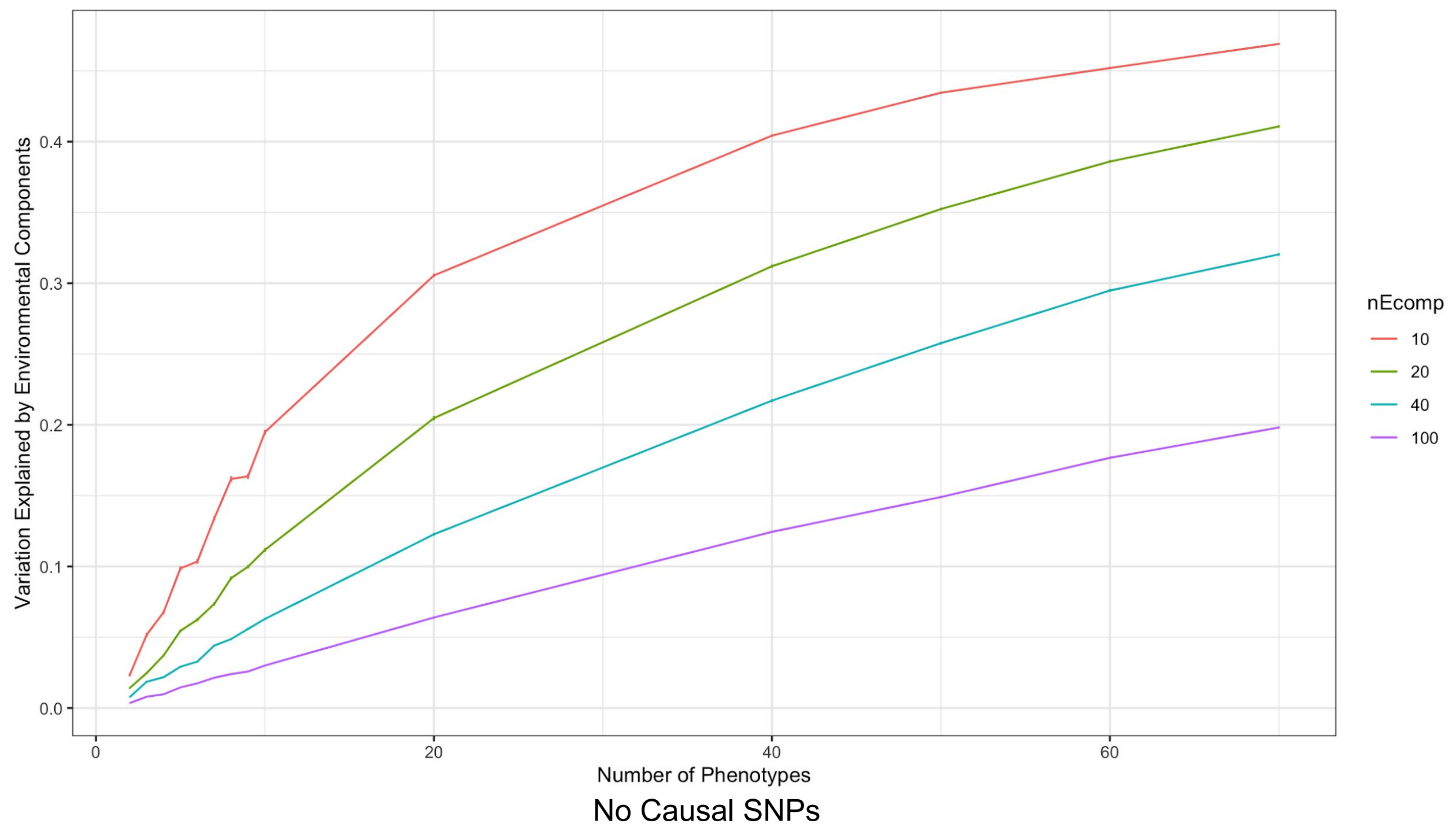
$$\begin{bmatrix} y_1^1 & \dots & y_1^T \\ y_2^1 & \dots & y_2^T \\ \vdots & \dots & \vdots \\ y_N^1 & \dots & y_N^T \end{bmatrix} = \begin{bmatrix} x_{11} & \dots & x_{1M} \\ x_{21} & \dots & x_{2M} \\ \vdots & \vdots & \vdots \\ x_{N1} & \dots & x_{NM} \end{bmatrix} \begin{bmatrix} \beta_1^{g1} & \dots & \beta_1^{gT} \\ \beta_2^{g1} & \dots & \beta_2^{gT} \\ \vdots & \dots & \vdots \\ \beta_M^{g1} & \dots & \beta_M^{gT} \end{bmatrix} + \begin{bmatrix} e_{11} & \dots & e_{1Q} \\ e_{21} & \dots & e_{2Q} \\ \vdots & \vdots & \vdots \\ e_{N1} & \dots & e_{NQ} \end{bmatrix} \begin{bmatrix} \beta_1^{e1} & \dots & \beta_1^{eT} \\ \beta_2^{e1} & \dots & \beta_2^{eT} \\ \vdots & \dots & \vdots \\ \beta_Q^{e1} & \dots & \beta_Q^{eT} \end{bmatrix} + \begin{bmatrix} e_1^1 & \dots & e_1^T \\ e_2^1 & \dots & e_2^T \\ \vdots & \dots & \vdots \\ e_N^1 & \dots & e_N^T \end{bmatrix}$$

Discovery Pipeline


Regress Tth trait on G
Assess significant SNPs
Standard model
Simulate T Traits
Multiple Phenotypes for Inferring the Environment (M-PIE)
Infer environmental components using T-1 traits
Remove effect of environment from Tth trait
Regress residual Tth trait on G
Assess significant SNPs

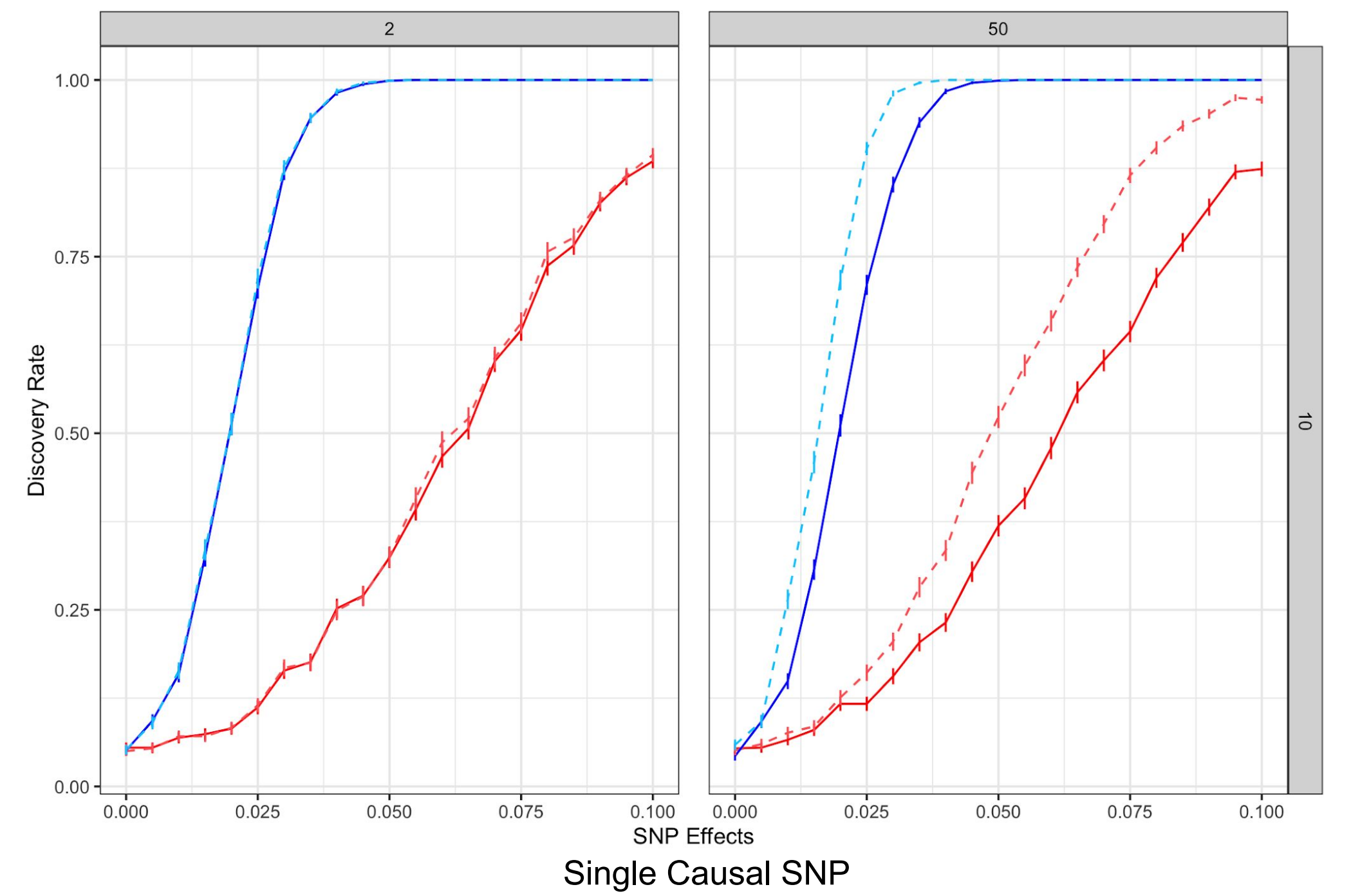
Results

Inferred environmental components increasingly capture environmental noise as the number of ascertained phenotypes increases.



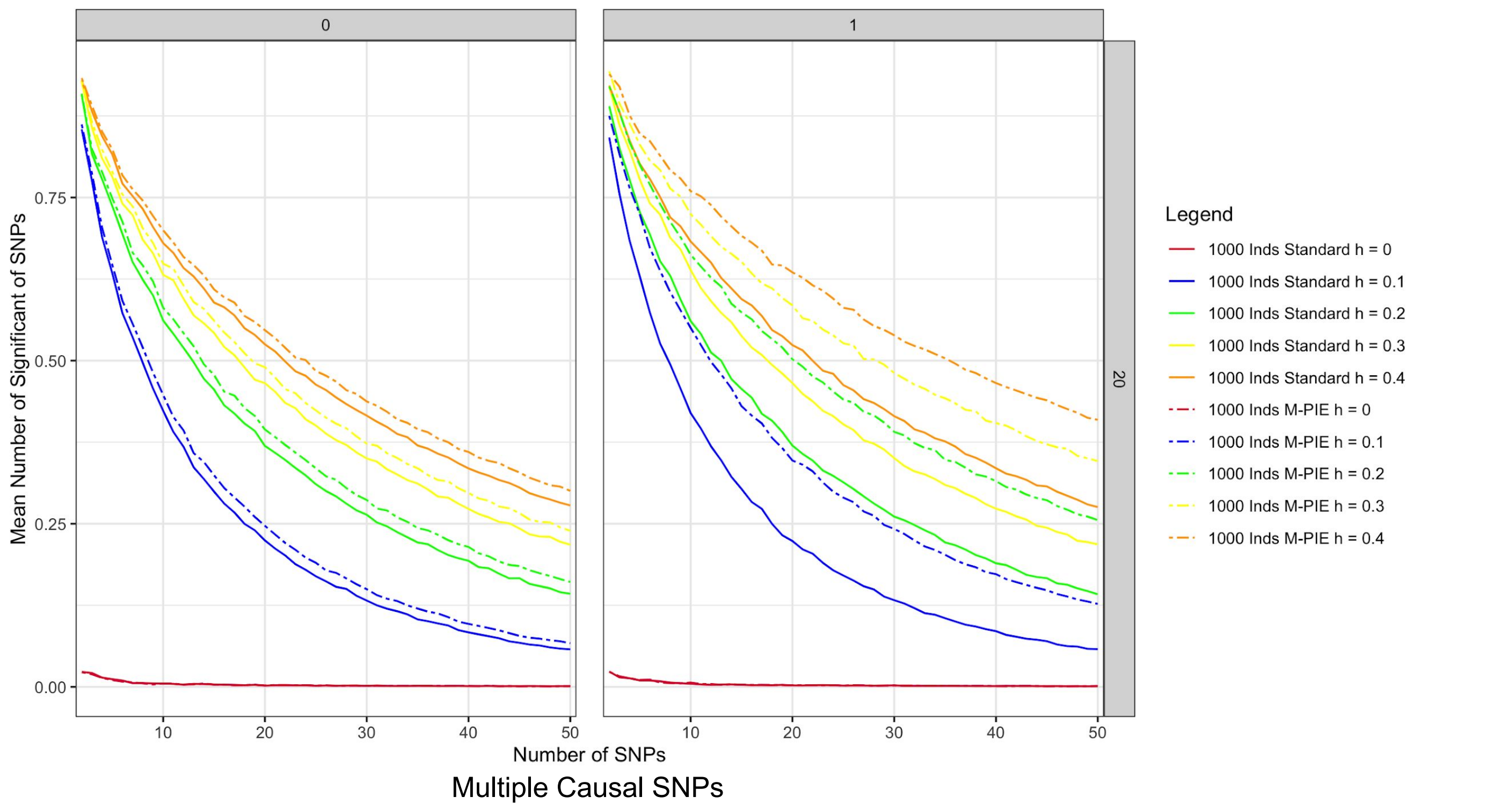
No Causal SNPs

Ascertaining a large number of phenotypes significantly increases the discovery rate. Here we show the difference in discovery rate between 2 and 50 phenotypes for 10 environmental components.



