Title

Statistical and Computational Methods in Epidemiological and Pharmacogenomic Studies: from Application to Method Development

Author

Chi, Calvin

Publication Date

2020

Peer reviewed|Thesis/dissertation

Statistical and Computational Methods in Epidemiological and Pharmacogenomic Studies: from Application to Method Development

by

Calvin Chi

A dissertation submitted in partial satisfaction of the

requirements for the degree of

Doctor of Philosophy

in

Computational Biology

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Haiyan Huang, Co-chair
Professor Lisa Barcellos, Co-chair
Associate Professor Nir Yosef
Associate Professor Maya Petersen

Summer 2020

Statistical and Computational Methods in Epidemiological and Pharmacogenomic Studies: from Application to Method Development

Abstract

Statistical and Computational Methods in Epidemiological and Pharmacogenomic Studies: from Application to Method Development

by

Calvin Chi

Doctor of Philosophy in Computational Biology

University of California, Berkeley

Professor Haiyan Huang, Co-chair

Professor Lisa Barcellos, Co-chair

Statistical and computational methods are seeing a growing role in genetic epidemiology and pharmacogenomics. Genetic epidemiology is the study of the interplay between genetic and environmental factors on human health and disease in populations. Pharmacogenomics is the study of how genes affect response to drugs. Chapter 2 illustrates an admixture mapping study of multiple sclerosis from local ancestry estimates provided by a linear-chain random conditional field. Chapter 3 shows the application of regression-based methods and causal inference principles to study the relationship between genotype and DNA methylation of human labial salivary glands in Sjögren's syndrome. Chapter 4 applies variational autoencoder to perform dimensionality reduction on DNA methylation data for discovery of clinically-relevant disease subtypes in Sjögren's syndrome. Chapter 5 applies sparse canonical correlation analysis to summarize gene expression-drug sensitivity associations and introduces a nuclear norm-based dissimilarity measure to compare associations from different cell line groups in pharmacogenomic studies. Finally, Chapter 6 presents a 1D convolutional neural network model for imputing human leukocyte antigen alleles from phased genotype data.

To my friends and family

Contents

List of Figures

List of Tables

Acknowledgments

I would like to start by thanking Kate Chase, the program coordinator, who encouraged me to apply to the Computational Biology program when I had doubts that came from transitioning from a basic sciences field to a computational one. Kate has run the Computational Biology program very well during these five years and has always been helpful via emails.

I would like to thank Brooke Rhead from my program, who provided me with guidance during my first few years in the Barcellos lab. I would like to thank my co-advisor Lisa Barcellos for her supervision of my research, encouragements, provision of the lab space, and general supportiveness. Others from the Barcellos lab whom I worked with closely at various stages of my PhD years include Xiaorong Shao, Indro Fedrigo, Hong Quach, and Diana Quach, all of whom were supportive and friendly. Lastly, I would like to thank Lindsey Criswell for the collaboration on a research project on sjögren's syndrome.

I would like to acknowledge the National Science Foundation for funding the last three years of my PhD training and acknowledge Sandrine Dudoit, Lisa Barcellos, and Carmay Lim for writing my letters of recommendation for the fellowship. I would also like to thank Gaston Sanchez, whom I served as a graduate student instructor, for the pleasant collaboration in teaching STAT20.

For my qualifying exam, I would like to thank faculties Lisa Barcellos, Maya Petersen, Nir Yosef, Jack Colford, Sandrine Dudoit, and Art Reingold for the support and participation. Thank you for helping me prepare and for making the day of the exam a pleasant experience. In addition to those mentioned, I would like to thank my peers Cam Adams, Shannon McCurdy, Jonathan Levy, and Wilson Cai for helping me prepare for the exam.

In my third year, I joined Haiyan Huang's group to engage in applied statistics research. I want to thank Yuting Ye and Haiyan Huang for working together on a research project. Haiyan has been a very accessible and hands-on advisor, who was even willing to hold late evening calls when I was away for my summer internship so that I could continue to make progress on my research. I want to thank Peter Bickel, whom I got to know through lab meetings, for taking an interest in my career development and for the casual chats in Evans.

I next want to thank my graduate school friends I met during my PhD journey, some of whom go to other schools - Brian Cheung, Benjamin Chu, Shi-Zhuo Looi, Ben Wormleighton, Juwon Lee, Emily Chung, Alvin Li, Nichole Luk, and Simon Chu. Your company has made my experience more pleasant. I enjoyed discussing tips for academic success and the occasional "duels" in select subject areas. I want to thank Josiah Davis, whom I met through causal inference class, for the successful referral to an internship, which jump-started my industry career. I want to thank Anthony Cao for your friendship and support during our time at Berkeley together, and Eddie Massey for encouraging me to persevere over some long phone calls.

Lastly, I would like to thank my parents for their persistent support, from supporting my decision to leave the medical profession to pushing me to persevere through the most difficult of times of my PhD journey. I can always count on you for having my best interests at heart.

Chapter 1

Introduction

This dissertation presents both the application and development of statistical and computational methods for genetic epidemiology and pharmacogenomics. Genetic epidemiology involves studying the interplay between genetic factors and environmental factors on human health and disease in populations. Pharmacogenomics is the study of how genes affect or are associated with people's response to drugs. Chapters 2 - 4 illustrate applying methods to answer epidemiological or biological questions concerning multiple sclerosis (MS) and Sjögren's syndrome (SS), and chapters 5 - 6 focus on developing methods for cancer pharmacogenomics and human leukocyte antigen (HLA) allele imputation respectively. Both classical and modern methods are encountered in this dissertation. Given the diversity of subjects present, Chapter 1 is devoted to introducing relevant background for genetic epidemiology and pharmacogenomics, as well as the main statistical and computational methods involved.

1.1 Admixture Mapping

When the prevalence of a genetic disease varies by ethnicity, it could suggest origination of the causal genetic variants from a particular ancestral population. One way to investigate this is to perform admixture mapping, which infers local genetic ancestry in admixed populations and tests if risk variants are predominantly of one ancestry in cases compared to controls (Figure 1.1). Significant association of ancestry with a phenotype is interpreted as evidence of ancestry-specific phenotype, provided the phenotype is genetic. Admixture occurs when individuals from two or more previously isolated ancestral populations interbreed. While most allele frequencies remain relatively constant between populations, some differ substantially by population [1]. Major admixed populations in the United States include African Americans, Hispanics, and Asian Americans.

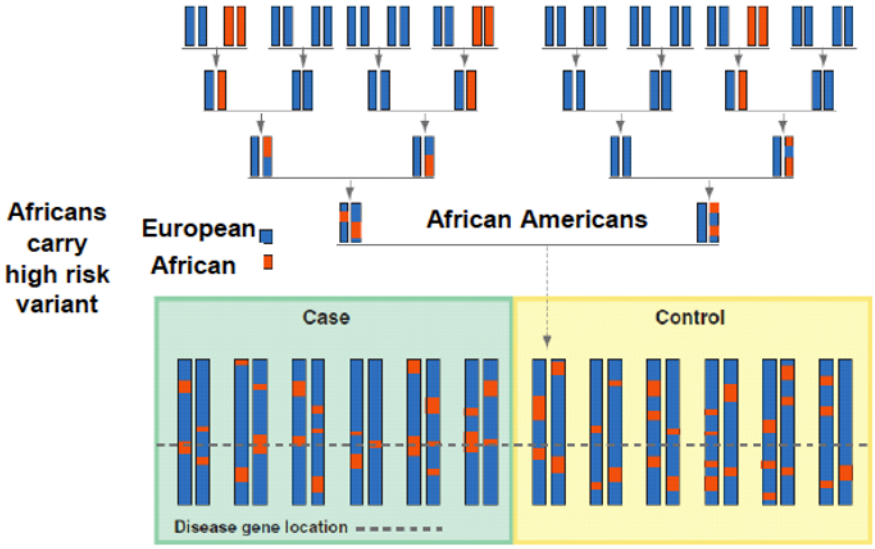


Figure 1.1: Admixture mapping study design [2]. The top half illustrates the admixture across 4 generations between African and European ancestry. The bottom half illustrates the study design of admixture mapping between cases and controls.

An admixture mapping study requires

- Reference panel containing genotype data from populations of known ancestry.
- Genotype data of target admixed population, whose local ancestry is to be inferred.
- Phenotype status of individuals from the target population.

Every admixture mapping study needs to control for population stratification, which is the systemic difference in allele frequencies between ancestral populations. Population stratification can cause confounding, where differential risk of an allele by ancestry could be due to a systematic difference in ancestry between cases and controls. Chapter 2 provides a published admixture mapping study of MS [3], which has highest prevalence in White, non-Hispanic populations.

Genotype Data

For the purpose of this dissertation, an individual's genotype data is the number of reference alleles at select loci across the genome, where the reference and alternate alleles at each locus are predefined. The genotype g_l at locus l can thus take on values of 0, 1, or 2. Genotype data is obtained from genotyping chips, which differ in the number of single nucleotide polymorphism (SNPs) genotyped and the selection of loci for which genotyping is to be performed. Typically, genotyping chips select "tag" SNPs that is representative of the

surrounding genetic region by linkage disequilibrium (LD) with surrounding SNPs, where LD is defined as the non-random association of alleles at nearby loci.

Ancestry Inference

Local ancestry inference involves inferring the ancestry of both alleles at select genetic loci, using a panel of reference genotypes of known ancestry. Methods for local ancestry estimation include HAPMIX [4], LAMP [5], LAMP-LD [6], LAMP-HAP [6], ELAI [7], and RFMix [8]. For this dissertation's admixture mapping study, we choose RFMix, one of the state-of-the-art methods for local ancestry inference, which has been shown to out-perform LAMP-LD and LAMP-HAP [8]. RFMix uses a conditional random field parameterized by random forests to infer local ancestry in phased admixed haplotypes using a panel of reference haplotypes.

Ancestry inference is preceded by phasing, which is the process to assigning alleles to paternal and maternal chromosomes from genotype data. This is commonly achieved with BEAGLE, a software that implements a hidden markov model capable of both phasing and allele imputation [9]. It has been shown that denser SNPs from imputation can increase accuracy of ancestry estimation [10].

1.2 DNA Methylation

DNA methylation refers to the methylation of cytosine in CpG sites, which are regions of DNA where a cytosine nucleotide is followed by a guanine nucleotide in the $5' \rightarrow 3'$ direction. DNA methylation is one of the mechanisms of epigenetic modification, where epigenetics is the study of changes in phenotype or gene expression without changes in DNA sequence. DNA methylation is generally associated with transcription silencing, although the underlying mechanisms are not necessarily identical at gene promoters, gene bodies or repeated sequences [11]. Regions of accessible chromatin are frequently lowly methylated or unmethylated, suggesting that transcription binding and DNA methylation are mutually exclusive [12]. In other contexts, such as within the gene-body, DNA methylation is positively correlated with transcription [13–15]. Hence, the relationship between DNA methylation and gene expression is complex. The phenotypic effects of DNA methylation regulation are profound, affecting Mammalian development [16, 17], cell type-specific gene regulation [18], and cancer [19].

The study of DNA methylation is of epidemiological interest for its value as biomarkers for disease or prognosis, involvement in disease mechanisms, relationship to environmental exposure, regulation of response to medication, etc. Specifically, DNA methylation is shaped by both DNA variation [20] and environmental exposure [21–26].

DNA methylation is measured for each CpG site as the ratio of fluorescent intensity from methylated alleles (M) to total fluorescent intensity from methylated alleles and unmethylated alleles (U), summarized as the β value in Equation 1.1.

$$\beta = \frac{\max(M, 0)}{\max(M, 0) + \max(U, 0) + \alpha} \tag{1.1}$$

Typically, $\alpha = 100$ in Equation 1.1. The β value ranges from 0 to 1, where 1 reflects complete methylation and 0 reflects no methylation. Depending on the analysis, some studies prefer to work with M-values of DNA methylation, which can be mapped from β values according to Equation 1.2.

$$M = \log_2 \frac{\beta}{1 - \beta} \tag{1.2}$$

M-values range from $-\infty$ to ∞. Although β-values are biologically interpretable, M-values are approximately homoscedastic and thus recommended for differential methylation analysis [27]. The experimental steps for a methylation assay are

1. Bisulfite treatment: treating denatured genomic DNA with sodium bisulfite, which deaminates unmethylated cytosines into uracils, while methylated cyosines remain unchanged.

2. Whole-genomic DNA amplication of bisulfite treated DNA.

3. Hybridization and single-base extension: bisulfite-converted amplified DNA are denatured and binds to the methylation-specific probe or non-methylation probe for each CpG site. Single-base extension is performed with hapten-labeled dideoxynucleotides for immunohistochemical assays.

4. Fluorescence staining: immunohistochemical assays are performed to show the intensities of the unmethylated and methylated fluorescent signals.

The above steps are depicted in Figure 1.2. Differential methlyation studies typically compare DNA methylation profiles between different experimental conditions, but in similar tissues due to the tissue specificity of DNA methylation [18, 28–31].

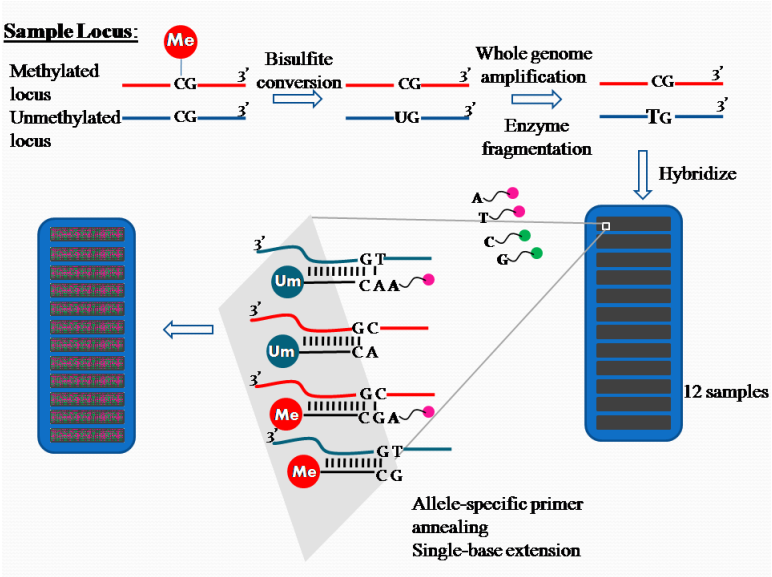


Figure 1.2: Workflow of the Infinium I assay for DNA methylation measurement [32].

DNA methylation data requires pre-processing before data analysis. Pre-processing steps typically correct for batch effects, remove CpG sites with cross-reactive probes [33–35], and remove CpG sites with SNPs in probes [36]. Cross-reactive probes and SNPs in probes decrease binding specificity and lead to spurious methylation signals. Following pre-processing, it is important to control for confounders such as age [37], smoking [38–41], and ancestry when analyzing DNA metyhlation data [42]. Chapters 3-4 present a study of the relationship between genotype, DNA methylation, and clinical phenotypes in SS.

1.3 Pharmacogenomics

Pharmacogenomics is the study of how an individual's genomic profile affects response to drugs, and is a branch of precision medicine. Although not always resembling their corresponding tissue *in vivo* [43], cancer cell lines have been the most widely used pre-clinical models to study the molecular basis of drug response [44]. Multiple datasets have been generated to that end, such as the NCI-60 project [45], the Cancer Cell Line Encyclopedia (CCLE) [46], the Genomics of Drug Sensitivity in Cancer [47, 48], and the Connectivity Map [49]. Along with profiling the drug response of cell lines across a panel of drugs, these studies either profile the molecular profiles of untreated cell lines from disease tissue, or profile molecular profiles of cell lines before and after treatment.

A pharmacogenomic dataset comprises of

- Molecular profiles: gene expression, DNA copy number, genotype, methylation, proteomics, etc.

- Drug sensitivity measurements.

Drug sensitivity is commonly measured in terms of cell line growth inhibition as a function of drug concentration. The smaller the concentration required to inhibit growth, the more sensitive the cell line is to the drug. Two popular measures of drug sensitivity are IC_{50} and area over the activity curve measuring dose response (Figure 1.3).

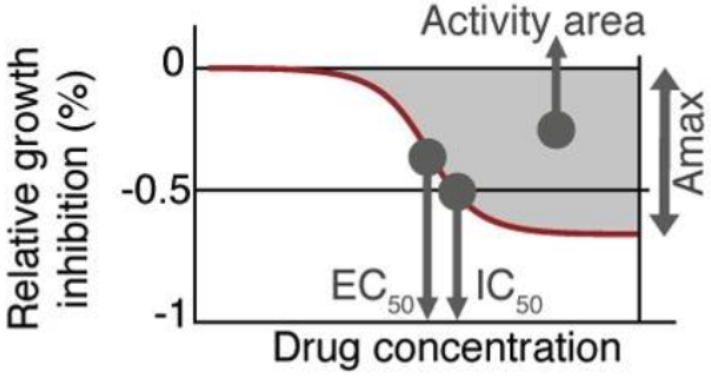


Figure 1.3: Dose-response curve. IC_{50} is the drug concentration required to inhibit 50% of cell line growth. Activity area (shaded in gray) is the area over the activity curve measuring growth inhibition as function of drug concentration [47].

Pharmacogenomic studies broadly fall into the categories of (1) identification of molecular features as drug-response associated biomarkers or (2) prediction of drug response from molecular profiles. Challenges to pharmacogenomic data analyses include heterogeneity of investigated cell lines, assay technologies, compounds screened, and type and quality of genomic data [44]. As a result, datasets generated from different studies are often not concordant. Another challenge is inconsistency in drug sensitivity measurements for the same cell line and drug across different studies [50]. Chapter 5 presents a statistical method for addressing the challenge of biomarker discovery in a heterogeneous set of cell lines.

1.4 Statistical and Computational Methods

Principal Component Analysis

Principal component analysis (PCA) is a method for finding a subspace in which the data approximately lies. In computational biology, PCA is often used to visualize genetic similarity between populations and as a way to control for genetic ancestry in epidemiological studies. Formally, given data $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$, PCA finds a subspace for the data $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)} \in \mathbb{R}^k$ with $k << n$ such that the original data is approximated well.

In PCA, this is achieved by projecting the data onto a subspace that optimally describes variance of the data. We start our derivation by finding the a vector $u : ||u||_2 = 1$ to project $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ onto such that the variance of the projections is maximized. The projection of $x \in \mathbb{R}^n$ onto u is defined in Equation 1.3.

$$proj_u x = \frac{x^\top u}{||u||_2} = u^\top x \tag{1.3}$$

The mean of the projections is $u^\top \hat{x}$, where $\hat{x} = \frac{1}{m}\sum_i x^{(i)}$. We can derive the variance of projections in terms of the symmetric covariance matrix $C \in \mathbb{R}^{n\times n}$ starting from the definition.

$$\begin{aligned}
\sigma^2 &= \frac{1}{m}\sum_{i=1}^{m}\left[u^\top(x_i - \hat{x})\right]^2 \\
&= \frac{1}{m}\sum_{i=1}^{m} u^\top(x_i - \hat{x})(x_i - \hat{x})^\top u \\
&= u^T\left[\frac{1}{m}\sum_{i=1}^{m}(x_i - \hat{x})(x_i - \hat{x})^\top\right]u \\
&= \frac{1}{m}u^\top C u
\end{aligned} \tag{1.4}$$

Now we finally solve for the vector $u : ||u||_2 = 1$ to project $x^{(1)}, \ldots, x^{(m)}$ onto such that the variance of the projections is maximized as a convex optimization problem

$$\max_{u:||u||_2=1} u^\top C u = \max_{u:||u||_2=1} u^\top U D U^\top u \tag{1.5}$$

$$= \max_{w:||w||_2=1} w^\top D w \tag{1.6}$$

$$= \max_{w:||w||_2=1} \sum_i \lambda_{ii} w_i^2, \tag{1.7}$$

where λ_{ii} is the ith largest eigenvalue from the diagonal matrix $D \in \mathbb{R}^{n\times n}$. The solution to Equation 1.7 is $w^* = e_1$, which is a vector of zeros except with a 1 in the first entry. This implies that $u^* = Uw^* = u_1$, the first eigenvector of C. The maximum objective value for σ^2 is then λ_{11}, the largest eigenvalue of C, with first principle axis u^*. The projections of $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ onto the first principle axis is called the first principal component. More generally, the ith principal component is found by projecting the data onto the vector that yields the ith maximum variance. From Equation 1.4, these vectors are the eigenvectors of C. In genetic epidemiology, principal components are often conditioned on in regression

models to control for genetic ancestry. An interpretation from PCA is that each original sample $x^{(i)} \in \mathbb{R}^n$ is a linear combination of the principal axes $u_1, \ldots, u_n \in \mathbb{R}^n$, which is expressed in Equation 1.8.

$$X = XUU^\top = XU \begin{bmatrix} u_1^\top \\ \vdots \\ u_n^\top \end{bmatrix} \tag{1.8}$$

Another popular dimensionality reduction technique is called multidimensional scaling (MDS), which has the same goal of finding $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)} \in \mathbb{R}^k$ from $x^{(1)}, \ldots, x^{(m)} \in \mathbb{R}^n$ where $k << n$. However, MDS does so by minimizing the sum of difference in pairwise distances between samples in Equation 1.9

$$\mathcal{L}(D, \hat{D}) = \sum_{i<j} (D_{ij} - \hat{D}_{ij})^2, \tag{1.9}$$

where $D \in \mathbb{R}^{m \times m}$ is the distance matrix from $x^{(1)}, \ldots, x^{(m)}$ and $\hat{D} \in \mathbb{R}^{m \times m}$ is the distance matrix from $\tilde{x}^{(1)}, \ldots, \tilde{x}^{(m)}$. When the dissimilarities are given by Euclidean distance, PCA and MDS are equivalent [51].

