A Bayesian Discovery of Complement Epistasis in the Natural History of Type 1 Diabetes

Shane Ridoux

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

University of Colorado-Denver

February 3, 2025

Contents

1 Introduction

1.1 Background

Type 1 Diabetes (T1D) is an autoimmune disorder where the immune system erroneously targets and destroys insulin-producing pancreatic islet-β cells, leading to a lack of insulin and elevated blood glucose levels. Often diagnosed in early childhood, T1D left untreated can result in serious complications such as diabetic ketoacidosis, a life-threatening condition. With the disease incidence rising, efforts to understand the pathogenesis and etiology of T1D continue with the hopes of phase-specific therapeutic intervention to mitigate life-threatening complications, eliminate the burden of insulin pumps, and improve quality of life.

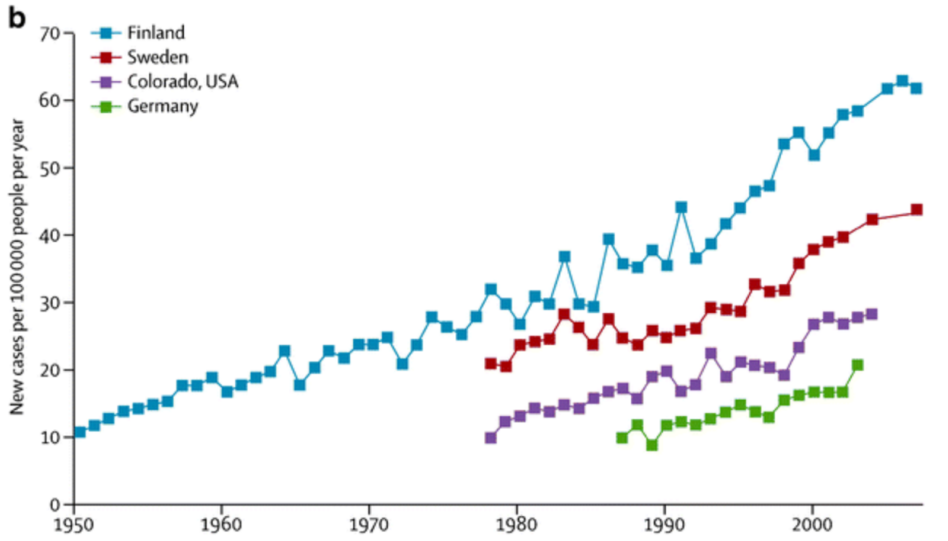

Figure 1: Incidence of T1D

Type 1 Diabetes is strongly influenced by genetic factors, with nearly 40% of the risk attributed to the highly polymorphic Major Histocompatibility Complex (MHC), which is associated with various diseases. This region includes genes encoding components of the complement system which acts as a rapid and targeted innate immune defense mechanism against pathogens, primarily through promoting inflammation and modulating the adaptive immune response. However, the efficiency and regulation of complement activity varies among individuals largely due to inherited genetic differences.

1.2 Motivation

While genetic polymorphisms within the complement system significantly affect its activity and regulation, their role in immune activation and T1D progression remains underexplored due to the oversight of epistatic interactions. Genetic epistasis is most commonly understood as gene interaction where the

contribution of one gene on the phenotypic outcome is dependent on genetic background (citation). The complotype is the total inherited set of genetic variants in complement genes (citation) and can be thought of as the genetic background for the complement system.

The complotype impacts the activation potential of the complement and immune system where an overactive complement system influences susceptibility to inflammation and autoimmune disease while an underactive complement system results in increased risk for infection. While individual polymorphisms within the complotype have a minor impact on activation potential, their aggregate effect is believed to be greatly amplified either synergistically or antagonistically (citation?). Polymorphisms in complement genes have been shown to be associated with type 1 diabetes however, epistasis in complement genes has yet to be studied in type 1 diabetes.

2 Methodology

2.1 Bayesian Motivation (Moti-Bayesian)

A Genome-Wide Association Study (GWAS) tests millions of representative single nucleotide polymorphisms (SNPs) from a population in a linear framework where each SNP is tested for association with a phenotype individually (citation). Testing interactions among these SNPs in pairs or tuples of length n drastically increase the already large number of hypotheses. A typical approach adjusting for multiple hypotheses in a GWAS would be to use Bonferoni correction for the determined number of independent SNPs in a genome with the standard significance threshold being set at 5×10^{-8}.

There are issues with this standard of practice. The standard genome-wide significance threshold faces challenges in evolving GWAS practices. It was developed for common variants, potentially lacking power for rare variants. It does not take into account the prior probability of a variant being associated with a phenotype, nor does it account for the statistical power of the test which both influence the interpretation of the p-value. Additionally, its foundation on the number of independent SNPs in a European reference population neglects the genetic diversity across populations and the influence of genetic context. These oversights can distort results, particularly when considering epistasis, where interactions between variants may not align with assumptions of independence. Applying the threshold universally risks missing meaningful associations while inflating false positives, especially in diverse populations or complex phenotypes. Bayesian methods have been applied to address false positives in GWAS especially in the context of epistasis.

2.2 Data Preparation

Talk about imputation and DAISY Cohort and potentially TEDDY Cohort

2.3 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 $Syn(A;B;D)$.

2.3.2 Diffusion Kernels

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

2.3.3 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:

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

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

$$f^{(h)}(X_i) = \sum_{j_1=1}^{p} \tilde{X}_{ij_1}\beta_{j_1}^{(h)} + \sum_{j_1=2}^{p}\sum_{j_2\leq j_1} \tilde{X}_{ij_1j_2}\beta_{j_1j_2}^{(h)} \tag{1.3}$$

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

2.4 Network Interpretation

3 Results

3.1 Plots

3.2 Limitations and challenges

4 Conclusion and possible extensions

5 Bibliography

References

citation

6 R-codes: