MS Thesis Proposal

BNA Title that tim is still thinking about

Tim Vigers

2020

Table of Contents

[Abstract 2](#_Toc45546113)

[Introduction 3](#_Toc45546114)

[DNA Methylation 3](#_Toc45546115)

[Figure 1: DNA Methylation Patterns 3](#_Toc45546116)

[Metabolomics 3](#_Toc45546117)

[Figure 2: Micronutrients in One-Carbon Metabolism and DNA Methylation 4](#_Toc45546118)

[Type 1 Diabetes 5](#_Toc45546119)

[Specific Aims 6](#_Toc45546120)

[Primary Aim 1 6](#_Toc45546121)

[Secondary Aim 1a 6](#_Toc45546122)

[Secondary Aim 1b 6](#_Toc45546123)

[Primary Aim 2 6](#_Toc45546124)

[Primary Aim 3 6](#_Toc45546125)

[Methods 6](#_Toc45546126)

[Methylation 6](#_Toc45546127)

[Metabolomics 6](#_Toc45546128)

[Feature Reduction 6](#_Toc45546129)

[Bayesian Network Analysis 7](#_Toc45546130)

[Figure 3: Graphical separation, conditional independence, and probability decomposition. 8](#_Toc45546131)

[Preliminary Results 9](#_Toc45546132)

[References 10](#_Toc45546133)

# Abstract

# Introduction

## DNA Methylation

Although all cells in an organism share the same DNA, it’s clear that not all genes are active at the same time in every cell. DNA and histone modifications that impact gene expression without altering the base pair sequence are collectively referred to as “epigenetics,” and are often inherited along with the raw genetic information. They are “fundamental to the regulation of many cellular processes” and “may be necessary for generating the large range of different phenotypes that arise from the same genotype.”1

Epigeneticists have long recognized the role of DNA cytosine methylation in repression of gene transcription and X chromosome inactivation. A cytosine followed immediately by a guanine nucleotide is called a CpG dinucleotide, and gene silencing is associated with regions of high CpG density. Certain transcription factors have been shown to be sensitive to DNA methylation (in one study, 22% of the 542 transcription factors examined demonstrated decreased binding to methylated motifs). Although mammalian genomes have relatively few CpG sites (~1%), about two thirds of mammalian promoters contain regions, called CpG islands (CGIs), of 200 base pairs (bp) or more in which 50% of the nucleotides (or more) are guanine or cytosine.2 So, DNA methylation can still have an outsized effect on expression despite low CpG density overall.

Methylation can also change heterochromatin structure through recruitment of DNMT proteins and chromatin remodelers, which in turn alters transcription by making it less likely that transcription factors will bind to DNA.3 Due to DNA methylation’s role in gene expression, abnormal methylation has been linked to many diseases including various cancers.1

### Figure 1: DNA Methylation Patterns

A screenshot of a video game

Description automatically generated

From Portela & Esteller (2010)1: Unmethylated CGIs (a) and CpG island shores (b) allow transcription factor binding which in turn leads to gene expression. Methylated sites prevent gene expression.

## Metabolomics

Metabolites are small molecule products of metabolism and are involved in many vital cellular processes including energy storage, cellular signaling and apoptosis, post-translational protein modification and transport, and maintenance of homeostasis in the cellular milieu. For example, “﻿metabolites such as ATP, acetyl-CoA, NAD+, and S-adenosyl methionine (SAM) can function as co-substrates, regulating post-translational modifications that affect protein activity” and “﻿fatty acids and hormones can interact with plasma proteins to enable their transport in the bloodstream.”4 Some metabolite-protein interactions also have a role in the initiation of signaling cascades.4 As a result of this far-reaching impact at multiple levels of human development, analysis of the metabolome can therefore quantify the integrated response to endogenous and exogenous disease factors or other physiological changes.5

In addition to these well-studied effects of the metabolome, there is mounting evidence that nutrition and metabolism also directly affect DNA methylation, and that the study of nutri-epigenetics may elucidate the role of diet in many diseases.6 S-adenosylmethionine (SAM) is the primary methyl group donor for enzymes that methylate nucleic acids and histones. Methylation of a substrate using this methyl group results in S-adenosylhomocysteine (SAH), another intermediate metabolite in the one-carbon pathway. Animal studies have shown that concentrations of SAM and SAH directly influence histone methylation in a way consistent with a signal transduction mechanism, meaning that “concentrations of SAM and SAH are on the order of the kinetic parameters that determine enzyme activity.”7 Also, human studies have confirmed that diet, specifically consumption of folate, choline, betaine, B vitamins and methionine, globally modifies methylation.6

### Figure 2: Micronutrients in One-Carbon Metabolism and DNA Methylation

A close up of a map

Description automatically generated

From Anderson et al. (2012)6: Substrates obtained via diet are highlighted in yellow.

## Type 1 Diabetes

Type 1 diabetes (T1D) is an autoimmune disease characterized by the production of antibodies which target pancreatic -cells. This leads to both micro and macrovascular complications, and individuals with T1D are at an increased risk of experiencing a major cardiac event (e.g. stroke, angina, and myocardial infarction) and are less likely to have satisfactory outcomes after a coronary event.8 The disease currently affects over 30 million people worldwide,9 and is increasing by 3-4% per year on average.10 However, the global burden of disease is difficult to estimate due to geographic variation in incidence.8,10

Genetic predisposition accounts for some of the etiology of T1D (sibling relative risk has been estimated at 15)11 and explains some geographic variation in incidence.10 Human leukocyte antigen (HLA) genes were the first to be linked to T1D and account for much of the genetic predisposition to the disease, but genome-wide association studies (GWAS) have also identified more than 40 other loci associated with increased relative risk of T1D. However, it is still unclear how these multiple loci interact with one another and the environment to eventually produce a T1D diagnosis. And in addition to the complex genetics of T1D, low monozygotic (MZ) twin concordance (approximately 50%) and studies of migrant populations support the theory that non-genetic factors play an important role in T1D development.9

Epigenetic differences may be important contributors to non-genetic T1D etiology. Changes in methylation have been associated with other autoimmune conditions, and monozygotic twins can be epigenetically heterogenous despite sharing identical genetic code.9 Rakyan et al. and Stefan et al. performed epigenome-wide association studies (EWAS) in discordant and concordant twin pairs and found that methylation profiles were more similar among participants with T1D than to unaffected twins. Epigenetics profiles were also combined with GWAS data and differentially expressed methylation sites were mapped to 6 well known T1D susceptibility genes, including two major histocompatibility complex (MHC) genes and several HLA loci.9,12 Finally, the Norris group has found that T1D cases had different longitudinal methylation patterns compared to controls prior to diagnosis.13

Environmental factors including viruses, diet, and the metabolome have also been linked with T1D etiology.8 Previous studies have found associations between T1D and differentially expressed phospholipids and sphingolipids, excretion of modified amino acids, and vitamin D (and related compounds on its metabolic pathway).5,8 Branched chain amino acids (BCAAs) have also been associated with insulin resistance.14

Integration of epigenetic and metabolomics data requires statistical methods capable of evaluating associations between many different variables in complex multilevel networks. Bayesian networks are becoming increasingly popular for analysis of large-scale systems biology (“omics”) data, including protein-protein interactions and gene regulatory networks.15 This approach can be used to generate intuitive graphical models, which represent probabilistic dependence between multiple variables15,16 and avoid many of the pitfalls of traditional mediation analyses.17 Also, probabilistic dependence relationships can be interpreted as causal pathways, which clarifies the biological interpretation of results. We hope to derive a graphical model of the relationships between metabolites and methylation sites, in order to illustrate and make inferences on the epigenetic and environmental pathways that lead to T1D.

# Specific Aims

## Primary Aim 1

Construct multi-Omics networks connecting T1D status, metabolite concentrations, and methylation levels at select CpG sites.

### Secondary Aim 1a

Compare the triad and SmCCNet approaches to feature reduction.

### Secondary Aim 1b

Compare BNA triad results from the “bnlearn” R package to likelihood-based causal mediation results from the “cit” R package.

## Primary Aim 2

Use BNA on reduced feature set to learn the structure of methylation and metabolomic pathways related to T1D. This final network structure will allow us to analyze the probabilistic dependence relationships between metabolites, methylation sites, and T1D.

## Primary Aim 3

Place the final network structure in biological context using known metabolomic and methylation pathways from previous research and public databases such as MethDB and the Human Metabolome Database (HMDB).

# Methods

## Methylation

DNA methylation was measured in peripheral whole blood, and study participants were split into two groups. One group was measured ﻿using the Infinium HumanMethylation450K Beadchip (“450 K”) and the other using ﻿the Infinium MethylationEPIC Beadchip (“EPIC”).13 Measurement platform will be included as a covariate in all linear models, in order to adjust for potential platform differences.

## Metabolomics

## Feature Reduction

First, we will select methylation-metabolite pairs by using linear models to correlate all combinations of probe and metabolite. We will also use logistic regression to find probes and metabolites that are significantly associated with the T1D phenotype. Candidate pairs that are strongly associated (nominal p-value < 0.001) and contain either a probe or metabolite (or both) that is significantly associated with the T1D phenotype (nominal p-value < 0.05) will continue on to the next step. We will not adjust these p-values for multiple comparisons because doing is complicated with many correlated variables, and there are no agreed upon best practices.18 However, careful network development with strict thresholds in later steps should prevent false positives in the final results, as described in Rudra et al.

Metabolite-methylation pairs fitting the above criteria will be further pared down using causal inference testing. The R package “cit” performs multiple statistical tests associated with causal mediation, and generates an omnibus p-value for the component tests.19 For example, let L represent a methylation probe, G represent a metabolite, and T represent the outcome of T1D. To test for causal relationships, the package fits three linear regression models (with i indexing subject):

And tests four null hypotheses:

The omnibus p-value is simply the highest p-value from the four tests above,20 and if all four null hypotheses are rejected (i.e. an omnibus p-value < 0.05) the methylation-metabolite pair will move on to the Bayesian network analysis.

## Bayesian Network Analysis

A graph consists of a non-empty set of vertices (or nodes) and a finite set

of pairs of vertices called arcs, links, or edges (unlike , this set can be empty). If arc is an ordered pair, the arc is said to be directed from to and is usually represented by an arrow of the form . Graphs themselves can also be described as directed if all arcs are directed.21 Sequences of arcs form a path from one vertex to another, and imply that the first node is a cause the and the other an effect.22 Bayesian networks are a specific case of graphical models, where each arc represents a probabilistic dependency between two variables , which are represented as nodes . The graphs must be directed and acyclic (i.e. no path of the form exists).21

Graphical separation , or the absence of an arc between two given nodes, can be linked to probabilistic separation in three ways. If each node corresponds to random variable and *P* represents the probabilistic dependence structure of , then graph G is an independence map (I-map) if:21

For all disjoint subsets A, B, and C of . Graph G is a dependence map (D-map) of P if:

G is called a perfect map if it is both an I-map and D-Map:

### Figure 3: Graphical separation, conditional independence, and probability decomposition.

A screenshot of a cell phone

Description automatically generated

From Nagarajan et al. (2013).

The formal definition of a Bayesian network is a directed acyclic graph (DAG) denoted if and only if G is a minimal I-map of P such that not a single arc can be removed without making G no longer an I-map. From this definition follow two important properties. First, the global distribution P(**X**) can be decomposed into:

Where is the set of parent nodes of . From this decomposition follows the local Markov property of BNs, which states that “each node is conditionally independent of its non-descendants (e.g. nodes for which there is no path from to ) given its parents.”22

The first step of fitting a Bayesian model is often called structure learning, and it involves trying to find the best (preferably the minimal I-map) graph structure. There are several popular methods for learning network structures, and these can be grouped into three categories: constraint-based, score-based, and hybrid structure learning algorithms.

For each pair selected by the causal inference testing, we will use a score-based algorithm to find the best network structure using the probe-metabolite-phenotype triad. With no additional restrictions on edge number or direction (aside from the assumptions of Bayesian networks), there are 24 possible network structures for each triple. We will compare all possible network structures using a Deviance information criterion (DIC) score, and the structure with the best (lowest) score will continue on to the next selection step. DIC can be considered a Bayesian equivalent to Akaike’s information criterion, a measure commonly used for model comparison in frequentist statistics. It is calculated as the posterior mean deviance (a measure of fit) plus double the effective number of parameters in the model (a measure of model complexity):23

Where

Based on the best networks from comparison of DIC, we will somehow build a larger network…

# Preliminary Results

# References

1. Portela A, Esteller M. Epigenetic modifications and human disease. *Nat Biotechnol*. 2010;28(10):1057-1068. doi:10.1038/nbt.1685

2. Gardiner-Garden M, Frommer M. CpG Islands in vertebrate genomes. *J Mol Biol*. 1987;196(2):261-282. doi:10.1016/0022-2836(87)90689-9

3. Greenberg MVC, Bourc’his D. The diverse roles of DNA methylation in mammalian development and disease. *Nat Rev Mol Cell Biol*. 2019;20(10):590-607. doi:10.1038/s41580-019-0159-6

4. Johnson CH, Ivanisevic J, Siuzdak G. Metabolomics: Beyond biomarkers and towards mechanisms. *Nat Rev Mol Cell Biol*. 2016;17(7):451-459. doi:10.1038/nrm.2016.25

5. Frohnert BI, Rewers MJ. Metabolomics in childhood diabetes. *Pediatr Diabetes*. 2016;17(1):3-14. doi:10.1111/pedi.12323

6. Anderson OS, Sant KE, Dolinoy DC. Nutrition and epigenetics: An interplay of dietary methyl donors, one-carbon metabolism and DNA methylation. *J Nutr Biochem*. 2012;23(8):853-859. doi:10.1016/j.jnutbio.2012.03.003

7. Mentch SJ, Mehrmohamadi M, Huang L, et al. Histone Methylation Dynamics and Gene Regulation Occur through the Sensing of One-Carbon Metabolism. *Cell Metab*. 2015;22(5):861-873. doi:10.1016/j.cmet.2015.08.024

8. Atkinson MA, Eisenbarth GS, Michels AW. Type 1 diabetes. *Lancet*. 2014;383(9911):69-82. doi:10.1016/S0140-6736(13)60591-7

9. Rakyan VK, Beyan H, Down TA, et al. Identification of type 1 Diabetes-associated DNA methylation variable positions that precede disease diagnosis. *PLoS Genet*. 2011;7(9):1-9. doi:10.1371/journal.pgen.1002300

10. Tuomilehto J. The emerging global epidemic of type 1 diabetes. *Curr Diab Rep*. 2013;13(6):795-804. doi:10.1007/s11892-013-0433-5

11. Polychronakos C, Li Q. Understanding type 1 diabetes through genetics: Advances and prospects. *Nat Rev Genet*. 2011;12(11):781-792. doi:10.1038/nrg3069

12. Stefan M, Zhang W, Concepcion E, Yi Z, Tomer Y. DNA methylation profiles in type 1 diabetes twins point to strong epigenetic effects on etiology. *J Autoimmun*. 2014;50:33-37. doi:10.1016/j.jaut.2013.10.001

13. Johnson RK, Vanderlinden LA, Dong F, et al. Longitudinal DNA methylation differences precede type 1 diabetes. *Sci Rep*. 2020;10(1):1-13. doi:10.1038/s41598-020-60758-0

14. Guasch-Ferré M, Hruby A, Toledo E, et al. Metabolomics in prediabetes and diabetes: A systematic review and meta-analysis. *Diabetes Care*. 2016;39(5):833-846. doi:10.2337/dc15-2251

15. Shaddox E, Stingo FC, Peterson CB, et al. A Bayesian Approach for Learning Gene Networks Underlying Disease Severity in COPD. *Stat Biosci*. 2018;10(1):59-85. doi:10.1007/s12561-016-9176-6

16. Sachs K, Perez O, Pe’er D, Lauffenburger DA, Nolan GP. Causal protein-signaling networks derived from multiparameter single-cell data. *Science (80- )*. 2005;308(5721):523-529. doi:10.1126/science.1105809

17. Rudra P, Shi WJ, Russell P, et al. Predictive modeling of miRNA-mediated predisposition to alcohol-related phenotypes in mouse. *BMC Genomics*. 2018;19(1):1-12. doi:10.1186/s12864-018-5004-3

18. Conneely KN, Boehnke M. So many correlated tests, so little time! Rapid adjustment of P values for multiple correlated tests. *Am J Hum Genet*. 2007;81(6):1158-1168. doi:10.1086/522036

19. Millstein J, Chen GK, Breton C V. Cit: Hypothesis testing software for mediation analysis in genomic applications. *Bioinformatics*. 2016;32(15):2364-2365. doi:10.1093/bioinformatics/btw135

20. Millstein J, Zhang B, Zhu J, Schadt EE. Disentangling molecular relationships with a causal inference test. *BMC Genet*. 2009;10:1-15. doi:10.1186/1471-2156-10-23

21. Nagarajan R, Scutari M, Lèbre S. *Bayesian Networks in R*.; 2013. doi:10.1007/978-1-4614-6446-4

22. Scutari M, Denis J-B. *Bayesian Networks*.; 2014. doi:10.1201/b17065

23. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *J R Stat Soc Ser B Stat Methodol*. 2002;64(4):583-616. doi:10.1111/1467-9868.00353