A Bayesian Detection of Epistasis Using Summarized Synergy Networks in Islet Autoimmunity

Shane E. Ridoux

Advised by Drs. Randi K. Johnson, Joshua P. French & Erin E. Austin

University of Colorado-Denver

April 28, 2025

Contents

Dedication

To Andrew Gavin Olguin, this is for you.

I could not have done it without you.

Acknowledgements

I would like to acknowledge the National Human Genome Research Institute's Pathways in Genomic Data Science Fellowship (R25HG012919) and the University of Colorado - Denver for their generous financial support of this research and my graduate education.

I would like to express my gratitude to the members of the Diabetes AutoImmunity Study in the Young (DAISY) for their invaluable contribution, without which this data would not exist.

I am truly grateful for the mentorship I have received from my committee members - Randi K. Johnson (Advisor), Erin E. Austin, & Joshua P. French.

I am very fortunate to have been a part of the RKJCollab. Thank you Sarah, Charlie, Kirk, and Lauren for your support, mentorship, and knowledge.

Thank you to my family and friends for helping me find balance en route to finishing this thesis.

Abstract

Type 1 Diabetes (T1D) is a complex autoimmune disease characterized by the erroneous destruction of islet β-cells resulting in life-long dependence on exogenous insulin. Although T1D is considered to be largely genetic, with the heritability of the disease around 80%, roughly 20% of the phenotypic variance is still unexplained. Epistasis, the interaction of genetic variants, is thought to contribute to this missing heritability in T1D, yet the epistatic landscape of T1D and its precursor Islet Autoimmunity (IA) remains understudied. A genome-wide analysis of the epistatic architecture of IA could inform potential biological pathways in the etiology of T1D in addition to elucidating aspects of its missing heritability.

After quality control, 359 DAISY children selected for nested case-control study with a customized Exome array imputed to the TOPMed r2 reference panel were included for analysis of genome-wide epistasis detection. Within-gene synergy networks were used to capture non-linear epistatic effects and summarized through diffusion kernel principal components (kPC). Between-gene epistasis was detected through a bayesian interaction analysis of kPC-summarized genes.

RESULTS

DISCUSSION

1 Introduction

1.1 Background

Epistasis, first coined by Bateson in 1909, has been defined and redefined and can carry with it some ambiguity in its definition. Often neglected, he also defined the term hypostasis[1,2]. These two terms were initially meant to describe a hierarchical relationship where a higher factor (epistatic gene/allele) controls whether a lower factor (hypostatic gene/allele) can be expressed. Here *epistatic* and *hypostasis* come from the Greek prefixes *epi* meaning "upon", *hypo* meaning "under", and root *stasis* meaning "standing". Thus, to Bateson, an epistatic gene "stands upon" a hypostatic gene, masking the hypostatic gene's expression. Epistasis, or *dual epistacy*, was redefined by Fisher in 1918 for a statistical context in which there is a deviation from linearity of allelic effects in relation to some quantitative phenotype D[2,3]. Here we already see differing definitions of epistasis just 9 years a part with differing implications. Fisherian statistical epistasis does not necessarily imply biological or functional epistasis, however, it can provide evidence for further investigation.

The modern definition of epistasis is more of an umbrella term for genetic interaction. Under this umbrella, we find *statistical epistasis*, which is the average epistasis measured over random genetic backgrounds at the population level, and *positive/negative epistasis*, which refers to the effect of two mutations where postive (negative) epistasis implies the effect of the double mutant is greater (less) than the sum of its parts, i.e. the sum of the effects of the individual mutations[4].

1.2 Motivation

2 Methodology

2.1 Overview

2.2 Data Preparation

Talk about imputation and DAISY Cohort and potentially TEDDY Cohort

2.3 Within-Gene Epistasis Detection

Talk about method generally

2.3.1 Computing Entropy-Based SNP Synergies

The concept of bivariate synergism quantifies the combined effect of two SNPs (A and B) on a disease (D) beyond their individual contributions using information theory. It is calculated as:

$Syn(A;B;D) = I(A,B;D) - [I(A;D) + I(B;D)]$

where $Syn(A;B;D)$ compares the joint contribution of SNPs A and B to the disease D with the additive contributions of the individual SNPs. The information gain I(A; D) about the disease D due to knowledge about SNP A and is defined as:

$I(A;D) = H(D) - H(D|A)$

$I(A,B;D) = H(D) - H(D|A,B)$

where $H(\cdot)$ is the entropy.

$H(D) = \sum_d p(d) log(\frac{1}{p(d)})$

$H(D|A) = \sum_{a,d} p(a,d) log(\frac{1}{p(d|a)})$

where $p(d)$ is the probability of having disease $D = d$ and $p(d|a)$ is the probability of having disease $D = d$ given SNP A has genotype a. A Network results from the bivariate synergy calculations of each SNP where the edge of nodes A and B has weight $Syn(A;B;D)$.

Entropy can be thought of as a measure for the uncertainty in a

2.3.2 Diffusion Kernels

Why diffusion kernels? Essentially allow for gene summaries that become the explanatory variable in the Bayesian model...

2.4 Between Gene Epistasis Detection

2.4.1 Bayesian Interaction Modeling

Detecting gene-gene interactions

- one continuous variable per gene
- Bayesian semi-parametric regression approach to infer non-linear gene-gene interaction
- As gene explanatory variables for individual i, used first kernel PCA X_{ij_1} for each gene j_1
- genes and gene-gene interactions are selected on Posterior Inclusion Probability (PIP)

Let p be the number of genes, k be a number sufficiently large $\ni$ all exposures are captured in the model, and X_{ij_1} be the first kernel PCA for gene j_1 as gene explanatory variable for individual i.

Assume:

$$Y_i \sim N(f(X_i, \sigma^2)) \tag{1.1}$$

$$f(X_i) = \sum_{k=1}^{p} f^{(h)}(X_i) \tag{1.2}$$

$$f^{(h)}(X_i) = \sum_{j_1=1}^{p} \tilde{X}_{ij_1} \beta_{j_1}^{(h)} + \sum_{j_1=2}^{p} \sum_{j_2 \leq j_1} \tilde{X}_{ij_1 j_2} \beta_{j_1 j_2}^{(h)} \tag{1.3}$$

$$\tilde{X}_{j_1} = g_1(X_{j_1}), \quad \text{where } g_1(\cdot) \text{ is the natural spline basis function} \tag{1.4}$$

3 Results

3.1 Overview

3.2 Plots

3.3 Limitations

4 Discussion

5 Conclusion

6 References

References

[1] Bateson, W. (1909). *Mendel's Principles of Heredity.* Cambridge, Massachusetts London, England: Cambridge University Press.

[2] Cordell, H. J. (2002, October). Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics 11*(20), 2463–2468.

[3] Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh 52*(2), 399–433.

[4] Johnson, M. S., G. Reddy, and M. M. Desai (2023, May). Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biology 21*(1), 120.

A Data Preprocessing

B Code

C Additional Figures & Tables