Logistic Regression

Logistic regression learns $\theta \in \mathbb{R}^{n+1}$ for the function in Equation 1.10

$$h(x; \theta) = g(\theta^\top x) = \frac{1}{1 + e^{-\theta^\top x}}, \tag{1.10}$$

with the mapping $h : \mathcal{X} \to \mathcal{Y}$, where $\mathcal{Y} \in \{0, 1\}$ and $x \in \mathbb{R}^n$. The learned weights $\hat{\theta}$ can be found as the weights maximizing the log likelihood of the data, using either gradient descent or Newton's method. Logistic regression has been ubiquitously used in association studies in genetic epidemiology, evaluating the association between a SNP with a binary phenotype. Specifically, the exponential of the SNP coefficient has the interpretation of being the odds ratio. To see this, let $p(x; \theta) = h(x; \theta)$, and start with the definition of odds

$$\frac{p(x^{(i)}; \theta)}{1 - p(x^{(i)}; \theta)} = \frac{\frac{e^{\theta^\top x^{(i)}}}{e^{\theta^\top x^{(i)}} + 1}}{\frac{1}{e^{\theta^\top x^{(i)}} + 1}} = e^{\theta^\top x^{(i)}}. \tag{1.11}$$

Using the result from Equation 1.11, the odds ratio (OR) between two sets of covariates $x^{(i)}$ and $x^{(k)}$ is

$$OR = \frac{p(x^{(i)};\theta)/(1-p(x^{(i)};\theta))}{p(x^{(k)};\theta)/(1-p(x^{(k)};\theta))} = \frac{e^{\theta^\top x^{(i)}}}{e^{\theta^\top x^{(k)}}} = e^{\theta^\top (x^{(i)}-x^{(k)})}. \tag{1.12}$$

If $x^{(i)} - x^{(k)} = [0, \cdots, 0, 1, 0, \cdots, 0]$, then $OR = e^{\theta_p}$ for some p. In other words, if $x_j^{(i)} = x_j^{(k)}$ for every $j \neq p$ and $x_p^{(i)} - x_p^{(k)} = 1$, then e^{θ_p} has the interpretation of being the odds ratio when x_p is increased by 1, holding all other variables the same.

Since $\hat{\theta}$ is the maximum likelihood estimate of θ, asymptotically $\hat{\theta}$ has the multivariate normal distribution in Equation 1.13

$$\hat{\theta} \xrightarrow{d} \mathcal{N}(\theta, (I_n^{obs}(\hat{\theta}))^{-1}), \tag{1.13}$$

from which we could conduct inference by testing $H_0 : \theta_i = 0$ versus $H_1 : \theta_i \neq 0$. In Equation 1.13, $I_n^{obs}(\hat{\theta})$ is the observed Fisher information from n samples. Let $X \in \mathbb{R}^{n\times(p+1)}$ be the data matrix with samples $x^{(1)}, \ldots, x^{(n)}$, then

$$I_n^{obs}(\hat{\theta}) = X^\top diag\left\{ p(x^{(1)};\hat{\theta})(1-p(x^{(1)};\hat{\theta})) \right\}_{i=1}^{n} X. \tag{1.14}$$

The $100(1-\alpha)\%$ confidence interval (CI) for the odds ratio of $x_i = c$ versus $x_i = c+1$ is then

$$100(1-\alpha)\%\ \text{CI} = \exp\left(\hat{\theta}_i \pm z_{1-\alpha/2} \times \hat{se}(\hat{\theta}_i)\right) \tag{1.15}$$

where $z_{1-\alpha/2}$ is the standard score. In genetic epidemiology, the possible conclusions regarding an allele based on a $1-\alpha$ OR confidence interval are

- Not significantly associated with case status: confidence interval contains 1.
- Protective allele: confidence interval lies below 1.
- Risk allele: confidence interval lies above 1.

RFMix: Conditional Random Field

RFMix is a discriminative method for local ancestry estimation that uses a conditional random field (CRF) parameterized by random forests trained on a reference panel of haplotypes [8]. It takes as input phased admixed and reference haplotypes and outputs local ancestry

estimation for each locus in the data. Let Y represent the random vector of local ancestry estimates and X be the phased admixed and reference haplotypes, then RFMix models $P(Y \mid X)$.

RFMix takes as input the genetic location in centimorgans (cM)[1] of each SNP and uses this genetic location to divide all haplotypes into W contiguous disjoint windows such that the maximum distance between all SNPs in any window is at most d cM. Assume R ancestries are present in N reference haplotypes, then $H \in \mathbb{R}^{N \times W}$ is a matrix where $H_{i,j}$ represents the sequence of alleles of haplotype i in window j. Thus, $H_{i,j}$ can be expanded to $H_{i,j}^{(1)}, H_{i,j}^{(2)}, ..., H_{i,j}^{(s_j)}$, where s_j represents the total number of alleles in window j. The local ancestries can be represented as $A \in \mathbb{R}^{N \times W}$, where $A_{i,j}$ represents the local ancestry of haplotype i in window j.

RFMix employs a linear-chain CRF to model $P(A \mid H)$ as an undirected graph of feature functions and identifies $\hat{A}$ to maximize $P(A \mid H)$. Independence is encoded by the graph and allows $P(A \mid H)$ to be written as factorized terms. Feature functions in CRF are functions that capture the relationship between dependent variables and maps to a real-valued output. The feature function itself has no probabilistic interpretation but is used to model $P(A \mid H)$.

Specifically, let $H_{i,*}$ and $A_{i,*}$ represent the alleles and local ancestry assignments of haplotype i across all windows. The CRF models $P(A_{i,*} \mid H_{i,*} : \Theta)$ according to Equation 1.16,

$$\begin{aligned} P(A_{i,*} \mid H_{i,*} : \Theta) = \frac{1}{Z(H_{i,*})} \exp \Bigg\{ & \sum_{w=1}^{W} \sum_{r=1}^{R} \sum_{h \in \mathcal{H}_w} \theta^A_{w,r,h} \mathbb{1}(A_{i,w} = r) \mathbb{1}(H_{i,w} = h) \\ & + \sum_{p=1}^{W-1} \sum_{j=1}^{R} \sum_{k=1}^{R} \theta^T_{p,j,k} \mathbb{1}(A_{i,p} = j) \mathbb{1}(A_{i,p+1} = k) \Bigg\} \end{aligned} \tag{1.16}$$

where $\mathcal{H}_w$ is the set of haplotypes in window w and $Z(H_{i,*})$ represents the partition function (normalization constant) summed over all $A_{i,*}$ to ensure $P(A_{i,*} \mid H_{i,*} : \Theta) \in [0, 1]$. The parameters θ^T and θ^A represent model parameters to be learned, which are defined in Equation 1.17.

$$\begin{aligned} \theta^A_{w,r,h} &= \ln(P(A_{i,w} = r \mid H_{i,w} = h)) \\ \theta^T_{p,j,k} &= \ln(P(A_{i,p} = j, A_{i,p+1} = k)) \end{aligned} \tag{1.17}$$

The feature functions are $\mathbb{1}(A_{i,w} = r)\mathbb{1}(H_{i,w} = h)$ and $\mathbb{1}(A_{i,p} = j)\mathbb{1}(A_{i,p+1} = k)$. The CRF can be represented graphically in Figure 1.4.

[1] A unit of centimorgan between two chromosomal positions represents the expected number of crossovers of 0.01 in a single generation.

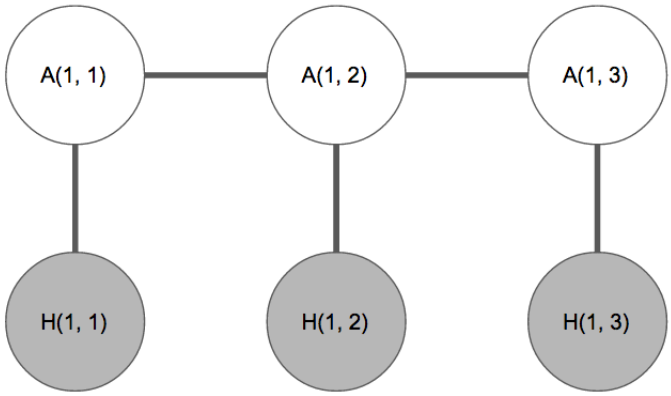


Figure 1.4: Linear-chain CRF. $A(i, j)$ and $H(i, j)$ represent the ancestry and allele sequence for haplotype i at window j respectively.

The parameter θ^A is obtained from training a random forest on reference haplotypes for each window, mapping SNPs to local ancestry. The parameter θ^T is derived from the joint probability distribution of the admixture model described by Falush *et al* [52].

$$P(A_{i,p} = j, A_{i,p+1} = k) = \begin{cases} q_j(\exp(-d_p G) + (1 - \exp(-d_p G))q_k) & \text{if } j = k \\ q_j(1 - \exp(-d_p G))q_k & \text{otherwise} \end{cases} \tag{1.18}$$

In Equation 1.18, where q_j is the proportion of ancestry j in the admixed population, G is the number of generations since admixture, and d_p is the distance between the middle windows p and $p + 1$.

Once initial estimates for θ^A and θ^T are obtained. Entries $A_{i,p+1}$ can be determined via dynammic programming, with the following recurrence relation in Equation 1.19.

$$A_{i,p+1} = \arg\max_k \left\{ \theta^A_{p,r,h} \mathbb{1}(A_{i,p} = r)\mathbb{1}(H_{i,p} = h) + \theta^T_{p,r,k} \mathbb{1}(A_{i,p} = r)\mathbb{1}(A_{i,p+1} = k) \right\} \tag{1.19}$$

With initial estimates for θ^T, θ^A, and local ancestries made, these estimates could be iteratively improved by using the expectation-maximization (EM) algorithm. This approach has the advantage of incorporating our admixed haplotypes to infer θ^T and θ^A in other admixed haplotypes.

In the maximization step, θ^A is updated by re-training random forests for each window. In each window, the set of chromosomes are divided into b sets such that each set has approximately the same number of haplotypes assigned to each ancestry. Given a window,

a random forest is trained on haplotypes from $b-1$ sets and used to update θ^A for the remaining set. This is repeated b times. The parameters θ^T is fixed and do not need to be updated. In the expectation step, the updated parameter θ^A is used to infer local ancestry by Equation 1.19.

Causal Inference

Causal inference is a statistical, and to a certain extent philosophical, discipline concerned with inferring the causal relationship between variables. A causal relationship is said to exist between A and Y when changing A would lead to a change in Y. The causal relationship has also been thought about as the difference in potential outcomes under one condition $Y(1)$ versus another condition $Y(0)$.

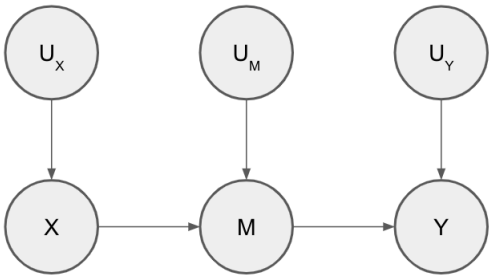
Following the causal inference framework from Judea Pearl [53], we introduce the basics of specifying a structural causal model (SCM) and the back-door criterion for determining which set of variables to condition on for the treatment effect between treatment A and outcome Y. A SCM is specified by the following

- **Endogenous variables** $X = \{X_1, \ldots, X_J\}$: variables meaningful for the study which could affect one another.
- **Background variables** $U = \{U_1, \ldots, U_J\}$: unmeasured variables not affected by variables in X but affects variables in X.
- **Functions** $F = \{f_{X_1}, \ldots, f_{X_J}\}$: functions defining structural equations mapping variables from U, X to variables in X.

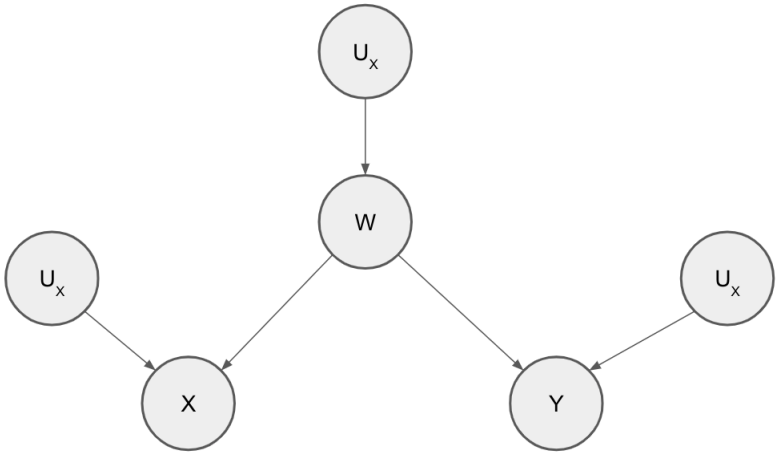
With $X = \{W, A, Y\}$ and $U = \{U_W, U_A, U_Y\}$, an example set of structural equations is

$$\begin{aligned} W &= f_W(U_W) \\ A &= f_A(W, U_A) \\ Y &= f_Y(W, A, U_Y) \end{aligned} \tag{1.20}$$

SCMs encode from domain knowledge, causal assumptions such as which endogenous variables affect each other and which variables are independent. A directed causal graph can be drawn from a specified SCM. For this dissertation we restrict to acyclic SCMs, which lead to directed acyclic graphs. Two causal structures leading to dependence between two variables X and Y are shown in Figure 1.5. Figure 1.5a shows a mediation relationship where the effect of X on Y occurs only through the mediator M. Figure 1.5b shows W being a shared cause of X and Y, which leads to a spurious association between X and Y.



(a) An effect of X on Y via mediator M.



(b) Common cause W between X and Y.

Figure 1.5: Causal structures that can lead to dependence between X and Y.

Conditioning on a causal intermediate or shared common cause between X and Y will remove sources of dependence. However, conditioning on a collider (and its descendents) between X and Y can induce an association between X and Y (Figure 1.6). Spurious association between X and Y due to conditioning on a collider is also referred to as Berkson's bias. Based on the above mentioned rules, we can remove spurious associations between X and Y by conditioning on a set of variables S satisfying

- No element of S is a descendant of X.
- The elements of S "block" all "back-door" paths from X to Y, which are paths that end with an arrow pointing to X.

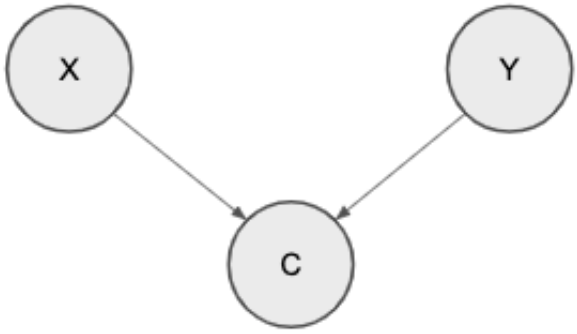


Figure 1.6: Collider C between variables X and Y.

The above mentioned criterion for S is called the back-door criterion.

Mendelian randomization in genetic epidemiology comes from the successful application of causal inference principles. Mendelian randomization uses genetic variation with known association with a modifiable exposure to investigate the causal effect of said exposure on disease. It is assumed that genetic variation only affects disease status through the exposure of interest. This corresponds to the causal structure in Figure 1.5a, where X becomes genetic variation, M becomes the exposure, and Y becomes disease status. Although there may be shared common cause between M and Y, it can usually be assumed that there is no shared common cause between X and Y. This is because genotypes are usually assigned randomly from parents to offspring, like in a randomized controlled trial. Thus, such a study design controls for reverse causation and confounding to determine the causal effect between an exposure with disease. Mendelian randomization has been successfully used show evidence that increasing levels of vitamin D leads to decreased risk of MS [54]. Chapter 3 provides a study applying causal inference to show that DNA methylation mediates surrounding genetic variation on risk of SS at the major histocompatibility complex.

Agglomerative Hierarchical Clustering

Agglomerative hierarchical clustering is an unsupervised learning algorithm that builds a binary tree of the data by successively merging similar clusters of points. In this fashion, every subtree represents a cluster. The resulting dendrogram, or binary tree, from hierarchical clustering provides a helpful visualization of the data. The hierarchical clustering algorithm is presented in Algorithm 1.1.

Algorithm 1.1 Agglomerative hierarchical clustering

```
procedure HIERARCHICAL CLUSTERING({x_1, ..., x_n})
    Place each data point from x_1, ..., x_n into its own singleton group
    while not all data merged into single cluster do
        Merge two closest group
    Return sequence of groups
```

Hierarchical clustering only requires a measure of dissimilarity between data points. Given a dissimilarity measure $d(a, b)$ between points a and b, the popular choices of intergroup dissimilarity between clusters $D(A, B)$ are

- **Complete linkage**: $D(A, B) = \max_{a,b} \left\{ d(a, b) : a \in A, b \in B \right\}$
- **Single linkage**: $D(A, B) = \min_{a,b} \left\{ d(a, b) : a \in A, b \in B \right\}$
- **Average linkage**: $D(A, B) = \frac{1}{|A||B|} \sum_{a \in A} \sum_{b \in B} d(a, b)$
- **Centroid linkage**: $D(A, B) = d(\mu_A, \mu_B)$, where μ_S is mean of S.

Generally, dissimilarity increases monotonically with each level of merge, but centroid linkage can cause inversions where a parent cluster is merged at a lower dissimilarity than its children. Different decisions about intergroup dissimilarity can lead to vastly different dendrograms. Single linkage is sensitive to outliers and can produce unbalanced trees. In contrast, complete linkage leads to most balanced trees out of the intergroup dissimilarities introduced. The comparisons between average linkage, complete linkage, and single linkage are illustrated in Figure 1.7.

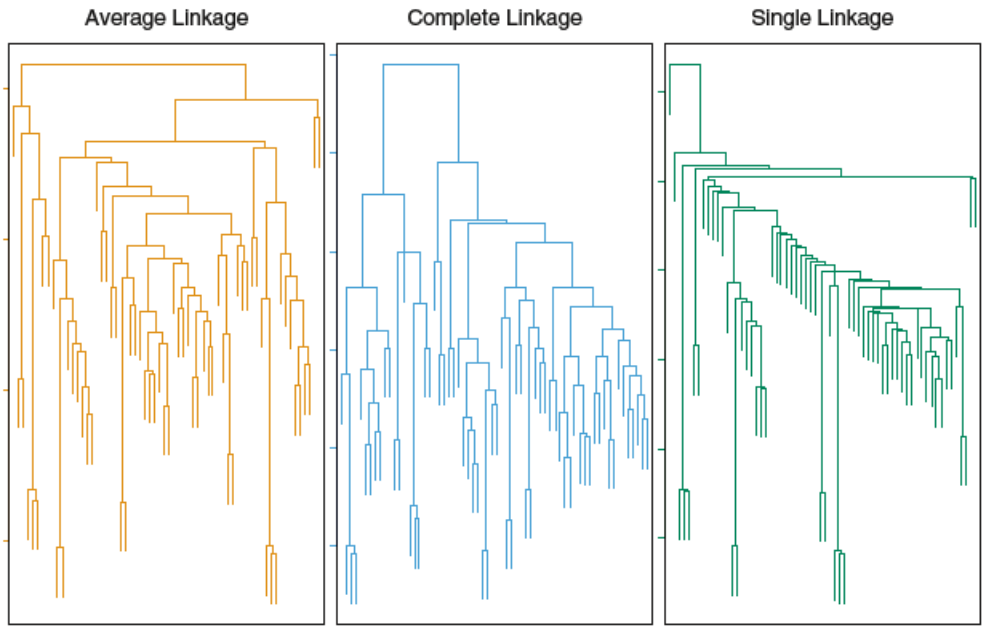


Figure 1.7: Comparison of average, complete, and single linkages. Complete linkage yields a well-balanced dendrogram while single linkage yields a relatively unbalanced dendrogram [55].

Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a method for measuring the associations between two sets of variables. Formally, given random vectors $X_{rv} \in \mathbb{R}^p$ and $Y_{rv} \in \mathbb{R}^q$ with zero means, CCA finds projection vectors $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^q$ that maximizes the correlation between $X_{rv}^\top a$ and $Y_{rv}^\top b$ (Equation 1.21).

$$\max_{a,b} corr(X_{rv}^\top a, Y_{rv}^\top b) = \max_{a,b} \frac{Cov(X_{rv}^\top a, Y_{rv}^\top b)}{\sqrt{Var(X_{rv}^\top a) Var(Y_{rv}^\top b)}} \tag{1.21}$$

The covariance term $Cov(X_{rv}^\top a, Y_{rv}^\top b)$ can be expressed in terms of $Cov(X_{rv}, Y_{rv})$ by a similar manipulation in Equation 1.4.

$$\begin{aligned} Cov(X_{rv}^\top a, Y_{rv}^\top b) &= \mathbb{E}[(X_{rv}^\top a - \mathbb{E}[X_{rv}^\top a])(Y_{rv}^\top b - \mathbb{E}[Y_{rv}^\top b])] \\ &= \mathbb{E}[a^\top (X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top b] \\ &= a^\top \mathbb{E}[(X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top] b \\ &= a^\top Cov(X_{rv}, Y_{rv}) b \end{aligned} \tag{1.22}$$

Since $Var(Z) = Cov(Z, Z)$ for any random variable Z, $Var(X_{rv}^\top a) = a^\top Cov(X_{rv}, X_{rv})a$ and $Var(Y_{rv}^\top b) = b^\top Cov(Y_{rv}, Y_{rv})b$. Thus, Equation 1.21 can be expressed as

$$\max_{a,b} corr(X_{rv}^\top a, Y_{rv}^\top b) = \max_{a,b} \frac{a^\top Cov(X_{rv}, Y_{rv})b}{\sqrt{a^\top Cov(X_{rv}, X_{rv})ab^\top Cov(Y_{rv}, Y_{rv})b}} \tag{1.23}$$

Without access to true distributions of X_{rv} and Y_{rv}, we estimate the covariance terms in Equation 1.23 from data matrices $X \in \mathbb{R}^{n\times p}$ and $Y \in \mathbb{R}^{n\times q}$. Here rows of X and Y are i.i.d. samples $x_i \in \mathbb{R}^p$ and $y_i \in \mathbb{R}^q$. Assume that each feature of X, Y have been centered to have zero mean, then

$$\begin{aligned} Cov(X_{rv}, Y_{rv}) &= \mathbb{E}[(X_{rv} - \mathbb{E}[X_{rv}])(Y_{rv} - \mathbb{E}[Y_{rv}])^\top] \\ &\approx \frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x})(y_i - \hat{y})^\top \\ &= \frac{1}{n}\sum_{i=1}^{n} x_i y_i^\top \\ &= \frac{1}{n}X^\top Y, \end{aligned} \tag{1.24}$$

where the approximation is due to law of large numbers. By similar reasoning, $Cov(X_{rv}, X_{rv}) \approx \frac{1}{n}X^\top X$ and $Cov(Y_{rv}, Y_{rv}) \approx \frac{1}{n}Y^\top Y$. Substituting these estimates, we arrive at an optimization problem that can be solved with data in Equation 1.25.

$$\max_{a,b} corr(X_{rv}^\top a, Y_{rv}^\top b) \approx \max_{a,b} \frac{a^\top X^\top Y b}{\sqrt{a^\top X^\top X a b^\top Y^\top Y b}} \tag{1.25}$$

Some literature refer to a, b as canonical vectors and Xa, Yb as canonical variates. In Chapter 5, we use a special version of CCA called sparse CCA that induces sparsity in a, b, as a means of summarizing the associations between gene expressions and drug sensitivities.

Convolutional Neural Network

Most of modern machine learning is focused on learning functions from data. Deep learning involves choosing a loss function, a neural network architecture as the function approximator, and optimizing the neural network weights with gradient descent. The convolutional neural network is a class of neural network that imposes the following infinitely strong priors on its weights:

- Shared weights for the same hidden activation layer.

- Weights are zero except at the small, spatially contiguous receptive field assigned to the hidden activation unit.
- Each hidden activation unit is invariant to small translations.

We start our introduction with fully-connected neural networks. Consider a supervised learning problem with labeled training data points $(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})$, where $x^{(i)} \in \mathbb{R}^n$ and $y^{(i)} \in \mathbb{R}$. The "neuron" is the most basic building block of a neural network (Figure 1.8).

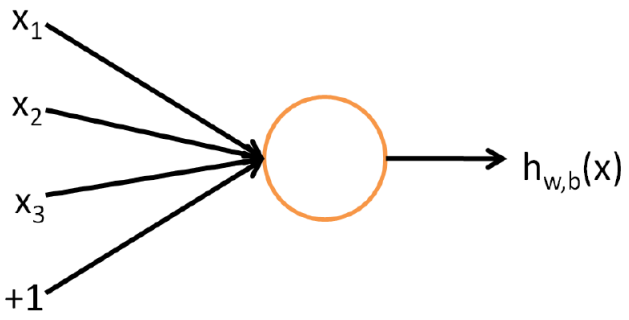


Figure 1.8: A neuron is the basic computational unit of a neural network. In this example, the neuron takes as input x_1, x_2, x_3 and the intercept term, and outputs $h_{W,b}(x) \in \mathbb{R}$.

The computation in Figure 1.8 is $h_{W,b}(x) = f(W^\top x + b)$, where $f : \mathbb{R} \to \mathbb{R}$ is a nonlinear activation function. Common choices for $f(\cdot)$ include the sigmoid, tanh, and ReLU functions, each with its own impact on training. When $f(\cdot)$ is the sigmoid function, then the single neuron in Figure 1.8 corresponds to logistic regression. Non-linear activation functions allow neural networks to approximate nonlinear functions between input and output.

A neural network is comprised of many neurons, with output of a neuron serving as input to the other. An example small neural network is illustrated in Figure 1.9.

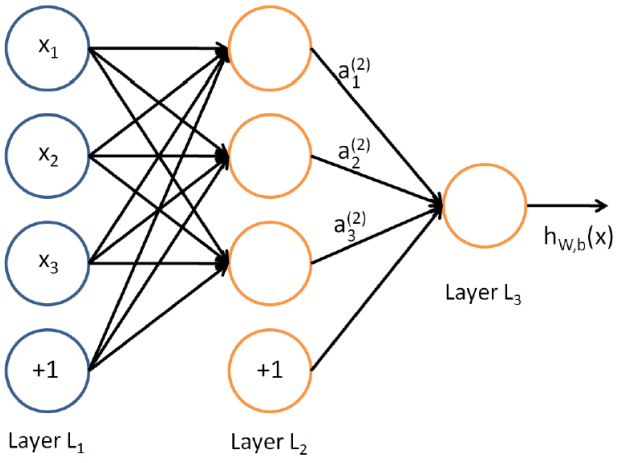


Figure 1.9: An example small neural network.

In Figure 1.9, circles denote inputs to the neural network. Circles labeled with "+1" are called bias units. The leftmost layer of the network is the input layer, the rightmost layer is the output layer, and the middle layers are called the hidden layers. A deep neural network is sometimes defined as a neural network with more than two hidden layers. The neural network in Figure 1.9 has trainable weights $(W, b) = (W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)})$, where $W^{(l)} \in \mathbb{R}^{p \times k}$ denotes the weights associated with layer l and $b^{(1)} \in \mathbb{R}^p$ denotes bias associated with layer l. We use $a_i^{(l)} \in \mathbb{R}$ to denote the activation (i.e. output) value of neuron i in layer l. For example, for the first layer $l = 1$, $a_i^{(1)} = x_i$. The computations performed by the network in Figure 1.9 are

$$a_1^{(2)} = f(W_{11}^{(1)} x_1 + W_{12}^{(1)} x_2 + W_{13}^{(1)} x_3 + b_1^{(1)}) = f(z_1^{(2)}) \tag{1.26}$$
$$a_2^{(2)} = f(W_{21}^{(1)} x_1 + W_{22}^{(1)} x_2 + W_{23}^{(1)} x_3 + b_2^{(1)}) = f(z_2^{(2)}) \tag{1.27}$$
$$a_3^{(2)} = f(W_{31}^{(1)} x_1 + W_{32}^{(1)} x_2 + W_{33}^{(1)} x_3 + b_3^{(1)}) = f(z_3^{(2)}) \tag{1.28}$$
$$h_{W,b}(x) = a_1^{(3)} = f(W_{11}^{(2)} a_1^{(2)} + W_{12}^{(2)} + W_{13}^{(2)} a_3^{(2)} + b_1^{(2)}) = f(z_1^{(3)}). \tag{1.29}$$

Equations 1.26 to 1.29 can be more compactly written as

$$z^{(2)} = W^{(1)} x + b^{(1)} \tag{1.30}$$
$$a^{(2)} = f(z^{(2)}) \tag{1.31}$$
$$z^{(3)} = W^{(2)} a^{(2)} + b^{(2)} \tag{1.32}$$
$$h_{W,b}(x) = a^{(3)} = f(z^{(3)}) \tag{1.33}$$

Before training, the weights (W, b) of a neural network are initialized randomly. The weights of a neural network are typically trained using stochastic gradient descent using a random batch of training samples at each iteration. Given the loss value $\mathcal{L}$, the weight updates at each iteration are

$$W^{(l)} := W^{(l)} - \alpha \frac{\partial \mathcal{L}}{\partial W^{(l)}} \tag{1.34}$$

$$b^{(l)} := b^{(l)} - \alpha \frac{\partial \mathcal{L}}{\partial b^{(l)}}, \tag{1.35}$$

where α is the learning rate. The partial derivatives in Equations 1.34 to 1.35 are efficiently computed using the backpropagation algorithm.

A convolutional neural network (CNN) starts with convolutional layers instead of full-connected layers. A convolutional layer implements the infinitely strong priors mentioned in the beginning of this section. CNNs are probably best understood in the context of computer vision, where the input data are images with dimensions $m \times n \times 3$. A convolutional layer typically carries out the following computations in the order listed

1. Convolution
2. Nonlinearity
3. Pooling

In CNNs, convolution involves sliding a filter (i.e. kernel) of fixed dimension across the input and outputting an activation value via each dot product with the overlapping input. These filters belong to the set of weights the CNN has to train. A 2D convolution is illustrated in Figure 1.10.

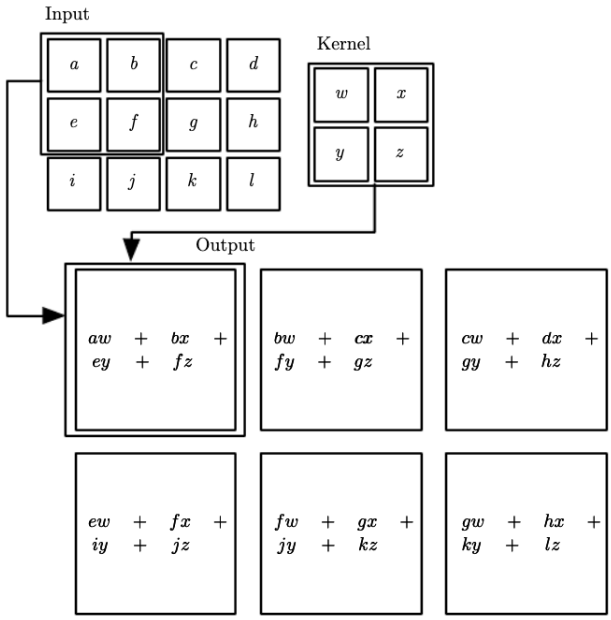


Figure 1.10: 2D convolution illustration [56].

Each filter results in a 2D activation map. A layer with k filters results in k such 2D activation maps. Thus, convolution implements the first two infinitely strong priors mentioned. The non-linearity step involves applying a nonlinear function element-wise to the activation maps. Pooling involves extracting a summary statistic from a fixed region of values. The most common pooling operation is max-pool, which extracts the maximum value of the input region. This operation is illustrated in Figure 1.11.

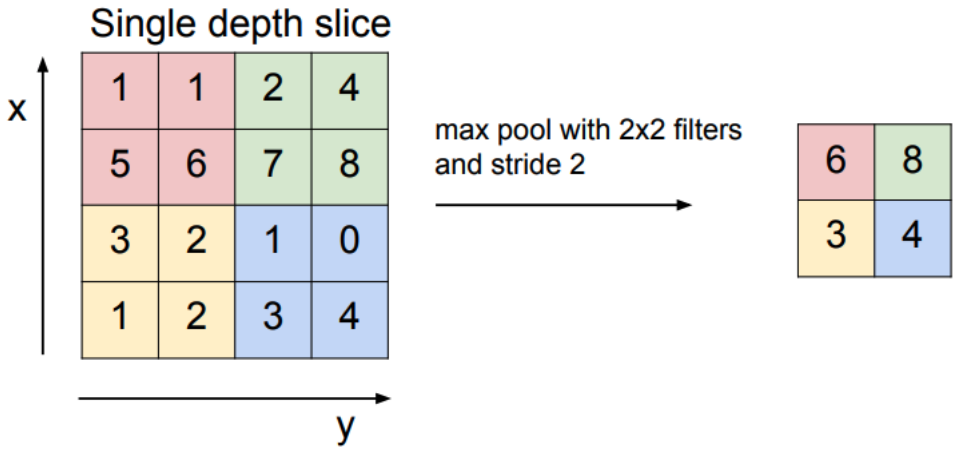


Figure 1.11: Max-pooling operation.

Pooling makes the representation approximately invariant to small translations of input. To gain intuition on how convolutions are helpful, the filters in a CNN can be thought of as being trained to detect certain features. In classical computer vision, edges are important features to detect because they outline the shape of objects in an image. Researchers used to design filters to detect edges of a certain orientation, such that these filters output a high value when a dot product is performed over a region containing an edge. The resulting activation map after convolving such a filter over an image is a dark image with edges outlined in white. An example hand-crafted filter is the Prewitt filter for detecting vertical edges

$$\textbf{Prewitt filter} : M = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

What researchers have found in CNNs is that many filters become trained to detect edges through the backpropagation algorithm. Thus, CNNs are able to learn filters for detecting important features without explicitly designing the filters. Again using an example from computer vision, pooling is helpful when we consider small translations of an object in an image to be equivalent. After starting out with a few convolutional layers for feature extraction, CNNs typically end with a few full-connected layers for final prediction. Chapter 6 provides a study using a 1D CNN to simultaenously impute allele haplotypes across the HLA loci *HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DPB1*, and *-DRB1* from phased SNP genotype data.

Chapter 2

Admixture Mapping Reveals Evidence of Differential Multiple Sclerosis Risk by Genetic Ancestry

2.1 Introduction

Multiple sclerosis (MS) is an autoimmune disease of the central nervous system that results in demyelination and tissue loss. Association studies in White, non-Hispanic populations have discovered human leukocyte antigen (HLA) alleles conferring strong risk and protective effects and 200 non-HLA genetic risk variants conferring modest risk of MS [57, 58]. Evidence that HLA class II alleles interact to confer greater risk of MS have been found [59]. Together, identified MS genetic risk factors are estimated to explain up to 30% of total heritability, of which most is accounted for by HLA alleles [57, 60].

The prevalence of MS varies across the globe but is highest in White, non-Hispanic populations. There is evidence that African Americans are at higher risk for developing the disease, and along with Hispanics, may have a more severe disease course. Incidentally, countries with majority White, non-Hispanic individuals and experience highest MS prevalence are located at higher latitudes. Past studies have not only established the association between ultraviolet radiation and MS prevalence, but have also found evidence supporting the causal role of low vitamin D on MS risk. In this study, we investigate another hypothesis—that the difference in MS prevalence across the globe can be explained by European ancestry. If European ancestry can explain this difference, then MS-associated alleles in admixed individuals can either be European or confer increased risk on a European haplotype compared to a non-European haplotype.

We investigate this by analyzing the genetic ancestry of MS-associated alleles in a large combined cohort totaling 1,471 MS cases and 10,913 controls including African American, Asian American, and Hispanic individuals. Previous studies have been able to replicate the association of the HLA risk allele *HLA-DRB1*15:01* in nearly all populations [61]. Addi-

tional HLA alleles have been found to be associated with MS in non-European populations, such as *HLA-DRB1*15:03* in African Americans and *HLA-DRB1*04:05* in the Japanese population [62]. Limited replication has been achieved for non-HLA genetic risk variants in other populations [63–65]. We found that most MS-associated alleles are cosmopolitan, but there is evidence that European risk alleles may confer more risk than non-European risk alleles, most notably for the major risk allele *HLA-DRB1*15:01*. Thus, there is evidence that the difference in MS prevalence could be explained by European ancestry. We also tested for the association of European ancestry with MS across the genome in African Americans, Asian Americans and Hispanics, and found a signal of association on chromosome 8 in Hispanics.

2.2 Materials and Methods

Sample Collection and Genotyping

Genotype data from a total of 21,647 subjects were collected from the Northern and Southern California Kaiser Permanente memberships, the U.S. Pediatric MS Network, the Genetic Epidemiology Research on Aging (GERA) cohort, and International Multiple Sclerosis Genetics Consortium (IMSGC). Table 2.1 shows the starting number of MS cases and controls by dataset. All cases met the diagnostic criteria for MS [66, 67]. Subjects from Northern California Kaiser Permanente and U.S. Pediatric MS Network were genotyped on the Illumina Human660W-Quad BeadChip, Infinium Human OmniExpress BeadChip, and Infinium Human OmniExpress Exome BeadChip. Subjects from Southern California Kaiser Permanente were genotyped on OmniExpress platforms. The 1,265 African American subjects from IMSGC were genotyped using the Illumina Immunochip and combined with other African Americans to study the ancestry of the major histocompatibility complex (MHC) region [64]. Note that IMSGC subjects were not genotyped genome-wide and were thus excluded from the genome-wide studies in this paper.

Source	Case (n)	Control (n)
Northern California Kaiser Permanente	1,069	637
Southern California Kaiser Permanente	645	636
U.S. Pediatric MS Network	792	413
IMSGC ImmunoChip	803	462
Genetic Epidemiology Research on Aging	0	16,168
Total	3,320	18,327

Table 2.1: Dataset sources for admixed populations. Starting number of cases and controls for each dataset source. n = number of individuals

.

Genotyping details for the GERA cohort are described elsewhere [68]. All genetic coordinates were converted to NCBI Build 37 before analysis. BEAGLE was used to obtain phased data for African Americans, Asian Americans, and Hispanics independently, using GRCh37 genetic map positions in centimorgans converted from GRCh37 genetic coordinates by BEAGLE utility software. Genetic map positions capture genetic linkage information and is used by RFMix for defining windows for local ancestry assignment. The reference panel used for phasing was constructed from selecting individuals from 1000 Genomes with ancestries present in each admixed population [9, 69]. The ancestries represented in our dataset were European (present in all groups), African (present in African Americans and Hispanics), East Asian (present in Asian Americans), and Native American (present in Hispanics).

Imputation

Genome-wide imputation of the dataset against the entire 1000 Genomes phase 3 reference panel was carried out using IMPUTE2 [69, 70]. For HLA imputation, SNP2HLA was used to perform 2-field imputation of alleles for *HLA-A*, *HLA-B*, *HLA-C*, *DRB1*, and *DQB1* using an admixed reference panel from the 1000 Genomes Project, comprised of 165 Native Americans, 155 Africans, 251 East Asians, and 303 Europeans [69, 71, 72]. The reference panel was tailored to contain ancestries represented by the target population to enhance imputation accuracy, and HLA alleles in each admixed population were imputed independently as previously described [73].

Quality Control

SNPs were filtered for minor allele frequency (> 0.01) and missingness on SNPs and samples (> 0.10) before and after imputation with IMPUTE2. Genotype probabilities from IMPUTE2 were converted to hard genotype calls using > 0.6 as the threshold, and SNPs were filtered for info score > 0.30. Additionally, A/T and C/G SNPs were discarded prior to local ancestry inference to avoid strand ambiguity. Related individuals ($\hat{\pi} > 0.25$) were removed from further analysis by identity-by-state, resulting in a total of 20,588 samples. For HLA imputation using SNP2HLA, we removed alleles with R^2 scores less than 0.80 and with allele frequencies below 0.005 from further analysis, filtering out 40, 66, and 63 HLA alleles to result in 70, 47, and 77 HLA alleles for African Americans, Asian Americans, and Hispanics, respectively. All quality control (QC) steps were performed using the PLINK software and R v3.3.1 (www.r-project.org) [74].

Analysis of Population Structure

Population structure was assessed using multidimensional scaling (MDS) and fastSTRUCTURE prior to genotype imputation in order to divide the samples into African American, Asian American, or Hispanic groups for further analysis [75]. MDS components captured ancestry to identify individuals likely to be African American, Asian American, or Hispanic,

using reference populations from the Human Genome Diversity Project (HGDP) [76]. Subjects that cluster with the European reference samples were identified as White, non-Hispanic and subsequently removed. Then, fastSTRUCTURE was used for each group to estimate global admixture proportions for individuals using independent SNPs and a HGDP reference panel tailored to the target population, with default parameters. A cutoff of at least 5% Native American global ancestry for Hispanics was imposed to further remove White, non-Hispanic individuals who were removed based on MDS. The 1,163 candidate Hispanic individuals who did not meet this requirement had an average 0.7% Native American ancestry and 96% European ancestry.

Local Ancestry Inference

We inferred local ancestry genome-wide separately for African Americans, Asian Americans, and Hispanics using RFMix software analysis of imputed and phased genotype data, and a reference panel from the 1000 Genomes Project tailored to the target population [8, 69]. The 1000 Genomes reference panel was selected over the HDGP reference panel as the appropriate reference because it has the required high genotype density for local ancestry inference. RFMix was run on recommended input parameters of 5 minimum number of reference haplotypes per tree node and 3 EM iterations. The number of generations of admixture used as input parameters for RFMix were 5, 6, and 11 for Asian Americans, African Americans, and Hispanics, respectively, according to previous estimates for populations in the United States [77].

Statistical Analysis

Association testing between case status and genetic ancestry was performed using the nonparametric test statistic proposed by Montana and Pritchard for admixture mapping [78].

$$T(l,k) = \frac{(\bar{z}_{l,d}(k) - \bar{z}_{l,c}(k)) - (\bar{q}_d(k) - \bar{q}_c(k))}{SD(z_{l,d}(k) - z_{l,c}(k))} \tag{2.1}$$

Briefly, the term $\bar{z}_{l,d}(k)$ represents the average local ancestry of cases at locus l for ancestry k and $\bar{z}_{l,c}(k)$ is similarly defined for controls. The term $\bar{q}_d(k)$ represents the genome-wide average of ancestry k among cases and $\bar{q}_c(k)$ is defined similarly for controls. Genome-wide ancestry estimates for this statistic are taken from local ancestry estimates from RFMix. This test statistic can be used to test for ancestry association at a single locus or at a region. Under the null, the test statistic follows the normal distribution and a p-value can be obtained through a z-test. The variance $Var(\bar{z}_{l,d}(k) - \bar{z}_{l,c}(k))$ of the test statistic at a given locus was empirically estimated as the sum of variance of average ancestry among cases and controls (see Appendix A.1). The standard deviation follows as the square root of the variance. The estimation of $SD(\bar{z}_{d,l}(k) - \bar{z}_{c,l}(k))$ corresponds to estimating the standard deviation of the average treatment effect, with disease status as treatment and ancestry as

outcome [79]. All terms of the test statistic were estimated from local ancestry estimates from RFMix. Complete details are described elsewhere [78].

Multivariate logistic regression was applied to evaluate the association of genetic variants with MS, using an additive model and adjusting for the first three MDS components to control for population stratification [64, 65]. ORs were used to characterize effect sizes of MS risk alleles. The Wilcoxon test was used to evaluate significance of global admixture proportion differences between cases and controls. All analyses were performed using PLINK and R v3.3.1 (www.r-project.org) [74].

Multiple hypothesis testing was addressed with Bonferroni correction. Bonferroni correction was used to establish significance for the study of non-HLA alleles, and adjusted p-values were provided for all multiple testing scenarios except when the number of tests is ten or less. For genome-wide association studies (GWAS), a significance level of $\alpha = 0.05$ with 15,282 tests results in a genome-wide significance level of 3.27×10^{-6}. Bonferroni correction was applied independently for the studies of African Americans, Hispanics, and Asian Americans.

Since local ancestry assignments span multiple loci, we reduced the burden of multiple hypothesis testing for ancestry association across the genome by only testing one locus per window defined by RFMix for inferring local ancestry, resulting in a total of 15,282 tests genome-wide. Complete details of how RFMix defines windows for local ancestry inference is described elsewhere [8].

Power Calculations

Power calculations are performed with the Genetic Association Study Power Calculator (http://csg.sph.umich.edu/abecasis/cats/gas_power_calculator/), which implements calculations from Skol *et al* [80]. We assume an additive disease model, a MS prevalence of 0.1% in the United States, significance level of 5%, and disease allele frequency of 10% [81]. For HLA alleles, we assume a relative risk of 2, and a relative risk of 1.2 for non-HLA alleles.

Comparison of SNP and Amino Acid Subsequences

SNPs and amino acids (AAs) imputed by SNP2HLA for European and African *HLA-DRB1*15:01* alleles in African Americans were aligned to the UCSC Genome Browser GRCh38 RefSeq Genes track, and the European subsequences were compared to the African subsequences. Note that "subsequence" refers to only the imputed SNPs and AAs, and not to contiguous DNA or AA sequence.

2.3 Results

Analysis of Population Structure

We performed MDS analysis on genotype data from 21,647 subjects to generate components used to control for population stratification in later analyses (Figure 2.1A). This analysis was done separately for African American samples which were genotyped using the Illumina Immunochip (Figure 2.1B). The first three components were sufficient to differentiate global ancestries and broadly categorize samples as African Americans, Asian Americans, or Hispanics. Component 2 was correlated with African ancestry in African Americans ($R = 1.00, p < 0.01$), component 1 was correlated with Native American ancestry in Hispanics ($R = -0.95, p < 0.01$), and component 1 was correlated with East Asian ancestry in Asian Americans ($R = 0.99, p < 0.01$).

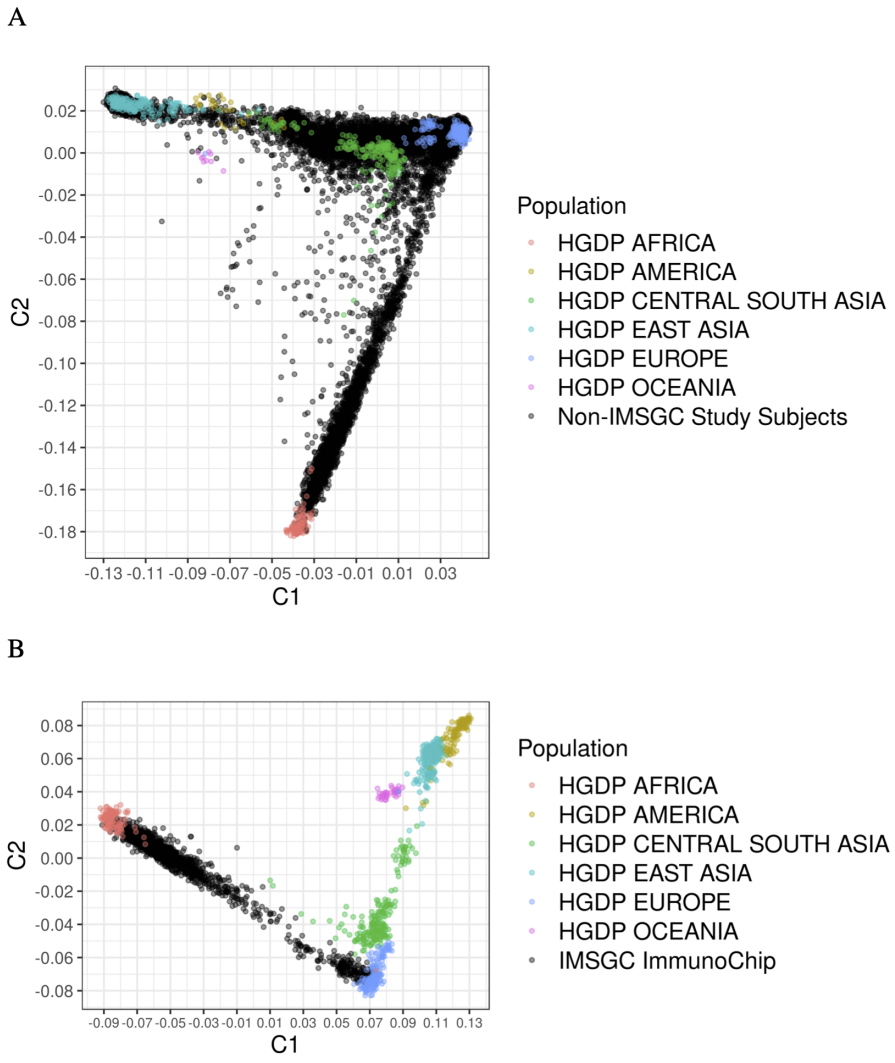


Figure 2.1: Multidimensional Scaling Analysis of Study Subjects with HGDP Reference Samples. (A) Study subjects from Northern California Kaiser Permanente, Southern California Kaiser Permanente, U.S. Pediatric MS Network, Genetic Epidemiology Research on Aging datasets, and (B) IMSGC Immunochip.

We used fastSTRUCTURE to estimate global admixture proportions for individuals from each admixed population. After eliminating White, non-Hispanic individuals and Hispanics with less than 5% Native American ancestry, a total of 3,692 African Americans, 4,915 Asian Americans, and 3,777 Hispanics comprised the final dataset (Table 2.2). African Americans were estimated to be 76.1% African and 23.9% European on average, Asian Americans were estimated to be 92.2% East Asian and 7.8% European on average, and Hispanics were estimated to be 68.4% European, 28.8% Native American, and 2.8% African on average, in

line with published estimates [77].

Population	Case (n)	Control (n)
African Americans	1,081	2,611
Hispanics	326	3,451
Asian Americans	64	4,851

Table 2.2: Number of cases and controls by admixed population. Number of MS cases and controls for African American, Hispanic, and Asian American datasets, after removing related individuals ($\hat{\pi} > 0.25$), White, non-Hispanic subjects, and Hispanics with less than 5% Native American ancestry. n = number of individuals.

The average global admixture for MS cases and controls is shown in Figure 2.2. We observed significant differences in global admixture proportions between cases and controls across all populations. African American cases had 5.0% increased African ancestry compared to controls ($p < 0.01$); Hispanic cases had 5.4% increased Native American ancestry ($p = 0.02$) and 11.3% decreased European ancestry ($p < 0.01$) compared to controls. Asian American cases had 23.0% increased European ancestry compared to controls ($p < 0.01$).

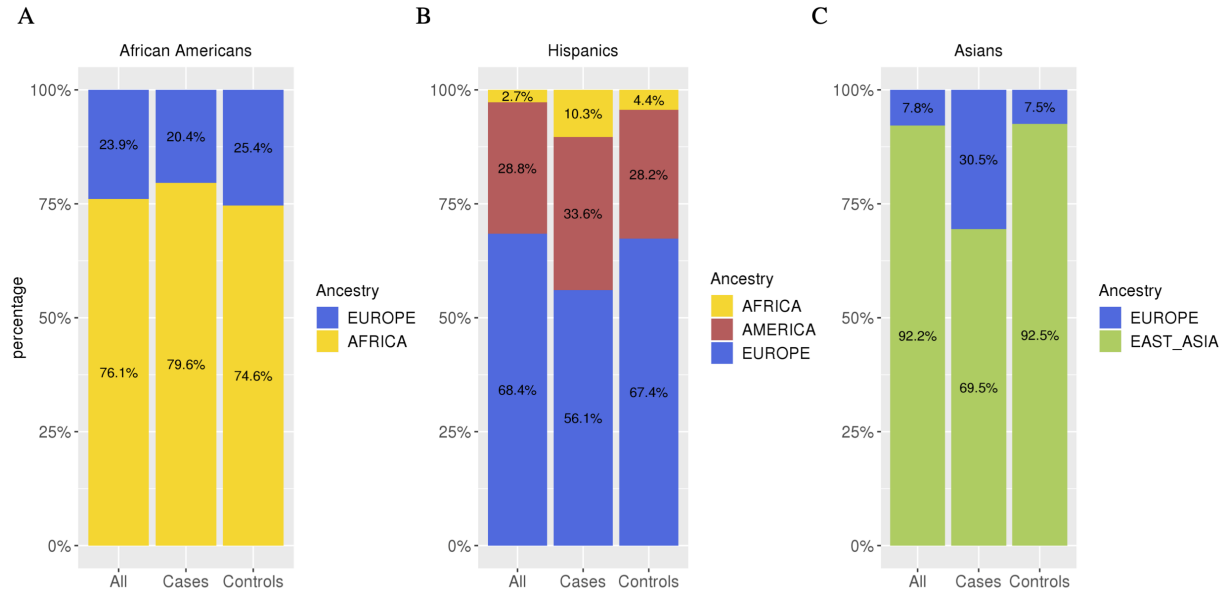


Figure 2.2: Global admixture proportions of study subjects. Global admixture proportion estimates by fastSTRUCTURE with HGDP reference samples. Proportions are shown by case/control status and with cases and controls combined for (A) African Americans, (B) Hispanics, and (C) Asian Americans. The x-axis label "All" denotes admixture proportions for cases and controls combined. See Table 2.2 for the sample numbers corresponding to each admixed population.

Ancestry Association at the MHC

In previous studies, up to eleven regions within the MHC have been identified to exhibit statistically significant independent effects of association with MS in White, non-Hispanic populations: six *HLA-DRB1*, one *HLA-DPB1*, one *HLA-A*, two *HLA-B* alleles, and one signal in a region spanning from *MICB* to *LST1* [82]. We tested each of these regions, in addition to regions spanned by *DQB1* and *HLA-C* and the broader regions class I, II, and III, for evidence of increased European ancestry in MS cases compared to controls. Results are summarized in Tables 2-4 and shown in Fig 3. In African Americans, cases exhibited increased European ancestry at the MHC region compared to controls, after accounting for global admixture proportion differences, with genes in the class I region and the *MICB-LST1* region reaching statistical significance ($p < 0.05$). In Hispanics, the direction of association was the same as in African Americans, but none of the regions reached statistical significance. In Asian Americans, the cases had decreased European ancestry at the MHC region compared to controls, with the regions *HLA-DQB1* and *HLA-DRB1* demonstrating evidence for statistical significance ($p < 0.05$).

MHC Region	$\bar{z}_{l,d}(k) - \bar{q}_d(k)$	$\bar{z}_{l,c}(k) - \bar{q}_c(k)$	z score	P-value
HLA-A	9.12E−3	−1.32E−2	2.04	2.06E−2
Class I	8.78E−3	−1.19E−2	1.91	2.83E−2
MICB-LST1	8.66E−3	−1.09E−2	1.78	3.79E−2
HLA-B	8.19E−3	−1.11E−2	1.75	4.05E−2
HLA-C	9.12E−3	−9.96E−3	1.72	4.24E−2
Class III	8.84E−3	−8.59E−3	1.59	5.57E−2
DQB1	1.00E−3	−7.28E−3	1.57	5.80E−2
DRB1	1.00E−3	−7.28E−3	1.57	5.80E−2
Class II	9.70E−3	−6.27E−3	1.47	7.10E−2
DPB1	9.58E−3	−2.87E−3	1.13	1.30E−1

Table 2.3: European ancestry association with MS at regions of the MHC in African Americans. Tests of European ancestry association with MS using test statistic for admixture mapping, sorted by p-value. The column $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ represents difference in average local and global European ancestry k proportions for cases d. The column $\bar{z}_{l,c}(k) - \bar{q}_c(k)$ is defined similarly as $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ for controls c. The z score is the admixture mapping test statistic calculated as described in Materials and Methods.

MHC Region	$\bar{z}_{l,d}(k) - \bar{q}_d(k)$	$\bar{z}_{l,c}(k) - \bar{q}_c(k)$	z score	P-value
Class I	−5.18E−2	−8.48E−2	1.57	5.76E−2
HLA-A	−5.06E−2	−8.39E−2	1.55	6.11E−2
HLA-C	−5.83E−2	−9.04E−2	1.51	6.51E−2
HLA-B	−5.98E−2	−8.89E−2	1.38	8.43E−2
Class III	−5.09E−2	−7.55E−2	1.22	1.12E−1
MICB-LST1	−6.29E−2	−8.60E−2	1.11	1.34E−1
DQB1	−4.60E−2	−6.72E−2	1.02	1.54E−1
DRB1	−4.60E−2	−6.72E−2	1.02	1.54E−1
Class II	−4.00E−2	−5.32E−2	0.66	2.54E−1
DPB1	−2.76E−2	−2.91E−2	0.07	4.71E−1

Table 2.4: European ancestry association with MS at regions of the MHC in Hispanics. Tests of European ancestry association with MS using test statistic for admixture mapping, sorted by p-value. The column $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ represents difference in average local and global European ancestry k proportions for cases d. The column $\bar{z}_{l,c}(k) - \bar{q}_c(k)$ is defined similarly as $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ for controls c. The z score is the admixture mapping test statistic calculated as described in Materials and Methods.

MHC Region	$\bar{z}_{l,d}(k) - \bar{q}_d(k)$	$\bar{z}_{l,c}(k) - \bar{q}_c(k)$	z score	P-value
DQB1	−1.60E−2	5.98E−2	-1.67	4.80E−2
DRB1	−1.60E−2	5.98E−2	-1.67	4.80E−2
HLA-A	−3.94E−2	3.21E−2	-1.56	5.94E−2
Class II	−1.67E−2	5.35E−2	-1.52	6.39E−2
Class III	−7.52E−3	5.83E−2	-1.48	7.00E−2
DPB1	−2.38E−2	4.32E−2	-1.43	7.69E−2
MICB-LST1	−8.18E−3	5.37E−2	-1.36	8.62E−2
Class I	−1.90E−2	4.07E−2	-1.33	9.24E−2
HLA-B	−3.69E−4	5.28E−2	-1.14	1.27E−1
HLA-C	−3.69E−4	5.27E−2	-1.14	1.27E−1

Table 2.5: European ancestry association with MS at regions of the MHC in Asian Americans. Tests of European ancestry association with MS using test statistic for admixture mapping, sorted by p-value. The column $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ represents difference in average local and global European ancestry k proportions for cases d. The column $\bar{z}_{l,c}(k) - \bar{q}_c(k)$ is defined similarly as $\bar{z}_{l,d}(k) - \bar{q}_d(k)$ for controls c. The z score is the admixture mapping test statistic calculated as described in Materials and Methods.

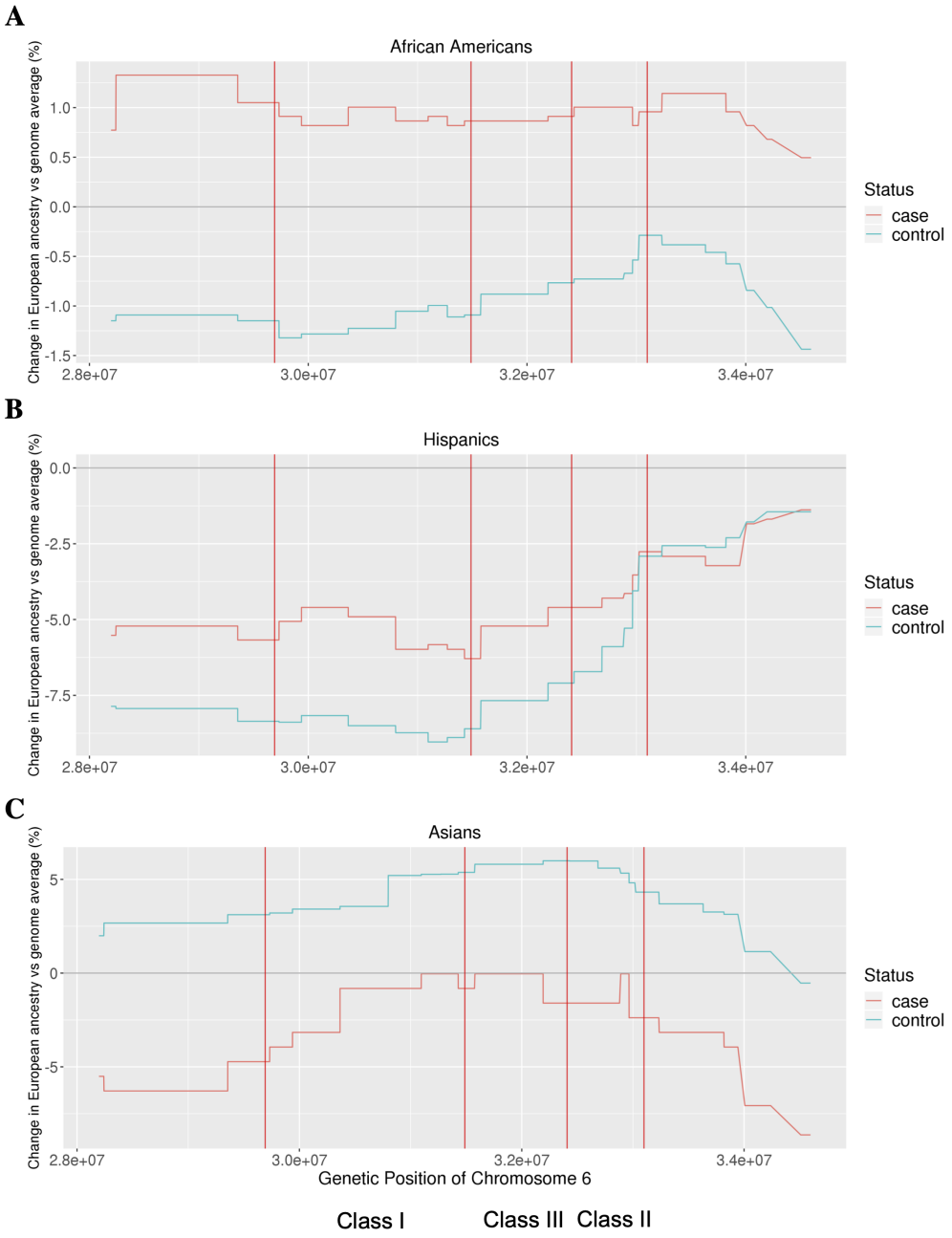


Figure 2.3: Deviation of local from global European ancestry at MHC, plotted for cases and controls. The red vertical bars denote borders of class I, II, and III of the MHC. Local and global ancestries are estimated with RFMix. For both (A) African Americans and (B) Hispanics, cases tended to have higher European ancestry than controls at the MHC. For (C) Asian Americans, cases tended to have lower European ancestry than controls at the MHC.

Ancestry of MS-Associated HLA Alleles

We investigated the ancestry of MS-associated HLA alleles to determine whether ancestry associations observed at the regions within the MHC could be explained. We first identified HLA alleles associated with MS in each admixed group using additive multivariate logistic regression, adjusting for the first three MDS components. We observed 14 alleles in African Americans, 15 alleles in Hispanics, and 4 alleles in Asian Americans that reached nominal significance of association ($p < 0.05$). *HLA-DRB1*15:01*, the strongest genetic association with MS observed in White, non-Hispanic individuals, to date, was a top signal across all three admixed populations, consistent with previous findings [61]. As expected, the African allele *HLA-DRB1*15:03* was significantly associated with MS in African Americans [83]. In African Americans, we further replicated the association of HLA risk alleles previously established in the White, non-Hispanic population: *HLA-DRB1*03:01*, *HLA-A*02:01*, *HLA-DRB1*14:01*, and *HLA-B*38:01* at nominal level significance ($p < 0.05$) [82]. In both Hispanics and Asian Americans, *HLA-DRB1*15:01* is the only established HLA risk alleles in White, non-Hispanics that was replicated. Figure 2.4 compares the p-values of significant MS-associated HLA alleles across different populations. With our sample sizes, we estimate close to 100% power of detection for African Americans and Hispanics and 80% power for Asian Americans. Assuming the MS HLA alleles found in the European population are also associated with MS in admixed populations, then 6, 7, and 4 HLA alleles are expected to be detected in African Americans, Hispanics, and Asian Americans respectively, post quality control.

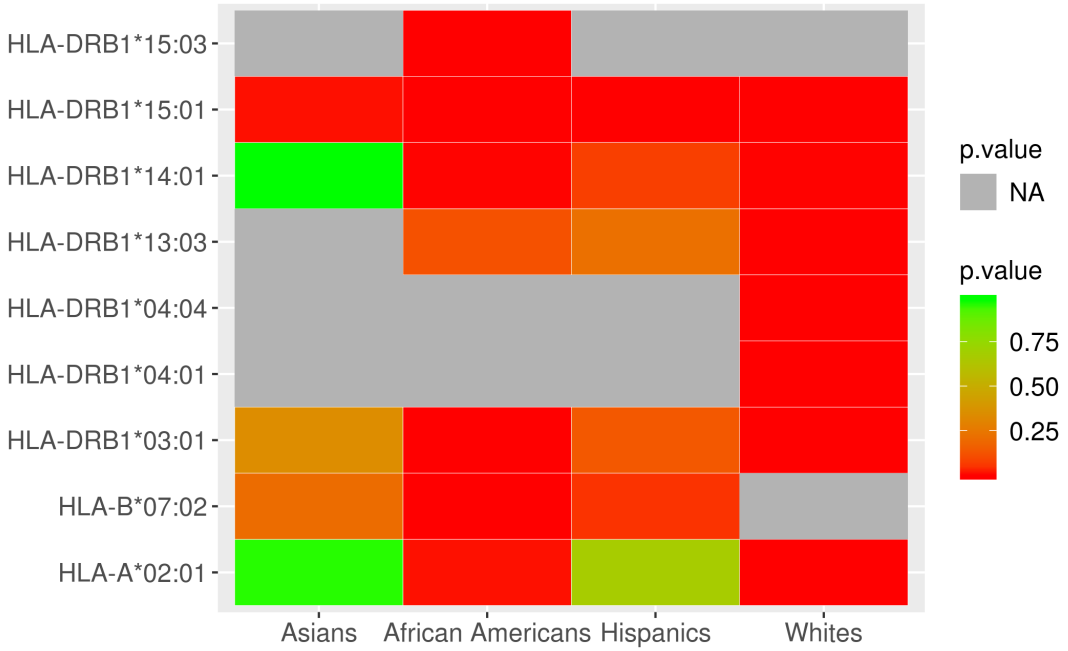


Figure 2.4: Comparison of MS-Associated HLA alleles across populations. P-value heat map for HLA alleles that reached Bonferroni significance in either Asian Americans, African Americans, Hispanics, or White, non-Hispanic individuals. P-values of HLA alleles associated with MS in White, non-Hispanic individuals that were already established in previous work [82]. The HLA allele *HLA-DRB1*15:01* was most consistently associated with MS across all four populations, followed by *HLA-DRB1*03:01*. Gray (NA) denotes an HLA allele that is missing due to not being present in the population or failed to pass HLA imputation QC.

Next, we estimated the admixture proportions of all the nominally-associated alleles using local ancestry estimates from RFMix. Analysis of HLA alleles and corresponding admixture proportions are shown in Tables 2.6 to 2.8, and Figure 2.5. Ancestry estimates for HLA alleles previously established to be ancestry-specific were in strong agreement: 98.4% East Asian for the East Asian allele *HLA-DRB1*04:05* in Asian Americans ($n = 692$ alleles), 96.2% European for the European allele *HLA-DRB1*01:01* in Hispanics ($n = 395$ alleles), 96.4% Native American for Native American allele *HLA-DRB1*14:02* in Hispanics ($n = 454$ alleles), and 99.5% African for African allele *HLA-DRB1*15:03* in African Americans($n = 881$ alleles) [84]. Most MS-associated HLA alleles are cosmopolitan across the admixed populations. The MS risk allele *HLA-DRB1*15:01*, which is more common in Europeans, was estimated to be 63.7% European in African Americans ($n = 512$ alleles) and 96.4% European in Hispanics ($n = 534$ alleles) [85]. However, it is striking that *HLA-DRB1*15:01* is 92.9% East Asian in Asian Americans ($n = 1,228$ alleles).

Allele	N	OR	P-value	Adj p-value	EUR	AFR
*HLA-DRB1*15:01*	512	2.00 (1.58-2.55)	1.26E−8	8.83E−7	0.64	0.36
*HLA-DRB1*03:01*	658	1.45 (1.22-1.72)	2.61E−5	1.83E−3	0.34	0.66
*HLA-B*07:02*	598	1.48 (1.23-1.78)	3.61E−5	2.53E−3	0.53	0.47
*HLA-DRB1*15:03*	881	1.37 (1.17-1.59)	8.11E−5	5.68E−3	0.00	1.00
*HLA-DRB1*14:01*	106	0.39 (0.22-0.70)	1.37E−3	9.58E−2	0.49	0.51
*HLA-A*03:01*	694	1.30 (1.10-1.54)	1.97E−3	1.38E−1	0.38	0.62
*HLA-A*02:01*	914	0.80 (0.68-0.94)	7.88E−3	5.51E−1	0.48	0.52
*HLA-C*08:02*	198	1.44 (1.07-1.94)	1.47E−2	1.00E0	0.21	0.79
*HLA-B*55:01*	45	0.24 (0.07-0.78)	1.74E−2	1.00E0	0.98	0.02
*HLA-DRB1*13:01*	476	0.76 (0.60-0.96)	2.09E−2	1.00E0	0.20	0.80
*HLA-C*04:01*	1704	0.87 (0.76-0.98)	2.35E−2	1.00E0	0.13	0.87
*HLA-C*07:02*	708	1.22 (1.02-1.45)	2.55E−2	1.00E0	0.48	0.52
*HLA-B*38:01*	43	0.28 (0.09-0.94)	3.87E−2	1.00E0	0.91	0.09
*HLA-C*03:02*	95	0.57 (0.33-0.98)	4.30E−2	1.00E0	0.00	1.00

Table 2.6: Ancestry of HLA alleles associated with MS in African Americans. HLA alleles that were nominally associated with MS ($p < 0.05$) and their ancestry proportions estimated from RFMix. Odds ratio (OR) of association for case-control comparison are also shown along with their 95% confidence interval. All tested HLA alleles passed imputation quality score ($R^2 > 0.80$) and have allele frequencies greater than 0.005. N = number of alleles; EUR = European; AFR = African; Adj p-value = Bonferroni adjusted p-value.

Allele	*N*	OR	P-value	Adj p-value	EUR	AMR	AFR
*HLA-DRB1*15:01*	534	2.45 (1.88-3.19)	2.59E−11	2.00E−9	0.96	0.02	0.02
*HLA-DRB1*16:02*	180	1.92 (1.23-2.99)	4.11E−3	3.17E−1	0.01	0.97	0.02
*HLA-DRB1*13:01*	314	0.47 (0.27-0.84)	1.13E−2	8.66E−1	0.81	0.03	0.17
*HLA-C*04:01*	1601	0.76 (0.61-0.95)	1.52E−2	1.00E0	0.22	0.68	0.10
*HLA-DRB1*01:01*	395	0.54 (0.33-0.89)	1.65E−2	1.00E0	0.96	0.01	0.03
*HLA-B*08:01*	318	1.54 (1.08-2.21)	1.71E−2	1.00E0	0.93	0.02	0.05
*HLA-B*40:02*	545	1.39 (1.05-1.84)	2.22E−2	1.00E0	0.09	0.91	0.00
*HLA-DRB1*14:02*	454	0.60 (0.39-0.94)	2.49E−2	1.00E0	0.03	0.96	0.01
*HLA-C*08:02*	327	1.52 (1.05-2.20)	2.53E−2	1.00E0	0.19	0.72	0.08
*HLA-DRB1*09:01*	160	0.39 (0.17-0.91)	2.90E−2	1.00E0	0.19	0.47	0.34
*HLA-A*01:01*	521	1.40 (1.03-1.91)	3.29E−2	1.00E0	0.97	0.02	0.00
*HLA-C*07:02*	976	1.28 (1.02-1.62)	3.55E−2	1.00E0	0.45	0.51	0.05
*HLA-B*14:01*	56	2.16 (1.03-4.55)	4.26E−2	1.00E0	0.95	0.00	0.05
*HLA-B*81:01*	47	2.38 (1.02-5.54)	4.44E−2	1.00E0	0.09	0.09	0.83
*HLA-B*07:02*	460	1.36 (1.00-1.85)	4.86E−2	1.00E0	0.90	0.05	0.06

Table 2.7: Ancestry of HLA alleles associated with MS in Hispanics. HLA alleles that were nominally associated with MS ($p < 0.05$) and their ancestry proportions estimated from RFMix. Odds ratio (OR) of association for case-control comparison are also shown along with their 95% confidence interval. All tested HLA alleles passed imputation quality score ($R^2 > 0.80$) and have allele frequencies greater than 0.005. N = number of alleles; EUR = European; AMR = American; AFR = African; Adj p-value = Bonferroni adjusted p-value.

Allele	N	OR	P-value	Adj p-value	EUR	EAS
*HLA-B*27:05*	50	6.74 (1.92-23.66)	2.92E−3	1.34E−1	0.98	0.02
*HLA-DRB1*15:01*	1228	1.88 (1.19-2.99)	7.26E−3	3.34E−1	0.07	0.93
*HLA-C*03:04*	1687	1.69 (1.04-2.76)	3.43E−2	1.00E0	0.04	0.96
*HLA-DRB1*12:02*	1139	0.32 (0.11-0.94)	3.75E−2	1.00E0	0.00	1.00

Table 2.8: Ancestry of HLA alleles associated with MS in Asian Americans. HLA alleles that were nominally associated with MS ($p < 0.05$) and their ancestry proportions estimated from RFMix. Odds ratio (OR) of association for case-control comparison are also shown along with their 95% confidence interval. All tested HLA alleles passed imputation quality score ($R^2 > 0.80$) and had allele frequencies greater than 0.005. N = number of alleles; EUR = European; EAS = East Asian. Adj p-value = Bonferroni adjusted p-value.

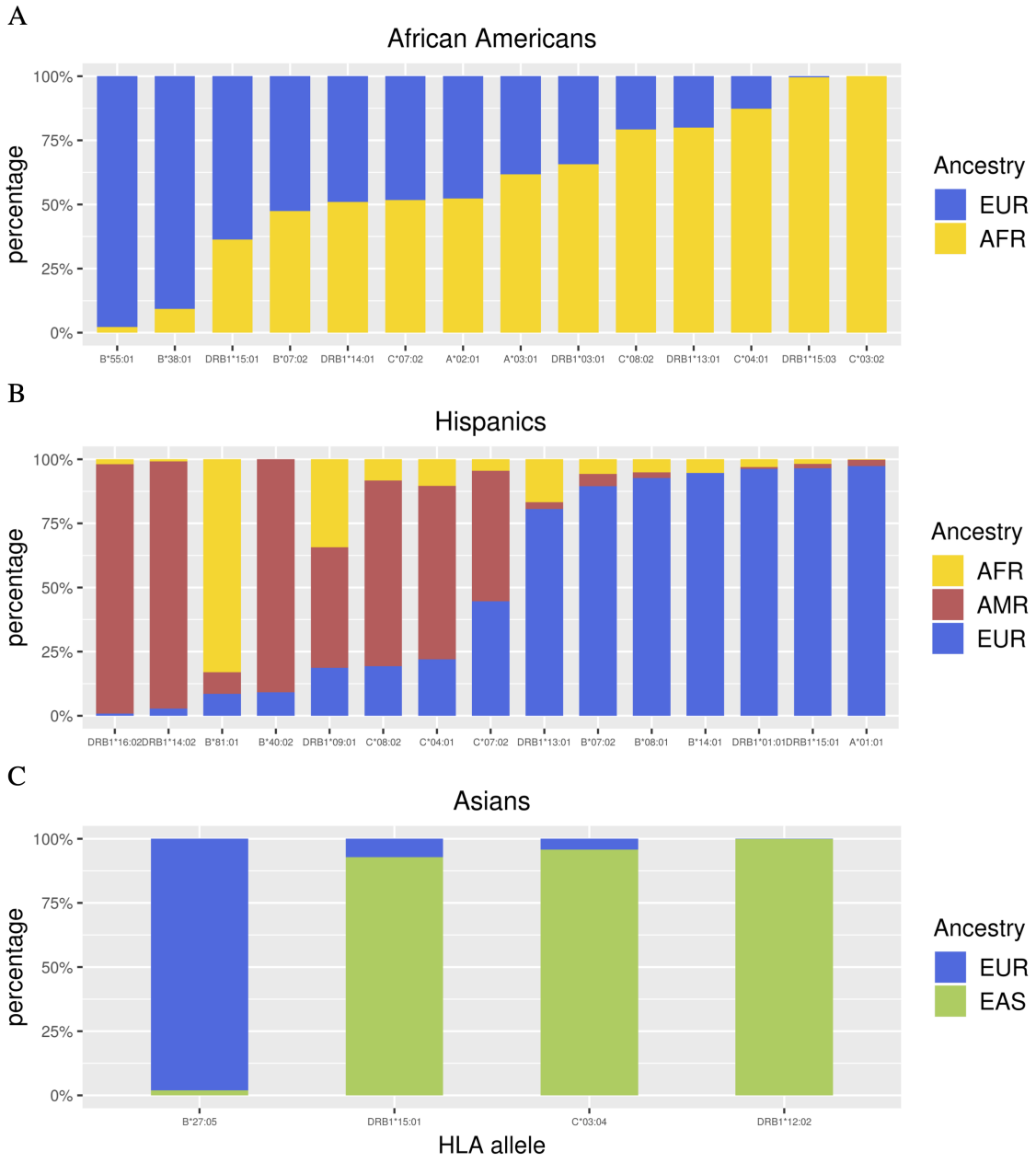


Figure 2.5: Admixture of HLA alleles associated with MS. Ancestry was inferred for MS-associated HLA alleles that passed QC using RFMix. MS-associated alleles were significant at the nominal level ($p < 0.05$), had imputation score $R^2 > 0.80$, and had minor allele frequency greater than 0.005. Other than *HLA-B*55:01* in (A) African Americans, *HLA-DRB1*15:01*, *HLA-DRB1*16:02*, *HLA-DRB1*01:01*, *HLA-DRB1*14:02*, *HLA-A*01:01* in (B) Hispanics, and *HLA-B*27:05* and *HLA-C*03:04* in (C) Asian Americans, HLA alleles associated with MS were cosmopolitan.

We searched for MS-associated HLA alleles that are potentially ancestry-specific, imposing a 96% ancestry cutoff because we were able to correctly estimate the ancestry of HLA alleles of known ancestry as 96% or greater. Briefly, we considered an MS-associated allele as a candidate ancestry-specific allele if 96% of its ancestry comes from a single ancestry across all admixed populations in which it exists, and/or is missing in the rest of admixed populations. An allele could be missing because it does not exist in other ancestries (e.g. African *HLA-DRB1*15:03*), or because it did not pass quality control for imputation. Using this approach, we classified *HLA-DRB1*14:02* and *HLA-DRB1*16:02* as Native American alleles, *HLA-DRB1*15:03* as an African risk allele, *HLA-DRB1*12:02* as an East Asian allele, and *HLA-B*55:01*, *HLA-B*27:05*, and *HLA-A*01:01* as European alleles.

Risk of MS between European and African HLA alleles in African Americans

Given that African Americans exhibit two-way admixture and many MS-associated HLA alleles in African Americans are relatively admixed, we studied the differential risk of HLA alleles in African Americans based on ancestry. We first performed a case-control study of the prominent MS risk allele *HLA-DRB1*15:01* in African Americans to determine whether there were any differences in risk conferred by *HLA-DRB1*15:01* alleles of European and African origin. We removed 12 alleles from the analysis, of which 6 were from cases and 6 were from controls, whose *HLA-DRB1*15:01* allele was not inferred to be completely European or African. Table 2.9 shows the final number of alleles by ancestry and by case status. The risk of MS conferred by the European *HLA-DRB1*15:01* allele was determined from logistic regression to be three times higher compared to the African *HLA-DRB1*15:01* allele (OR = 3.00, 95% CI: $1.90 - 4.76$, $p = 2.49 \times 10^{-6}$), after adjusting for the first 3 MDS components. We restricted the logistic regression to alleles from individuals with one copy of *HLA-DRB1*15:01* so that the association was not confounded by number of *HLA-DRB1*15:01* alleles.

*HLA-DRB1*15:01* Ancestry	Case (n)	Control (n)	
European	129	191	319
African	43	137	180
	171	328	499

Table 2.9: *HLA-DRB1*15:01* of European origin confers greater risk of MS compared to *DRB1*15:01* of African origin. Two-by-two table of counts of *HLA-DRB1*15:01* alleles by ancestry and case/control status. The *HLA-DRB1*15:01* allele passed imputation quality score ($R^2 > 0.80$) and had allele frequency greater than 0.005. Alleles that are not completely European or African were removed.

We continued the same analyses for other alleles in Table 2.6. Alleles with a sample size

less than 50 or with a predominant ancestry of more than 90% are excluded from the analysis. This analysis further revealed that European *HLA-B*07:02* (OR = 1.66, 95% CI: 1.12 − 2.47, $p = 1.18 \times 10^{-2}$) and *HLA-A*03:01* (OR = 1.54, 95% CI: 1.04 − 2.29, $p = 2.97 \times 10^{-2}$) conferred a greater risk of MS compared to their African counterparts for $p < 0.05$. However, for the risk allele *HLA-DRB1*03:01*, the European allele is protective (OR = 0.64, 95% CI: 0.43 − 0.96, $p = 3.03 \times 10^{-2}$) compared to the African allele. Hence, this provides additional evidence that the European haplotype confers more risk of MS compared to the African haplotype for other HLA alleles, although this is not true for every allele Table A.1.

SNP2HLA imputes SNPs and AAs for the exons of HLA alleles, with a 1-to-1 mapping between a SNP and AA subsequence (see Materials and Methods). Given that European *HLA-DRB1*15:01* conferred three times the odds of MS compared to African *HLA-DRB1*15:01*, and without evidence that this finding was due to *HLA-DQB1*06:02* (Tables A.2 to A.3), we compared the most representative SNP and AA subsequences for European and African *HLA-DRB1*15:01* alleles to look for differences. A large majority (94.1%) of European *HLA-DRB1*15:01* alleles shared the same SNP and AA subsequences, whereas African *HLA-DRB1*15:01* subsequences were more diverse. Figure 2.6 shows the comparison of the most frequent (94.1%) SNP and AA subsequence for European *HLA-DRB1*15:01* alleles against the top two most frequent (59.4% and 27.8% respectively) subsequences for African *HLA-DRB1*15:01* alleles, respectively. All differences between the European subsequence and the most frequent African subsequence were within exon 1. When compared against the second most frequent African subsequence, differences were found in exons 1, 3, and 6.

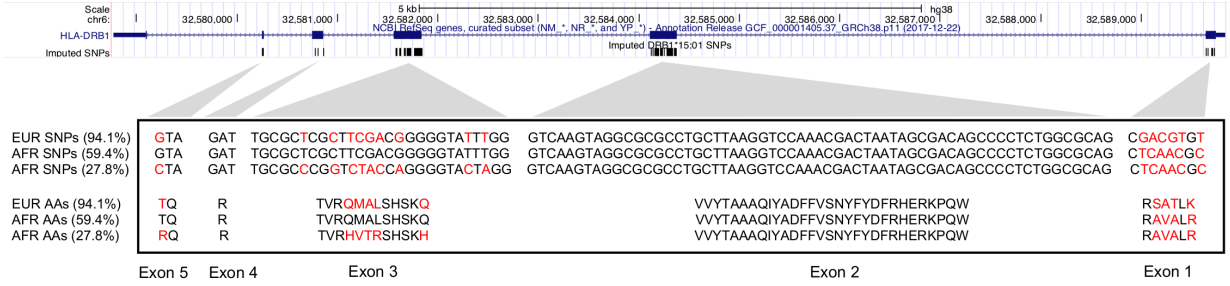


Figure 2.6: European and African *HLA-DRB1*15:01* subsequence comparison. Comparison of SNP and AA subsequences imputed by SNP2HLA for European and African *HLA-DRB1*15:01* alleles in African Americans. Note: subsequence implies the SNPs and AAs are not necessarily contiguous. The subsequences were aligned by position (GRCh38) with the UCSC genome browser NCBI gene track; the imputed positions correspond to exons 1-5 of *HLA-DRB1*. The top frequent (94.1%) SNP and amino acid subsequence for European *HLA-DRB1*15:01* was compared against the top two frequent (59.4% and 27.8%) subsequences for African *HLA-DRB1*15:01*. Red indicates a mismatch between any two given positions between an African and European allele. AA = amino acid; EUR = European; AFR = African.

Ancestry Association at non-HLA MS Genetic Risk Loci

We evaluated the association of European ancestry with MS for 200 established non-HLA genetic risk loci identified in White, non-Hispanic individuals. Following QC, 165 MS risk variants were available in African Americans, 167 MS risk variants in Hispanics, and 154 MS risk variants in Asian Americans for analysis. We tested each risk variant for association with MS and tested each genetic locus for association between European ancestry and MS. Increased East Asian ancestry in MS cases compared to controls for SNPs rs405343 ($p = 5.53 \times 10^{-13}$) and rs6670198 ($p = 6.13 \times 10^{-8}$) was observed in Asian Americans. No other genetic risk locus showed evidence of increased ancestry in cases compared to controls in any admixed population after adjustment for multiple tests. The risk allele T for SNP rs405343 was significantly associated with MS (OR=2.55, 95% CI: $1.70 - 3.83$, $p = 6.87 \times 10^{-6}$) in Asian Americans; however, the risk allele T for SNP rs6670198 showed no evidence for association. A small proportion of MS risk alleles overall demonstrated a nominal level of association at $p < 0.05$: 13 SNPs in African Americans, 21 SNPs in Hispanics, and 28 SNPs in Asian Americans. With our sample sizes, the powers of detection for African Americans, Hispanics, and Asians are estimated to be 21.5%, 26.5%, and 11.7%, respectively. Assuming the established MS non-HLA alleles are also associated with MS in admixed populations, then 35, 44, and 18 non-HLA alleles are expected to be detected in African Americans, Hispanics, and Asian Americans respectively, post QC.

We determined whether European ancestry, both globally and locally at the non-HLA genetic risk loci, was correlated with a cumulative genetic risk score in African American, Asian American, and Hispanic MS cases. Figure 2.7 shows results for each admixed population.

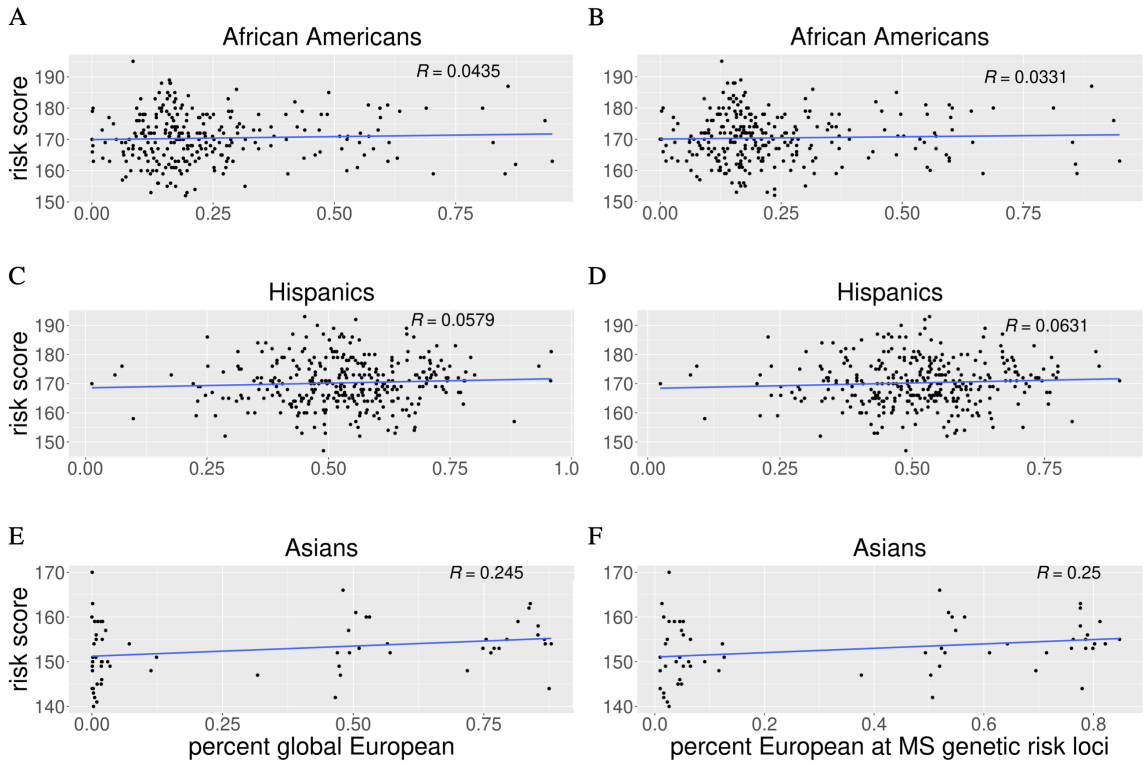


Figure 2.7: Plot of unweighted genetic risk score of the MS genetic risk variants passing QC versus European ancestry globally and locally at the corresponding MS genetic risk loci. All ancestries were estimated with RFMix. Panels (A), (C), and (E) show the relationship between risk score and percentage European global ancestry for African Americans, Hispanics, and Asian Americans respectively. Panels (B), (D), and (F) show the relationship between risk score and percentage European local ancestry calculated at the MS genetic risk loci for African Americans, Hispanics, and Asian Americans respectively. There was little correlation ($R < 0.30$; p-value > 0.05) between genetic risk score and European ancestry.

Globally, no evidence for significant correlation was observed in African Americans ($R = 0.04, p = 0.47$), Hispanics ($R = 0.06, p = 0.30$), or Asian Americans ($R = 0.25, p = 0.05$); similar results were observed for local ancestry in all populations. Admixture estimates showed that the majority of the non-HLA variants investigated here were cosmopolitan; local admixture was reflective of global admixture patterns (Figure 2.8).

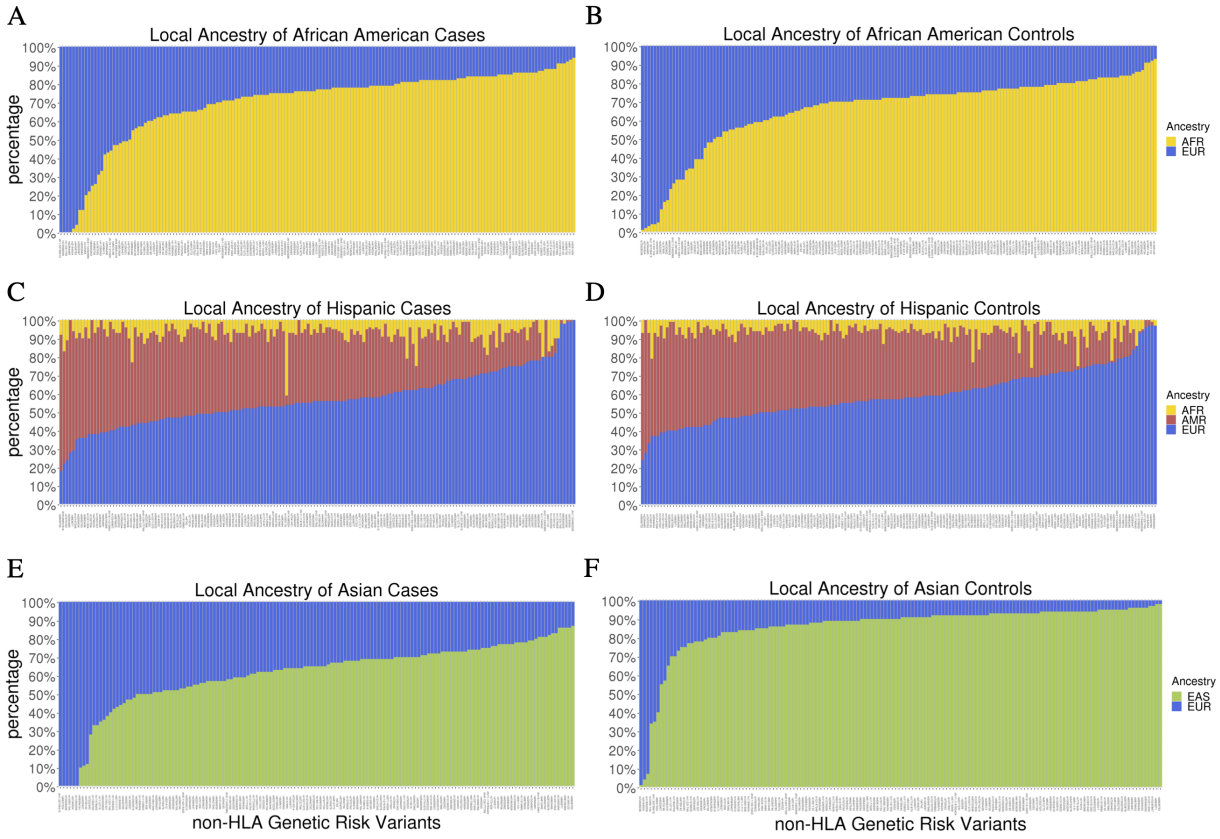


Figure 2.8: Local ancestry estimates from RFMix for the non-HLA risk variants that passed QC, sorted in order of increasing European ancestry. The admixture proportions of risk variants were estimated separately in (A) African American cases, (B) African American controls, (C) Hispanic cases, (D) Hispanic controls, (E) Asian American cases, and (F) Asian American controls. The ancestry proportions of risk variants in cases and controls were largely reflective of global admixture proportions in cases and controls, respectively.

Whole-genome Association Scan

We searched across the genome in African Americans, Asian Americans, and Hispanics to identify regions where individuals with MS had a higher proportion of European ancestry compared to controls using the test statistic in Equation 2.1. The Q-Q plots in Figure 2.9A and Figure 2.9 show that the admixture mapping test statistics are approximately normally distributed except at the tails. The test statistics are least normally distributed for Asian Americans, which exhibits the most imbalance between cases and controls.

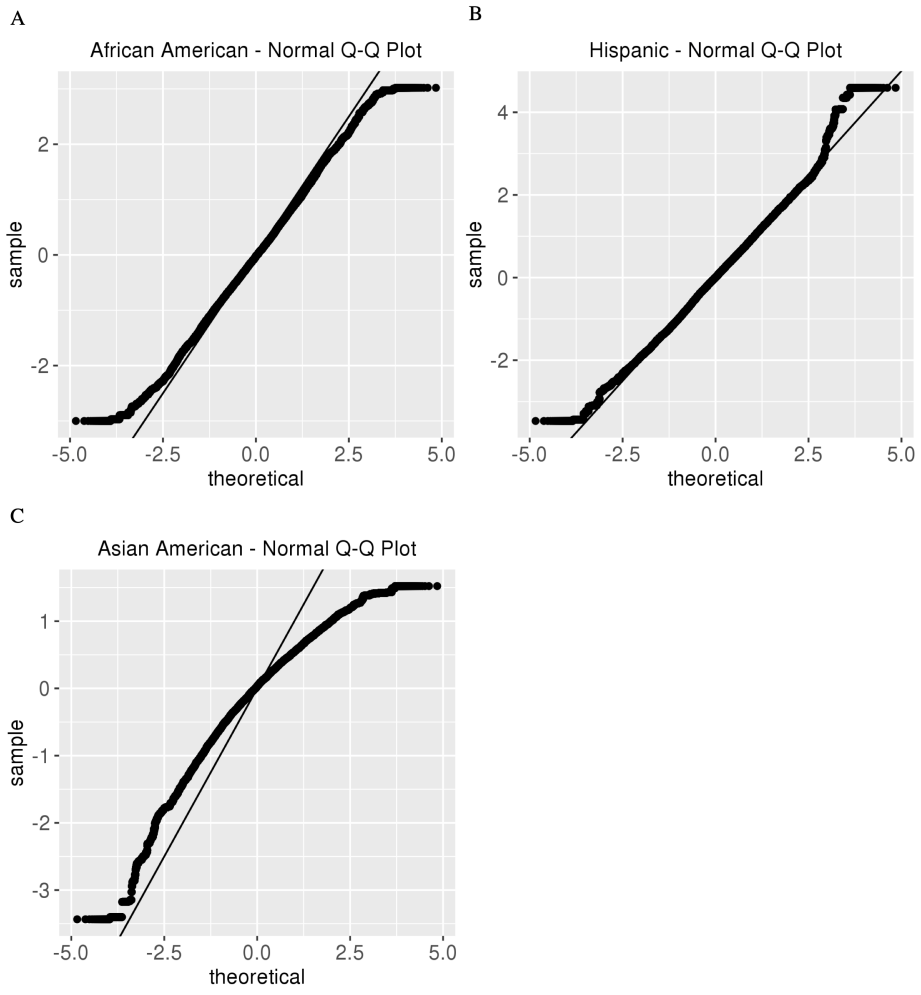


Figure 2.9: Q-Q plot of admixture mapping test statistic for (A) African Americans, (B) Hispanics, and (C) Asian Americans. The line $y = x$ represents the theoretical Q-Q plot if the test statistics are perfectly normally distributed.

The strongest peak of association observed was identified in a single region at chromosome 8 from $207,207-314,620$ (GRCh37) in Hispanics that corresponds to an increase in European ancestry in cases compared to controls (Figure 2.10). This is the only peak that reached genome-wide significance with a Bonferroni adjusted p-value of 3.36×10^2. The closest gene to this region is *ZNF596*, a zinc finger protein 9.8 kb downstream that is most highly expressed in the brain and cerebellum out of 20 different human tissues whose total RNA was sequenced [85].

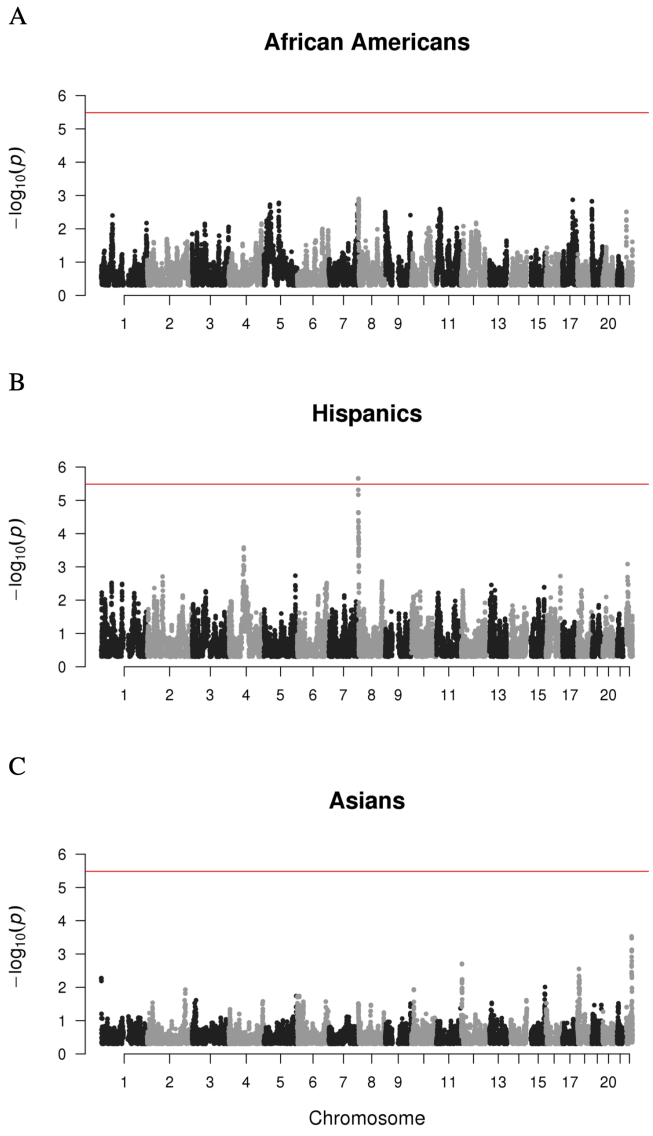


Figure 2.10: Genome-wide association of European ancestry with MS. P-values from testing of association between European ancestry and MS using non-parametric test statistic proposed by Montana and Pritchard, as described in Materials and Methods. One locus was selected from each 0.2 cM window used by RFMix for ancestry inference to reduce the burden of multiple hypothesis testing, resulting in $15,282$ tests. The red horizontal line indicates the negative log of the Bonferroni p-value ($p = 3.27 \times 10^{-6}$) for establishing significance. (A) None of the loci tested for African Americans demonstrated evidence for significant association. (B) In Hispanics, a region spanning from 2Mb to 3Mb on chromosome 8 showed evidence for a significant association. (C) None of the loci tested for Asian Americans were significantly associated.

2.4 Discussion

The genetic contribution to MS susceptibility is very complex; most studies have focused on populations of Northern European descent, and to date, the involvement of genes within and outside the MHC region has been established. Admixed individuals are derived from distinct ancestral populations; global and local genetic ancestry estimates can be used to test for association between the genome, a genetic locus or specific allele and a phenotype of interest [83, 86, 87]. This is one of the first studies to examine the relationship between genetic ancestry, HLA and non-HLA alleles and MS in three admixed populations: African Americans, Hispanics, and Asian Americans.

Within the MHC, we were first able to replicate the association of some previously established HLA risk alleles with MS [82]; *HLA-DRB1*15:01* was the most significant finding across all three admixed populations [61]. Here, the ORs for *HLA-DRB1*15:01* observed in admixed populations (1.88–2.45) were slightly lower than described in previous reports for White, non-Hispanic individuals (2.92) [82], but the direction of effect is consistent. In African Americans, we further replicated the association and direction of effect of HLA alleles previously established in the White, non-Hispanic population: *HLA-DRB1*03:01*, *HLA-A*02:01*, *HLA-DRB1*14:01*, and *HLA-B*38:01* at nominal level significance ($p < 0.05$) [82]. Additionally, we replicated the African HLA risk allele *HLA-DRB1*15:03* in African Americans [65]. A similar study by Isobe, *et al.* also replicated the association of HLA alleles *HLA-DRB1*15:01*, *HLA-DRB1*03:01*, *HLA-DRB1*15:03*, and *HLA*02:01*. Although *HLA-DRB1*14:01* was not found by Isobe to be significantly associated ($p = 0.070$), its protective effect is consistent with what is observed in this study. In summary, we detected association for 5 of the 6 established HLA MS alleles expected to be replicated under power calculations, and this supports the hypothesis that the MS genetic risk in African Americans partially overlaps with that of Europeans [65]. In both Hispanics and Asian Americans, *HLA-DRB1*15:01* is the only established HLA risk allele in White, non-Hispanics that was replicated [82], which suggests a smaller overlap in MS genetic risk between Hispanics and Asian Americans with that of Europeans.

At a nominal level of significance ($p < 0.05$), analysis of the HLA alleles identified five candidate risk alleles and four candidate protective alleles for African Americans, nine candidate risk alleles and four candidate protective alleles for Hispanics, and two candidate risk alleles and one candidate protective allele for Asian Americans. All directions of effect (risk or protective) of candidate MS HLA alleles are the same if found in more than one admixed population. In total, four of the nine protective HLA alleles novel in this study for MS belong to class I genes and five are class II *DRB1* alleles. It is plausible that the lower prevalence of MS in some admixed populations could be partially explained by the effects of protective alleles.

Of the significantly associated HLA haplotypes and alleles reported by Mack, *et al.* in Europeans, three were nominally associated with MS in at least one admixed population in this study [88]. In particular, the *HLA-DRB1*03:01* and *HLA-A*02:01* alleles in African Americans exhibited similar ORs and direction of effect (Table 2.6). However, the *HLA-*

*C*03:04* allele in Asian Americans conferred risk (OR = 1.69) instead of a protective effect (Table 2.8). It is plausible that this disagreement is because an overwhelming majority (95.7%) of *HLA-C*0304* alleles in Asian Americans are of East Asian origin in this study, while the investigation by Mack, *et al.* was in European Americans (Figure 2.5C, Table 2.8). The exon differences observed between European and African *HLA-DRB1*15:01* suggests that future high-resolution HLA analysis could further explain the differences in risk and protective effects that is due to ancestry.

The entire MHC region spanning $29,570,005 - 33,377,701$ (GRCh37) had a higher proportion of European ancestry in MS cases compared to controls for both African American and Hispanic populations. In Asian Americans, the MHC region had a higher proportion of East Asian ancestry in cases compared to controls. Interestingly, the local MHC ancestry associations observed in the current study for African Americans and Hispanics contrasted with global ancestry—African American and Hispanic cases demonstrated less European ancestry compared to controls when the whole genome was taken into consideration, and Asian American cases demonstrated more European ancestry compared to controls. To investigate these associations further, we characterized the admixture proportions of MS-associated HLA alleles. Figure 2.5 shows that a majority of HLA alleles, including *HLA-DRB1*15:01*, were inferred to exist in multiple ancestries and could thus be considered cosmopolitan. African American cases were not significantly European at the class II region compared to controls likely due to the contribution of the common African allele *HLA-DRB1*15:03*. In Asian Americans, *HLA-DRB1*15:01* and *HLA-C*03:01* conferred risk of MS and accounted for 68.6% of HLA alleles associated with MS. Together, these two alleles had an average of 94.7% East Asian ancestry which helps explain why cases tended to have a higher proportion of East Asian ancestry compared to controls within the MHC region.

We find it noteworthy that the European *HLA-DRB1*15:01* allele confers three times the odds of MS compared to the African *HLA-DRB1*15:01* allele in the African Americans we studied. A similar effect has been observed for European *HLA-B*07:02* and *HLA-A*03:01*. Together these findings provide evidence that in some genetic regions, the European haplotype could confer more risk of MS than haplotypes derived from other ancestries. In these cases, it is plausible that disease-causing genetic variants can come from only one ancestral population. However, it must be noted that this has not been found to be true for all admixed MS-associated alleles we examined (Table A.1), and that for alleles such as the African MS risk allele *HLA-DRB1*15:03*, the African haplotype confers more risk than the European haplotype. These findings together further highlight the complex genetic ancestry of MS-associated alleles in admixed populations.

Given that *HLA-DRB1*15:01* is in very strong linkage disequilibrium with *HLA-DQB1*06:02* in Europeans, we investigated whether the increased risk of MS in African Americans conferred by European *HLA-DRB1*15:01* could possibly be due to *HLA-DQB1*06:02*, despite the limitation that *HLA-DQB1* did not pass our imputation quality cutoff (average $R^2 = 0.53$ across all *DQB1* alleles). As expected, 99.5% of *HLA-DRB1*15:01* haplotypes that include *HLA-DQB1*06:02* in African Americans are of European ancestry. Table A.1 shows that *HLA-DQB1*06:02* was not associated with MS in African Americans (OR =

1.14, 95% CI: 0.68-1.92, p = 0.72). Further, Table A.3 shows that comparison of European *HLA-DQB1*06:02* alleles with African *HLA-DQB1*06:02* alleles, in the absence of *HLA-DRB1*15:01*, did not demonstrate evidence for a significant association (OR = 0.48, 95% CI: 0.23-1.00, p = 0.07); the direction of effect is, in fact, protective. Results from the current study are consistent with a previous report showing the association of MS with the *HLA-DRB1*15:01-DQB1*06:02* haplotype is due to the *DRB1* locus independent of *DQB1*06:02* [89].

A comparison of the most commonly imputed SNP and AA subsequences between European and African *HLA-DRB1*15:01* alleles revealed mismatches at exons 1, 3, and 5. Each of these exons help encode the DR beta 1 heterodimer, with exon 1 encoding the leader peptide and exon 5 encoding the cytoplasmic tail of the membrane protein. Exon 3, together with exon 2, encode the two extracellular domains [90]. Further investigation into whether genetic variation in these exons have functional consequences for peptide presentation in the context of MS is warranted. Our case study of *HLA-DRB1*15:01* illustrates how admixture mapping can be broadly applied to better characterize risk alleles in admixed populations.

Consistent with previous attempts to replicate the association of non-HLA genetic risk variants, we also failed to replicate association of most non-HLA genetic risk variants across all three admixed populations, except for rs405343 and rs6670198 in Asian Americans, which exhibit the same direction of effect as in whites [63–65]. Without correction for multiple testing with significance established at $p < 0.05$, we replicated the association of 13 SNPs in African Americans, 21 SNPs in Hispanics, and 28 SNPs in Asian Americans. For African Americans and Hispanics, we replicated less associations than is expected under power calculations. For Asian Americans, more associations were replicated than is expected. The majority of non-HLA MS risk variants identified so far appears to be cosmopolitan and their observed ancestry proportions are reflective of global admixture proportions (Figure 2.8). European global ancestry and European local ancestry at the non-HLA genetic risk loci was not correlated with the unweighted genetic risk score comprised of the non-HLA variants (Figure 2.7). Although our investigation showed that the majority of non-HLA MS genetic risk variants reported for the White, non-Hispanic population do not demonstrate strong associations with MS in African Americans, Asian Americans, and Hispanics, our study is under-powered to detect most associations. Besides lacking power due to small sample and effect sizes, there are multiple other explanations for why we may fail to replicate many associations of the non-HLA genetic risk variants with MS [65]. One explanation is that differences in minor allele frequencies reduced the power to detect associations in admixed populations. Another explanation is that the smaller haplotype blocks of African Americans and Hispanics may have caused many non-HLA genetic risk variants to fail tagging the putative causative variant of MS. Lastly, the absence of replication could simply be due to genetic heterogeneity across populations, which further justifies the need for GWAS in non-White populations.

A genome-wide search for European ancestry differences between MS cases and controls in all three admixed populations resulted in one region of chromosome 8 from $207,207$ to $314,620$ (GRCh37) in Hispanics only. The closest gene to this region is *ZNF596*, a zinc

finger protein 9.8 kb away that is highly expressed in the brain and cerebellum. Lesions in brain tissue as well as brain atrophy are pathological hallmarks of MS [91], and available data suggest Hispanics may have a more severe disease course than White, non-Hispanic individuals [92]; however, these findings await replication. Further investigation of this region in a larger independent dataset and full interrogation of nearby genes and determining whether *ZNF596* could be involved in MS pathogenesis from a functional perspective are warranted.

Some important strengths of this study included comprehensive analyses of a large, well-characterized dataset comprised of 12,384 admixed MS cases and controls with high quality genetic data, the application of rigorous QC procedures, genetic imputation methods for both SNP and HLA loci, probabilistic graphical modeling for local admixture estimation across the genome, and non-parametric statistical testing to identify local admixture differences between cases and controls that accounts for global differences. In the current study, the combined analysis of SNP and HLA genotypes in African Americans revealed for the first time, strong evidence that the European *HLA-DRB1*15:01* allele confers three times the MS risk compared to the African *HLA-DRB1*15:01* allele. This finding indicates increased risk attributed to the European 15:01 allele could be due to functional differences within *DRB1* itself, or possibly due to variant(s) present on the European *HLA-DRB1*15:01* haplotype that are not found on the African haplotype.

Some limitations must also be acknowledged. The diagnosis of MS cases in this large dataset occurred over a twenty-five year period and in different clinical settings; both prevalent and incident cases were included. Although all cases fulfilled established diagnostic criteria, is not known whether local genetic ancestral proportions (of particular importance in the current study) would be expected to change for cases diagnosed at different time points; larger investigations would be needed. We performed MDS analysis of genotype data to broadly categorize samples as African Americans, Asian Americans, or Hispanics for case-control analysis; careful matching on self-reported race/ethnicity was not possible for all individuals. MDS components were therefore used in each analysis to control for potential confounding; however, it is possible that population stratification could still contribute to some of our findings. The Asian MS case sample utilized in the current study was small compared to the other groups, reflecting the low prevalence of disease in this population, which reduced power to detect to modest effects.

In conclusion, results from the current study reveal a complex picture of genetic ancestry for MS-associated alleles in African Americans, Asian Americans, and Hispanics. Our study shows that the higher prevalence of MS in populations of northern European ancestry cannot simply be explained by the European ancestral origin of genetic risk factors. Rather, any difference in prevalence due to genetics might be partially explained by a combination of European risk alleles exerting greater risk (i.e. *HLA-DRB1*15:01*) compared to non-European risk alleles, or the presence of protective alleles in individuals of non-European ancestry. However, this does not rule out the possibility that observed prevalence differences could result from the influence of environmental risk factors or socioeconomic status, including differences in access to neurologists and diagnostic protocols using MRI, that may be

population-specific.

Chapter 3

Hypomethylation of immune genes mediates methylation quantitative trait loci at the major histocompatibility complex in Sjögren's syndrome

3.1 Introduction

Sjögren's syndrome (SS) is an autoimmune disease characterized by the lymphocytic infiltration of salivary and lacrimal glands, resulting in dryness of the mouth and eyes, fatigue, and joint pain. The prevalence of SS is estimated to be 3% in individuals aged 50 years or older and 0.6% overall, with a 9:1 female-to-male predominance [93]. When SS occurs in isolation, it is referred to as primary SS; secondary SS co-occurs with other systemic autoimmune diseases [94]. Environmental factors including infectious agents, stress, air pollution, and silicone are implicated in disease pathogenesis [95–98]. Genetic association studies have established genetic loci both within and outside the major histocompatibility complex (MHC) [99–101].

Differential methylation has been a consistent theme reported by multiple studies of CD4+ T cells, CD19+ B cells, whole blood, and labial salivary glands (LSGs) in SS [102–111]. Specifically, a general hypomethylation of immune-related genes have been discovered, along with implications for altered gene expression. Some of these studies have found evidence suggesting genetic control of DNA methylation. Imgenberg-Kreuz *et al.* identified methylation quantitative trait loci (meQTL), or loci where genetic variation is associated with DNA methylation, in whole blood. However, this analysis was performed in controls only instead of both cases and controls [111]. Another study reported an overlap of differentially methylated probes with established genetic risk loci [112]. Since the relationships were

associative and observed in datasets from different patients, evidence for genetic control is suggestive at best.

We formally investigated evidence for genetic control in SS in the largest study of LSGs from 64 primary SS cases and 67 symptomatic non-cases from the Sjögren's International Collaborative Clinical Alliance (SICCA) registry. Since non-cases are symptomatic for SS phenotypes, DNA methylation differences are more likely to reflect disease pathology than if we compared against healthy controls. Our overall approach first uses *bumphunter* to identify differentially-methylated regions (DMRs), or genetic regions differentially methylated in the same direction. Then, we identify meQTLs as SNPs ±250 kb away from a DMR where genetic variation is associated with DNA methylation levels. Finally, we perform the causal inference test developed by Millstein *et al.* to find DMR-meQTL pairs where the DMR mediates the risk of the surrounding meQTL on SS [113]. With this study, we report (1) CpG sites within the genome that showed evidence of mediating nearby genetic associations with SS, which by extension also revealed (2) CpG sites whose methylation levels were independent of neighboring genetic variation. Methods for site-specific epigenome editing are currently under development [114], and by providing an understanding of the genetic factors that influence differential methylation in SS, our study is essential for the potential application of such therapeutic approaches to SS.

3.2 Materials and Methods

Study subjects and clinical evaluation

A total of 131 female, non-Hispanic white individuals were selected from the SICCA registry for this study. All individuals from the SICCA registry exhibited at least one symptom related to SS, specifically symptoms of dry eyes or dry mouth, prior suspicion/diagnosis of SS, positive serum anti-SSA, anti-SSB, rheumatoid factor or antinuclear antibody results, increase in dental caries, bilateral parotid gland enlargement, or a possible diagnosis of secondary SS [115]. Case status was determined according to the 2016 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) criteria for SS [116]. "Non-case" from the SICCA registry with at least one, but not all, SS symptoms or signs were also included. More specifically, "non-cases" did not meet ACR/EULAR for SS but were enrolled in SICCA due to the presence of 1 or more symptoms or signs suggesting possible SS. Based on these criteria, we studied 64 SS cases and 67 non-cases.

Methylotyping and preprocessing

DNA was extracted from the LSG tissue collected from each study subject as previously described [104]. DNA methylation was measured for each subject using the Illumina 450K Infinium Methylation BeadChip (450K) platform for 28 subjects and the Infinium MethylationEPIC (EPIC) platform for 103 subjects. The 450K and EPIC chips allow for high-

throughput interrogation of more than 450,000 and 850,000 highly informative CpGs sites respectively, spanning 22,000 genes across the genome.

Methylation data processing was performed using `Minfi`, a Bioconductor package for the analysis of Infinium DNA methylation microarrays [117]. Background subtraction with dye-bias normalization was performed on methylated and unmethylated signals with the *noob* procedure, followed by quantile normalization with `preprocessQuantile` [118, 119].

For joint analysis of all 131 samples, the intersection of CpGs from 450K and EPIC chips was selected for analysis, resulting in a starting number of 452,832 CpGs. Probes where more than 5% of samples had a detection p-value > 0.01 were removed, to retain probes where signal is distinguishable from negative control probes. To remove probes with ambiguous methylation measurements due to incomplete binding between the DNA strand of interest and probe strand DNA, probes with SNPs with minor allele frequency greater than 0% at either the probe site, CpG interrogation site, or single nucleotide extension were removed. Finally, probes identified with probe-binding specificity and polymorphic targets problems, or cross-reactive probes, were removed [33, 34]. The final preprocessed dataset consisted of 336,040 CpG sites. Since no subject had more than 5% of probes with detection p-value > 0.01, all 131 subjects were retained. Both M-values and β-values were used in subsequent analyses (see Appendix B.1).

Removing unwanted DNA methylation variation

We identified array type (450K or EPIC), genetic ancestry, self-reported age of SS syndrome onset, collection phase, smoker status, anticholinergic drug use, and co-morbidities as potential confounders. Of these, array type and genetic ancestry were found to be strongly associated with DNA methylation and case status respectively, and analytical models were adjusted accordingly (Figures B.1 and B.2). However, case status was not associated with array type, because the distribution of cases and non-cases were similar between 450K and EPIC with 46.4% cases and 50.0% non-cases respectively. Wilcoxon's rank sum test of difference in ancestry MDS component values between cases and non-cases revealed a significant association at p-value ≤ 0.05 for components 2 - 4 and at p-value ≤ 0.10 for component 1. Unwanted methylation variation due to array type and genetic ancestry was removed from β-values and M-values using `ComBat` from the `SVA` package, which applies an empirical Bayes, model-based location/scale batch adjustment [120, 121]. See Appendix B.1 for details of `Combat` usage.

Genotyping and quality control

The subject genotypes were taken from the genotypes of the larger SICCA cohort, which was genotyped on the Illumina HumanOmni2.5-4v1 or Illumina HumanOmni25M-8v1-1 arrays from DNA extracted from whole blood. All quality control steps performed have been previously described [99]. The final genotype dataset consisted of $1,392,448$ SNPs.

Dimensionality reduction

Principal component analysis (PCA) was performed on the centered and scaled β-value matrix $X \in \mathbb{R}^{n \times p}$, where n and p are the number of subjects and CpG sites, respectively. We determined the relative influence of each CpG site on a principal component by analyzing principal component loadings (see Appendix B.1).

Multidimensional scaling (MDS) was performed to detect population structure using lower dimensions that explain observed genetic distance. With genotype data as reference allele counts, pairwise genotype dissimilarity is summarized by the distance matrix $D = J - IBS \in \mathbb{R}^{n \times n}$, where $IBS \in \mathbb{R}^{n \times n}$ is the identity-by-state similarity matrix and $J \in \mathbb{R}^{n \times n}$ is the all-ones matrix. MDS of genotypes from the 131 subjects and reference European subpopulations from the Human Genome Diversity Project (HGDP) [76] was performed using PLINK 1.9 to assess association between genetic ancestry and case-control status [122].

Identification of differentially methylated regions

Differentially methylated regions (DMRs) were identified using *bumphunter*, which searches for bumps, or contiguous CpG sites consistently hypermethylated or hypomethylated in one group of subjects compared to the other [123]. The linear regression specified for bumphunter was

$$M \sim outcome + array\ type + C1 + \cdots + C5, \tag{3.1}$$

which controlled for array type and genetic ancestry. Here, "M" is the M-value without batch correction with `Combat`, *outcome* is SS case status, *array type* indicates array (450K or EPIC), and $C1 - C5$ indicate the first five MDS components of genotype data. The number of bootstrap resampling B was set to 1,000 for generating null distribution of candidate DMRs for establishing significance. Significant SS DMRs were stringently selected as those with $fwerArea \leq 0.05$, defined as proportion of bootstraps with maximum bump area greater than observed DMR area, and consists of at least two CpG sites. See Appendix $B.1$ for details on choice of *bumphunter* hyperparameters and annotation of DMRs.

Gene set enrichment analysis

Since methylation at transcription start sites and gene bodies has been shown to regulate gene expression [124], we restricted gene set enrichment analysis (GSEA) to genes differentially methylated at the promoter or gene body. DMR genes were tested for enrichment of gene ontology (GO) gene sets from the Molecular Signatures Database [125] combined with SS-related gene sets from past studies using the hypergeometric test (see Appendix $B.1$ for gene set details). False discovery rate was controlled with the Benjamini-Hochberg procedure [126]. Since genes in the same pathway tend to be up or down-regulated together, GSEA

was performed separately for hypermethylated and hypomethylated DMR genes in cases compared to non-cases [127].

Identification of DNA methylation quantitative trait loci

MeQTLs are loci whose genotypes are associated with DNA methylation. In this study, candidate meQTLs for a DMR are defined as independent SNPs in the genomic region ±250 kb away from the start and end positions of the DMR respectively. Independent SNPs were randomly pruned using PLINK to satisfy pairwise correlation $r^2 \leq 0.5$ in a $250,000$ bp window, with a window stride of $25,000$ bp [122]. The association between a candidate meQTL and DMR was established by regressing the M-value, averaged across CpG sites of the DMR, against genotype encoded as 0, 1, or 2 copies of the reference allele. The DNA methylation values used for identifying meQTLs were batch-corrected for array type and genetic ancestry. Significance of association was evaluated using t-test from linear regression. Multiple hypothesis testing was addressed with the Benjamini–Hochberg procedure [126].

Mediation analysis with causal inference test

We used the CIT to determine whether the influence of meQTLs on SS was mediated by DNA methylation levels[22,54]. We evaluate evidence for the causal mediation model where genetic variation influences SS disease status through DNA methylation levels (Figure 3.1).

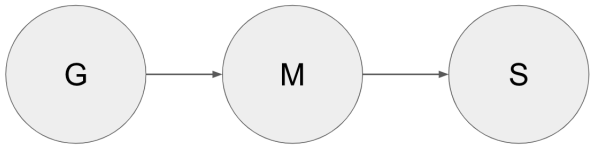


Figure 3.1: Causal mediation model. G = genotype; M = methylation; S = Sjögren's syndrome case status.

The CIT evaluates a set of statistical tests where rejection of the null supports a causal mediation relationship. The statistical tests evaluate the following necessary and sufficient conditions for the causal mediation model involving genotype "G", methylation "M", and case status "S",

1. $S \sim G$
2. $G \sim M|S$
3. $M \sim S|G$
4. $S \perp\!\!\!\perp G|M$,

where "$\sim$" denotes associated with and "$\perp\!\!\!\perp$" denotes independent of. The maximum p-value from these four statistical tests is the CIT p-value. See Millstein *et al.* and Appendix *B*.1 for additional details on the CIT [113]. The CIT genotype is encoded as 0, 1, or 2 copies of the reference allele, DNA methylation value is the batch-adjusted M-value, and SS is binary case status. False discovery rate was controlled at or under 5% using the permutation-based q-value developed and implemented by Millstein *et al.* [128, 129]. See Appendix *B*.1 for usage details of the CIT.

3.3 Results

Genome-wide DNA methylation profiles distinguish cases from non-cases

The 131 individuals in this study comprise of 64 SS cases and 67 non-cases. PCA of β-values adjusted for methylation array type and genetic ancestry revealed that principal component 1 (PC1) alone separated most cases from non-cases (Figure 3.2).

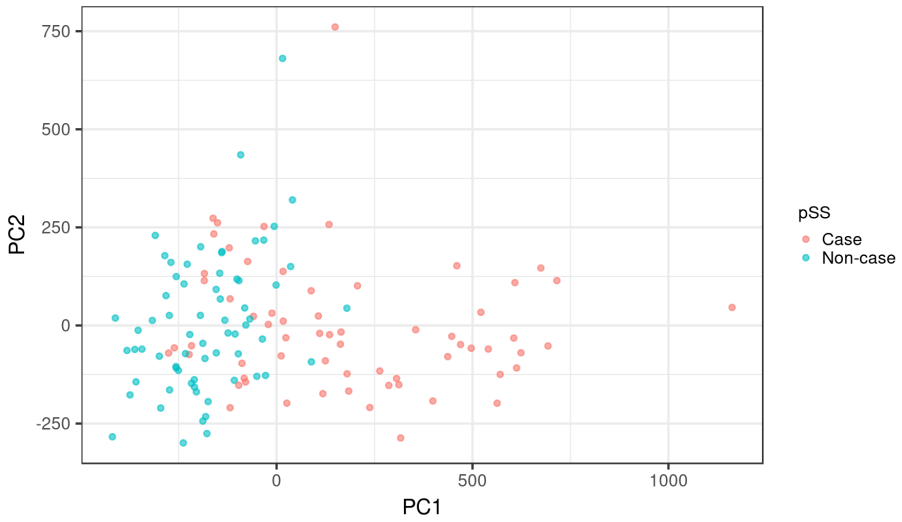


Figure 3.2: PCA of preprocessed, batch-corrected, β-values. SS case status, as determined by the 2016 ACR/EULAR diagnostic criteria, is indicated by color. Cases and non-cases show strong separation on PC1 values.

In contrast, PC2 did not provide as clear a separation. To understand this further, we analyzed the PC1 loadings, which specify the contribution of each CpG site to PC1 (see Appendix *B*.1). We observed that CpG sites in DMRs significantly contributed to PC1 on average, with an average absolute loading percentile of 94% (Figure 3.3C). In contrast, the average contribution of the DMR CpG sites to PC2 was relatively low. Together, this

indicated that case status explained a significant portion of variation in DNA methylation, and that PC1 captured most of this variation explained by case status.

Hypomethylation of genes involved in immune response

Analysis with *Bumphunter* identified 215 significant DMRs from 2,747 candidate "bumps". Of the 215 DMRs, 169 were hypermethylated regions and 46 were hypomethylated regions, in cases. Approximately 84% of DMRs were located in either promoters or gene bodies (Fig 2A), locations where differential methylation tends to influence transcription [124]. The top three DMR-contributing chromosomes were chromosomes 1, 6, and 17, and a majority of DMRs on chromosome 6 overlapped or surrounded the MHC (Figure 3.3B).

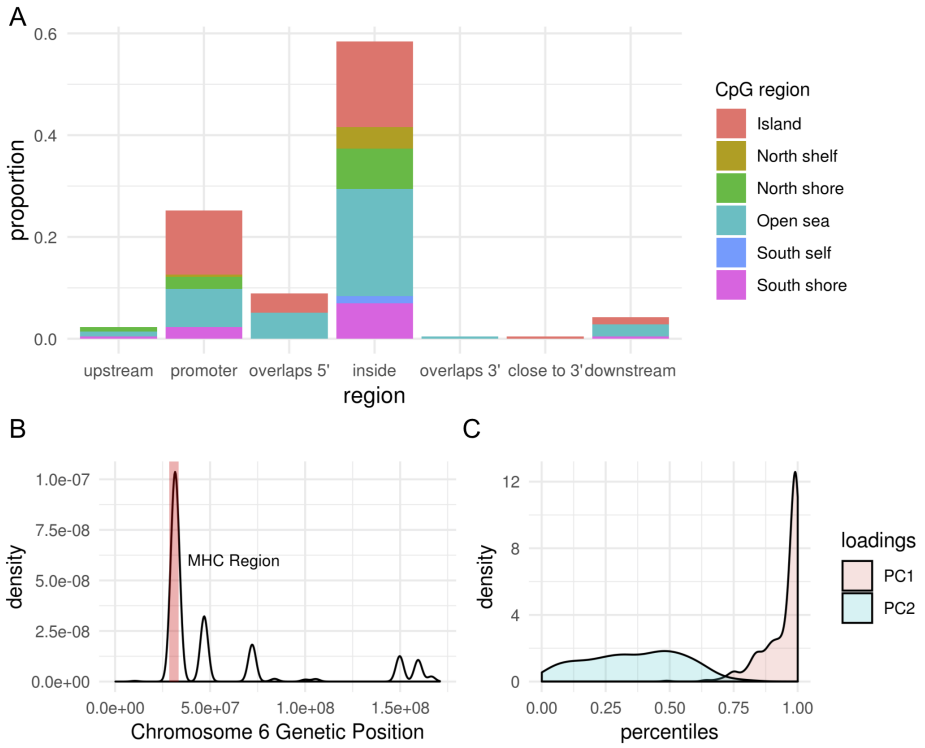


Figure 3.3: DMR characteristics. (A) Proportion of SS DMR locations relative to closest gene, and CpG type proportions at each DMR location; most DMRs are located either in the gene body (inside) or promoter, and most DMR CpG sites are either in the CpG island or the open sea. (B) Density plot of SS DMR locations on chromosome 6, where a DMR's location is represented by GRCh37 genetic coordinates of its first CpG site to last CpG site. The shaded red region denotes the MHC region. (C) Density plot of SS DMR CpG site loading percentiles for PC1 and PC2.

Genes near hypomethylated regions in cases were enriched for gene sets associated with immune function (Table 3.1), with the top gene sets almost exclusively related to immune response. This is expected given many DMRs were concentrated at the MHC. *IRF5*, which resides on chromosome 7 and is the strongest genetic risk factor for SS outside the MHC, was not the nearest gene for any DMRs. Of the 131 individuals in our study, 26 were in a previous LSG study by Cole *et al.*, which identified 57 genes whose promoters were hypomethylated in SS [104]. Eight of those 57 genes were among the hypomethylated DMR genes identified in this study. One DMR gene, *PSMB9*, was one of the 45 genes that previously demonstrated differential expression between SS cases and non-cases [130].

gene set	n	**overlap genes**	**p-value**	**adj. p-value**
SS DMP genes	8	*TAP1*, *LTA*, *PSMB8*, *AIM2*, *NCKAP1L*, *LINC00426*, *LCP2*, *ARHGAP25*	3.80E−18	1.71E−14
Antigen processing and presentation of endogenous peptide antigen	4	*HLA-E*, *HLA-B*, *TAP1*, *ABCB1*	1.60E−12	3.59E−9
Antigen processing and presentation of peptide antigen via MHC class I	6	*PSMB9*, *HLA-E*, *PSMB8*, *HLA-B*, *TAP1*, *ABCB1*	4.74E−12	5.53E−9
Antigen processing and presentation of endogenous antigen	4	*HLA-E*, *HLA-B*, *TAP1*, *ABCB1*	4.92E−12	5.53E−9
Negative regulation of innate immune response	4	*HLA-E*, *HLA-B*, *TAP1*, *NLRC5*	3.93E−10	3.24E−7
Negative regulation of natural killer cell mediated immunity	3	*HLA-E*, *HLA-B*, *TAP1*	4.32E−10	3.24E−7
Antigen processing and presentation via MHC class IB	3	*HLA-E*, *TAP1*, *ABCB1*	1.19E−10	7.64E−7
Positive regulation of antigen processing and presentation	3	*ABCB1*, *CCR7*, *TAP1*	1.58E−9	7.92E−7
Positive regulation of humoral immune response	3	*LTA*, *TNF*, *CCR7*	1.58E−9	7.92E−7
Negative regulation of cell killing	3	*HLA-B*, *HLA-E*, *TAP1*	2.66E−9	1.20E−6

Table 3.1: Top gene sets enriched for hypomethylated genes in SS. Candidate gene sets include GO gene sets from the Molecular Signatures Database [125], a set of genes previously reported to harbor differentially methylated CpG sites between SS cases and non-cases (SS DMP genes) [104], and a set of genes previously reported to be differentially expressed between SS cases and healthy controls (SS DE genes) [130]. n = number of overlapping genes; adj. p-value = Benjamini-Hochberg adjusted p-value.

In contrast to hypomethylated regions, genes near hypermethylated regions were enriched

for gene sets with miscellaneous functions, so the overall picture for hypermethylation in cases is less clear. Table 3.2 shows that the top gene sets were associated with nervous system development and cellular transport and signaling.

gene set	n	**overlap genes**	**p-value**	**adj. p-value**
Positive regulation of transporter activity	6	*WNK4*, *ATP1B2*, *RELN*, *HAP1*, *CACNB2*, *TRPC6*	1.36E−8	6.12E−5
Diencephalon development	5	*ETS1*, *GSX1*, *GLI2*, *HAP1*, *SLC6A4*	4.17E−7	9.38E−4
Hypothalamus development	3	*ETS1*, *GSX1*, *HAP1*	1.73E−6	2.59E−3
Vasoconstriction	3	*EDN3*, *HTR1A*, *SLC6A4*	3.29E−6	3.42E−3
Modulation of excitatory postsynaptic potential	3	*ZMYND8*, *CELF4*, *RELN*	4.38E−6	3.42E−3
Somatic stem cell population maintenance	4	*WNT98*, *LRP5*, *PBX1*, *BCL9*	4.59E−6	3.42E−3
Nerve development	4	*HOXB3*, *COL25A1*, *TFAP2A*, *SLITRK6*	5.32E−6	3.42E−3
Peptide Transport	4	*EDN3*, *SLC15A2*, *FAM3B*, *TAPBP*	7.06E−6	3.97E−3
Anatomical structure regression	2	*LRP5*, *GLI2*	1.03E−5	4.86E−3
ERBB2 signaling pathway	3	*ERBB2*, *GRB7*, *SHC1*	1.28E−5	4.86E−3

Table 3.2: Top gene sets enriched for hypermethylated genes in SS. Candidate gene sets include GO gene sets from the Molecular Signatures Database [125], a set of genes previously reported to harbor differentially methylated CpG sites between SS cases and non-cases (SS DMP genes) [104], and a set of genes previously reported to be differentially expressed between SS cases and healthy controls (SS DE genes) [130]. n = number of overlapping genes; adj. p-value = Benjamini-Hochberg adjusted p-value.

DNA methylation mediates the effect of MeQTL on SS at the MHC

We tested for association between average DMR methylation M-values and independent SNPs in surrounding ±250 kb neighborhoods for each, which yielded 20,754 unique DMR-SNP pairs. A total of 26 meQTL-DMR associations were identified with Benjamini-Hochberg adjusted p-value ≤ 0.05, with one each from chromosomes 3, 11, 12, 16, and two from chromosome 4; the rest were located within the MHC region on chromosome 6. Note that a

meQTL can be associated with multiple DMRs, and a DMR can be associated with multiple meQTL. The average meQTL-DMR distance was 153 kb, although this statistic was also partially dependent on the SNP pruning process (Figure 3.4C).