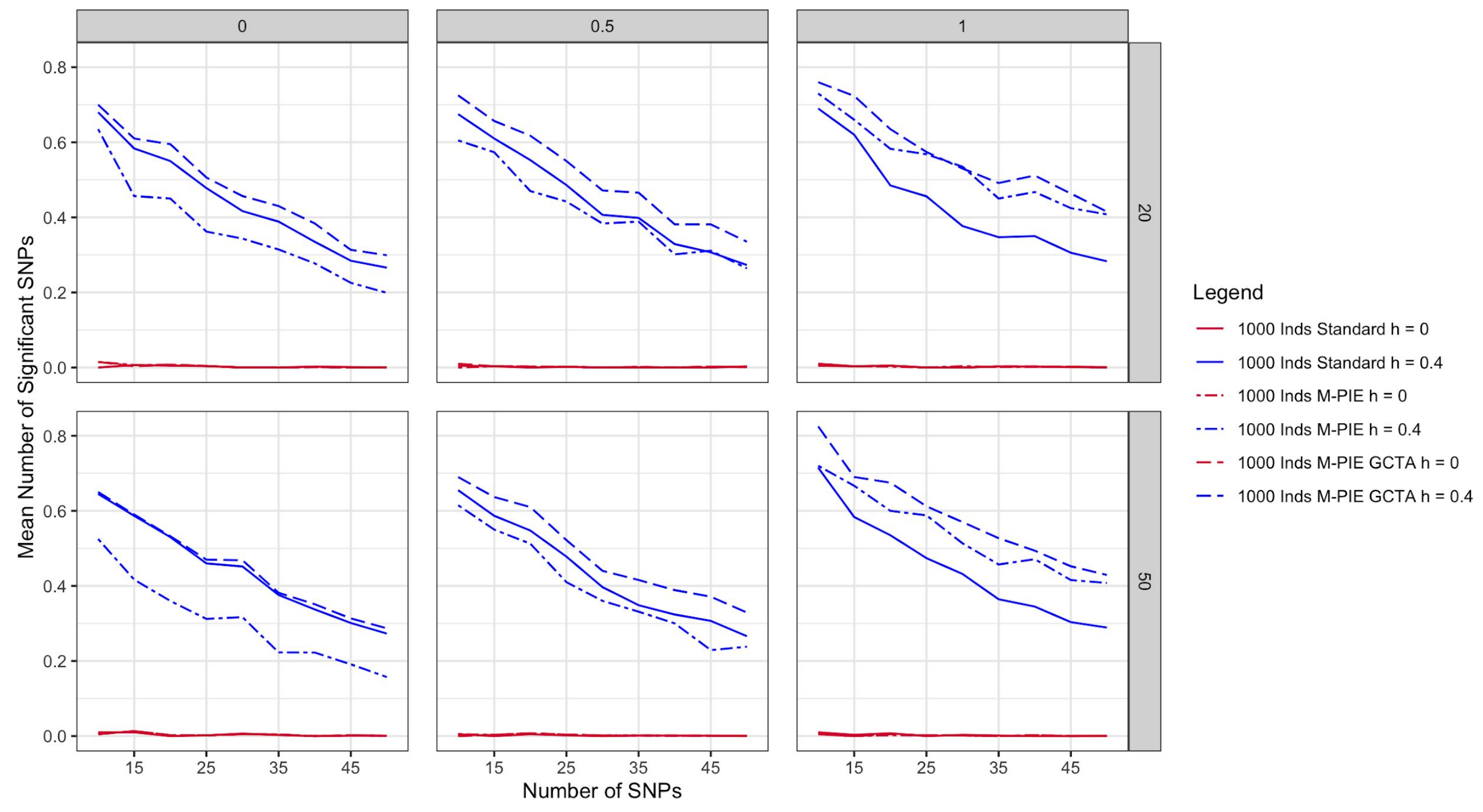
Single Causal SNP

0% and 100% correlation within environmental component blocks for 20 environmental components. Increasing correlation decreases the effective number of environmental components, resulting in increased discovery rate for 20 environmental components and 50 phenotypes.



Multiple Causal SNPs

Using the GCTA tool, we first regress the genetics out of phenotypes to ascertain the inferred environment using principal component analysis. This further increases the power of discovering significant SNPs by ensuring that PCs do not capture the genetic effect.



Infinitesimal model

Conclusion

In our work, we demonstrate that with many ascertained phenotypes, it is possible to infer unmeasured environmental components. We show that by regressing out the inferred environmental components, we reduce variation within the phenotypes of interest, which consequently increases genetic discovery power. This work has wide applications to emerging biobanks and medical databases where many individuals have been deeply phenotyped.

Future Directions

- Apply our model to real data from UK Biobank or other databases with deep phenotyping.
- Correlate inferred environmental components to real environmental exposures.
- Explore gene-by-environment interactions using inferred environmental components.

Acknowledgements

UCLA QCBio BIG BIO Sriram Lab