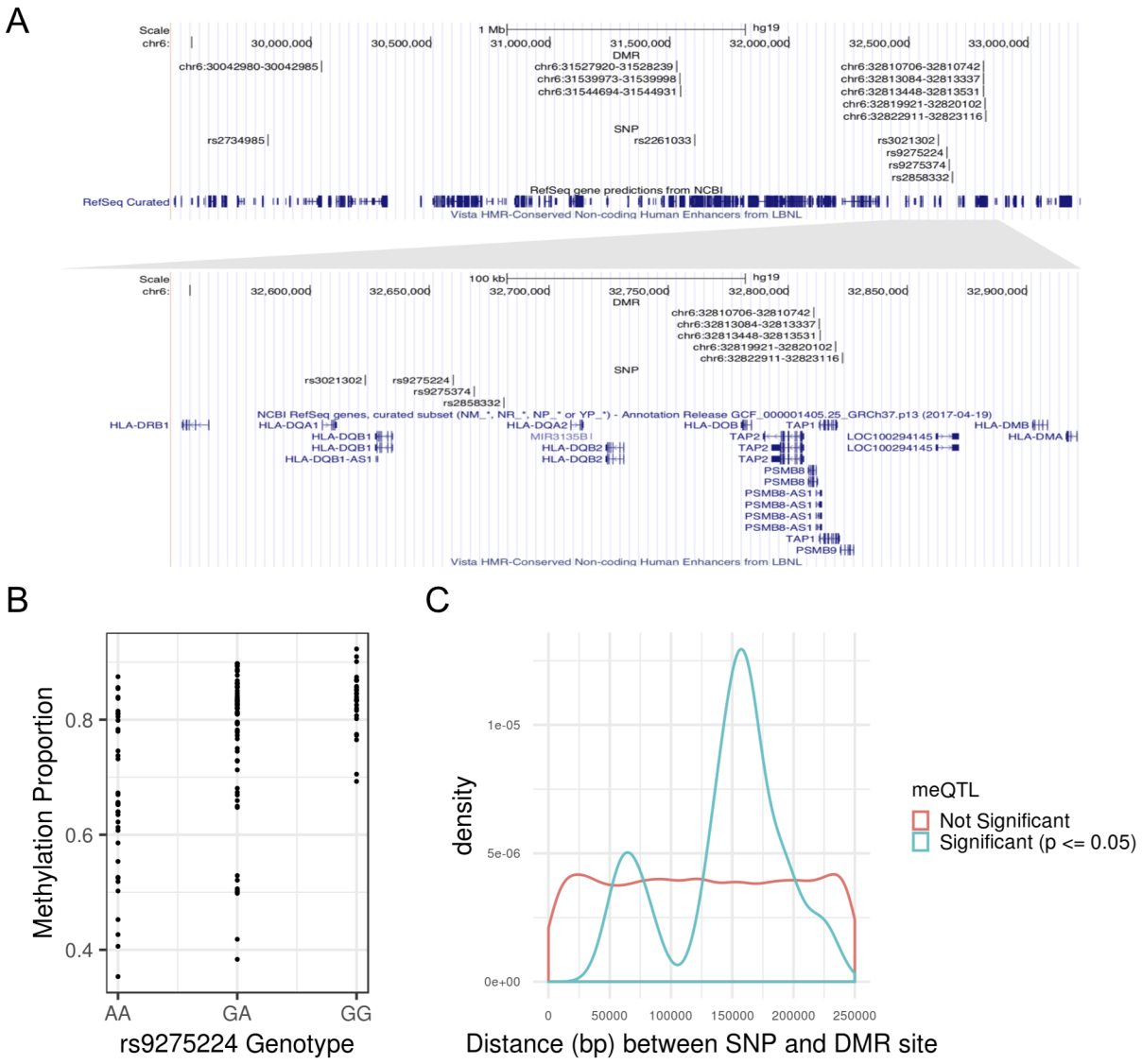


Figure 3.4: MeQTLs associated with SS DMR methylation M-values. (A) Sixteen of the 19 meQTL-DMR pairs supporting the causal mediation model lie in the MHC region. The top panel displays all DMRs and meQTLs on chromosome 6, and the bottom panel zooms in on a region with the meQTL-DMR pairs. Each DMR is specified by its chromosome, starting position, and ending position, in GRCh37 genetic coordinates. (B) SNP rs9275224 is a meQTL associated with average M-value of a DMR at 32,810,706 - 32,810,742 (GRCh37) on chromosome 6. (C) Density plot of associated and unassociated SNP-DMR pairs by absolute distance. While distance is approximately uniformly distributed for unassociated SNP-DMR pairs, the distance of associated SNP-DMR pairs is concentrated around 153 kb.

Of these 26 meQTL-DMR pairs, the CIT identified 19 with significant evidence supporting the causal mediation model (q-value ≤ 0.05), with one pair each from chromosomes 3, 12, and 16, and the rest from chromosome 6 (Table 3.3). At the MHC, the region spanning the *HLA-DQA1*, *HLA-DQB1*, and *HLA-DQA2* loci contained a high density of DMR-meQTL pairs, with five DMRs and four meQTLs (Figure 3.4A). In total, these meQTL-DMR pairs represent 12 unique DMRs and 9 unique SNPs.

SNP rs ID	**SNP position**	**A1**	**A2**	**SS DMR**	**distance**	**p.cit**	**q.cit**
rs9275224	32659878	G	A	chr6:32810706-32810742	150828	1.00E−3	2.11E−3
rs9275224	32659878	G	A	chr6:32819921-32820102	160043	1.00E−3	2.11E−3
rs9275224	32659878	G	A	chr6:32822911-32823116	163033	1.00E−3	2.11E−3
rs9275224	32659878	G	A	chr6:32813084-32813337	153206	1.00E−3	2.11E−3
rs9275224	32659878	G	A	chr6:32813448-32813531	153570	1.00E−3	2.11E−3
rs2261033	31603591	G	A	chr6:31544694-31544931	58660	1.17E−3	2.11E−3
rs2261033	31603591	G	A	chr6:31527920-31528239	75352	1.89E−3	2.11E−3
rs2734985	29818662	G	A	chr6:30042980-30042985	224318	1.99E−3	2.11E−3
rs9275374	32668526	A	G	chr6:32810706-32810742	142180	3.99E−3	3.47E−3
rs2261033	31603591	G	A	chr6:31539973-31539998	63593	5.25E−3	4.17E−3
rs13335209	87860446	A	C	chr16:87636539-87636594	223852	5.78E−3	4.30E−3
rs3021302	32623150	G	A	chr6:32810706-32810742	187556	7.84E−3	4.89E−3
rs3021302	32623150	G	A	chr6:32819921-32820102	196771	1.47E−2	9.29E−3
rs2858332	32681161	C	A	chr6:32819921-32820102	138760	1.63E−2	1.05E−2
rs17407659	24238010	A	G	chr12:24104007-24104115	133895	1.74E−2	1.35E−2
rs3021302	32623150	G	A	chr6:32813084-32813337	189934	2.49E−2	1.64E−2

rs3021302	32623150	G	A	chr6:32822911-32823116	199761	2.69E−2	1.74E−2
rs2858332	32681161	C	A	chr6:32810706-32810742	129545	3.36E−2	2.14E−2
rs76027985	112439220	G	A	chr3:112359488-112359557	79663	3.65E−2	2.44E−2

Table 3.3: Top causal inference test results for meQTLs of SS DMRs. All genetic positions are based on GRCh37 coordinates, and DMRs are denoted by the chromosome, start position, and end position. Distance refers to base pair distance between DMR and meQTL. A1 = allele 1; A2 = allele 2; SS DMR = differentially-methylated regions for Sjögren's syndrome; p.cit = causal inference test p-value; q.cit = permutation-based q-values from the causal inference test.

We provide further evidence that the meQTLs at the MHC, which we discovered in our 131 study subjects, are risk alleles for SS. Five of the six meQTLs at the MHC exhibited genome-wide significant associations with SS in a previous European GWAS (Figure B.6)[99]. Furthermore, there is evidence that three of these six meQTLs exhibited independent associations with SS based on results from logistic regression considering all six meQTLs included as predictors (Table B.1). These results add functional relevance to previously established SS-associated SNPs at the MHC.

3.4 Discussion

We investigated the casual mediation relationships between genetic variation, DNA methylation, and SS in the largest LSG study of SS to date. Despite comparing SS cases against symptomatic non-cases instead of healthy controls, our results show that significant differential methylation exists and is primarily driven by case status. The results of our DMR analysis are consistent with the general theme of hypomethylation previously seen in LSG [104]. Using the CIT applied to genotype and DNA methylation data from the same patients, we conclude that exists genetic control of differential methylation, especially at the MHC.

General hypomethylation of genomic regions involved in the immune response in LSG remains one of the most significant findings, with many DMRs located in the MHC region. Many of these hypomethylated genes have biological roles closely related to SS pathophysiology. For example, dendritic cells in the glands produce high levels of interferons [93], and *PSMB8* and *PSMB9*, whose expressions are induced by gamma interferon, were both hypomethylated in SS cases compared to non-cases. Genes *PSMB8* and *PSMB9* encode catalytic subunits of the immunoproteasome that is involved in peptide presentation on the surface of antigen-presenting cells [131]. Hypomethylation of *PSMB9* may have a causal role

in increasing expression levels in SS [130]. Previous studies have suggested that differential DNA methylation in SS could be controlled by B cells infiltrating the LSG, which in turn may affect the expression of inflammatory genes [106, 107]. However, since the LSG consists of a mixture of epithelial and inflammatory cells, it remains to be concluded which exact cell types contribute to the observed differential methylation.

Although the overall picture for hypermethylated regions in cases is less clear than that for hypomethylated regions, GSEA suggested some degree of neurological involvement in SS (Table 3.2). Peripheral neuropathy is the most common neurological complication of SS, but involvement of the central nervous system has also been observed, including cognitive disorder meningitis and optic neuritis [132]. The pathological mechanism by which SS leads to damage of the nervous system is not well-established, but it is thought to involve inflammatory infiltration of the dorsal root ganglia [93, 132].

Evidence of allele-specific methylation over extended genomic regions has been previously reported by tissue, developmental stage, and ancestry [133]. In this study, we identified DMRs whose methylation levels are under genetic control using the CIT. Twelve of the 215 DMRs demonstrated evidence of causal dependence on neighboring genotypes, with the majority residing in the MHC. Furthermore, 9 of the 16 DMRs in the MHC region showed evidence of mediation, suggesting a general theme of genetic control of DNA methylation at the MHC. However, the mechanism by which genetic variation influences methylation in the MHC remains unclear. Since HLA alleles are highly polymorphic, larger studies are needed to investigate whether established SS HLA risk alleles exhibit a mediation relationship with nearby DMRs. Table 3.3 shows multiple DMRs under genetic control of the same SNP (e.g. rs9275224) at the MHC, potentially reflecting genetic control by linkage disequilibrium (LD) blocks on multiple DMRs. Although we cannot definitely conclude this, Figure B.7 shows that regions of high DMR density partially overlap with genetic LD blocks.

Evidence that DNA methylation can mediate genetic risk, and specifically genetic risk conferred by the MHC, has been found in a number of other autoimmune diseases. Differential methylation encompassing exon 2 of *HLA-DRB1*15:01* has been shown in monocytes to the mediate effect of the *HLA-DRB1*15:01* allele on its expression and risk of multiple sclerosis [134]. In psoriasis, the majority of reported MeQTLs also reside in the MHC, although target CpG loci were located more than 500 kb away from their corresponding MeQTLs. Using the CIT, 11 SNP-CpG pairs were found to exhibit a methylation-mediated relationship with psoriasis in skin tissue [135]. In rheumatoid arthritis, DNA methylation levels were found to mediate genetic risk within the MHC in whole blood [136].

DNA methylation is currently thought to be influenced by genetic factors, age, environment and lifestyle, and tissue-type [137–140]. By distinguishing differentially methylated CpG sites under genetic control from those without, we provide information that could be essential for the potential therapeutic application of site-specific epigenetic editing for SS [114]. For example, it may be important to prioritize testing of CpG sites without genetic control first before moving on to those under genetic control. To date, DNA methyltransferase inhibitors and histone deacetylase inhibitors are the two classes of epidrugs approved by the FDA for clinical use in the United States, most commonly for cancer [141]. Side

effects are a major concern for these drugs because their effects are not locus specific, motivating a need for designing epidrugs with improved target specificity [142]. Partly due to a lack of understanding of the causal relationships involving epigenetic modifications and SS, methylation-modifying therapies have not been used clinically for SS [143].

In conclusion, this is the first study to conclude genetic control of differential DNA methylation in SS by performing a formal CIT on genotype and DNA methylation datasets obtained from 131 individuals from SICCA. We replicated the hypomethylation observed in many immune-related genes in cases, particularly those at the MHC. This study also suggested the potential involvement of neurological processes from the study of hypermethylated regions in cases. By performing CIT on DMRs and their nearby meQTLs, we found most DMRs at the MHC were mediators of nearby risk alleles for SS. We could not find similarly strong evidence of mediation for DMRs at other non-MHC locations. Through a formal study of the causal relationships between genetic variation, DNA methylation, and SS case status, we hope to provide valuable insights for any future endeavors to develop cite-specific methylation-modifying therapies for SS.

Chapter 4

Epigenetic stratification identifies clinically-relevant disease subgroups in Sjögren's syndrome with differential genetic risk at the major histocompatibility complex

4.1 Introduction

Primary Sjögren's syndrome (SS) is a common systemic autoimmune disease with a female-to-male ratio of 9:1. The hallmarks of SS are dryness of the mouth and eyes, fatigue, and joint pain. However, classification of SS based on these hallmarks alone remains challenging because these hallmarks are common in the general population [93]. SS is also a heterogenous disease, and there does not exist a formal criteria for separating cases into disease subgroups [116]. This heterogeneity poses a challenge for diagnosis, management and therapeutic development [144], and effective treatment options for SS are limited.

To address disease heterogeneity, a study based on the UK Primary SS Registry stratified patients by self-reported symptoms of depression, anxiety, pain, fatigue, and dryness [144]. Their study discovered four patient subgroups and demonstrated the importance of proper stratification for detection of treatment effects in clinical trials. However, since the clustering was based on disease symptoms, it is not certain whether these subgroups correspond to distinct pathobiology.

Gene expression, DNA methylation, and genetic variation have all been shown to capture disease variation [99, 130, 143]. In this study, we perform a cluster analysis of DNA methylation data from labial salivary glands (LSG) tissue collected from 64 primary SS cases and 67 symptomatic non-cases from the Sjögren's International Collaborative Clinical Alliance (SICCA) registry. LSG is a prominent target of autoimmune attack in SS and LSG

biopies are used for disease diagnosis and classification [93]. To address clustering in high dimensions, we apply a variational autoencoder (VAE) model to perform dimensionality reduction of DNA methylation data [145]. The VAE approach for dimensionality reduction is attractive for its ability to learn statistically independent latent variables in a smooth latent space [146], allowing more meaningful application of clustering distance measures. Following dimensionality reduction, we apply agglomerative hierarchical clustering to the latent variables to identify patient clusters.

We report the identification of patient clusters that partition cases into two subgroups, and investigate differences in clinical phenotype, genetic risk, and regions of variable methylation. Our multi-dimensional clinical data include serological assays, histopathologic examination, oral and ocular tests, and self-reported symptoms. We also investigate the effectiveness of each phenotypic criterion from the 2016 American College of Rheumatology/European League Against Rheumatism (ACR/EULAR) classification criteria in distinguishing severe from mild cases, providing a basis of potential revision.

4.2 Materials and Methods

Study subjects and clinical evaluation

Study subjects included 64 SS cases and 67 symptomatic non-cases, all of whom are female, non-Hispanic White individuals from SICCA, with well-characterized clinical phenotypic data. Phenotypic data include salivary, oral, ocular, serological test outcomes, and self-reported symptoms (Table C.1). These self-reported symptoms cover the categories of dryness, fatigue, pain, anxiety, and depression. All individuals from SICCA exhibit at least one symptom or sign related to SS, specifically symptoms of dry eyes or dry mouth, prior suspicion/diagnosis of SS, positive serum anti-SS-A, anti-SS-B, rheumatoid factor or antinuclear antibody result, increase in dental caries, bilateral parotid gland enlargement, or a possible diagnosis of secondary SS [115]. However, only primary SS cases are included in this study. Case-control status was determined according to the 2016 ACR/EULAR criteria for SS [116].

Methylotyping and preprocessing

DNA was extracted from the labial salivary gland tissue of each study subject as previously described [104]. DNA methylation was measured for each subject using the Illumina 450K Infinium Methylation BeadChip (450K) platform for 28 subjects and Infinium MethylationEPIC (EPIC) platform for 103 subjects. The 450K and EPIC chips allow for high-throughput interrogation of more than 450,000 and 850,000 highly informative CpGs sites respectively, spanning ∼22,000 genes across the genome.

Methylation data processing was performed using `Minfi`, a Bioconductor package for the analysis of Infinium DNA methylation microarrays [117]. Background subtraction with dye-

bias normalization was performed on methylated and unmethylated signals with the "noob" procedure, followed by quantile normalization with `preprocessQuantile` [118, 119].

For joint analysis of all 131 samples, the intersection of CpGs from 450K and EPIC chips were selected for analysis, resulting in an initial set of 452,832 CpGs. Probes where more than 5% of samples had a detection p-value > 0.01 were removed, to retain probes where signal is distinguishable from negative control probes. To remove probes with ambiguous methylation measurements due to incomplete binding between DNA strand and probe, probes with SNPs with minor allele frequency greater than 0% at either the probe site, CpG interrogation site, or single nucleotide extension were removed. Finally, cross-reactive probes, or probes with probe-binding specificity and polymorphic targets problems, were removed [33, 34]. The final preprocessed dataset consisted of 336,040 CpG sites. Since no subject had more than 5% of probes with detection p-value > 0.01, all 131 subjects were retained.

Both methylation measures of β-values and M-values were used for this study. A β-value is the ratio of the methylated probe intensity to the sum of methylated and unmethylated probe intensities, and reflects the proportion of methylation at a CpG site. The M-value can be derived from a β-value as $\log_2 \frac{\beta}{1-\beta}$, and was used for identifying DMRs due to less severe heteroscedasticity [27].

Genotyping and quality control

The subject genotypes were taken from the genotypes of the larger SICCA cohort, which was genotyped on the Illumina HumanOmni2.5-4v1 or Illumina HumanOmni25M-8v1-1 arrays from DNA extracted from whole blood. All quality control steps performed have been previously described [99]. The final genotype dataset consisted of 1,392,448 SNPs.

Removing unwanted DNA methylation variation

Since subjects were methylotyped on both the 450K and EPIC chip, we adjusted for batch effect due to array type (Figure C.1). We applied parametric empirical Bayes using `ComBat` from the `SVA` package to adjust β-values for array type [120, 121]. Since `ComBat` requires no missing values, missing methylation values were mean imputed per CpG site before adjustment, then missingness restored afterwards.

VAE summary

We use a VAE to perform a non-linear projection of methylation data onto a low dimensional latent space. The VAE achieves this by mapping input data to a distribution of latent variables whose samples are used to reconstruct the input data [145]. The VAE is comprised of an encoder that estimates the parameters of the latent variable distribution, and a decoder that attempts to reconstruct the data from the latent features. The encoder and decoder are typically parameterized by neural networks acting as effective function approximators.

We provide a brief review of VAE, and refer Kingma and Welling for complete details [145]. Given an input dataset $X = \{x_i\}_{i=1}^n$ where $x_i \in \mathbb{R}^p$ (in our case p is the number of CpG sites), the VAE learns a distribution of latent variables $z \in \mathbb{R}^m$ where $m < p$. Let $q_\phi(z|x)$ denote the latent variable distribution specified by an encoder with parameters ϕ, and let $p_\theta(x|z)$ denote the output distribution from a decoder with parameters θ. Then, the VAE method maximizes a lower bound of the log likelihood known as the evidence lower bound (ELBO)

$$\log p_\theta(x) \geq \mathbb{E}_{q_\phi(z|x)}[\log p_\theta(x|z)] - D_{KL}(q_\phi(z|x)||p(z)), \tag{4.1}$$

where $D_{KL}(q_\phi(z|x)||p(z))$ is the Kullback-Leibler (KL) divergence between distributions $q_\phi(z|x)$ and $p(z)$. The distribution $p(z)$ is chosen to be the standard multivariate normal distribution $\mathcal{N}(0, I)$. The ELBO terms have straightforward interpretations—maximizing the first term minimizes the reconstruction loss and minimizing the KL divergence constrains the latent variable distribution to be close to $p(z)$. Thus, $q_\phi(z|x)$ is chosen to belong to the multivariate normal distribution family with diagonal covariance, and the encoder estimates the mean and variance terms that specify the distribution. Reconstruction is learned by minimizing average binary cross entropy between $p_\theta(x|z)$ and input β-values, where $p_\theta(x|z)$ is chosen to be sigmoid activation for each CpG site.

The choice of $p(z)$ as a standard multivariate normal allows the VAE to learn latent variables with desirable properties. These properties are (1) statistical independence of latent variables and, depending on the decoder, (2) smoothness of the latent space [146]. In other words for (2), interpolation in the latent space corresponds to interpolation in the feature space of the data. This has led to the application of VAEs to extract meaningful latent factors from DNA methylation and RNA-seq data [147, 148]. For our application of clustering in the latent space, the smoothness property is important because it makes the distance measures more meaningful.

We use the VAE implementation *Tybalt* and its hyperparameters [148], with a few exceptions. In particular, we train with a batch size of 16 and a maximum of 50 epochs. We use the means outputted by the encoder as latent features of methylation data. Refer to Way and Greene [148] for implementation details.

Hierarchical clustering

All clustering was performed using agglomerative hierarchical clustering with Ward's minimum variance method as the link function [149]. At each merge iteration in hierarchical clustering, Ward's method merges the pair of clusters that leads to the minimum increase in total within-cluster variance after merging. Similar to other link functions such as complete linkage, Ward's method tends to produce more balanced dendrograms and is less sensitive to outliers. Euclidean distance between latent features is used for hierarchical clustering in the latent space of DNA methylation data. In contrast, the baseline hierarchical clus-

tering method uses the average absolute difference in β-values to compare a given pair of individuals.

Statistical Testing

The Wilcoxon Rank Sum test was used to test the difference in ordinal or continuous clinical phenotypes between severe and mild cases, and the Kruskal-Wallis test was used to test the difference between four patient clusters. The chi-square test of independence was used to test association between categorical clinical variables (i.e. nominal and dichotomous) and patient clusters or disease subgroup. Logistic regression was used for genetic risk allele association analysis for disease subgroups.

Focus score was not measured for patients whose LSG biopsy diagnosis was within normal limits, non-specific chronic inflammation, or sclerosing chronic sialadenitis, and their scores were assumed to be zero for statistical analysis. No other phenotype analyzed has more than two missing values, and missing values were omitted from statistical tests. Tear break-up times of greater than or equal to 10 seconds were considered healthy, so these times were truncated and set to 10 seconds.

Identification of differentially methylated regions

DMRs were identified using *bumphunter*, which identifies regions of CpG sites which are all hypermethylated or hypomethylated in one group of subjects compared to the other [123]. In this study, a candidate DMR is required to have at least two CpG sites and have an effect size of at least 1.0, where the effect size is the expected change in methylation from one group to the other. The linear regression specified for *bumphunter* was

$$M \sim outcome + array\,type, \tag{4.2}$$

controlling for array type. Here "M" is the M-value without batch correction, *array type* is an indicator variable for whether a subject was methylotyped on 450K, and *outcome* is whether the subject belongs to the "severe" or "mild" disease subgroup. The number of permutations was set at $B = 1,000$ for generating a null distribution of candidate DMRs for establishing significance, with `nullMethod = bootstrap` to control for the adjustment covariate. Significant DMRs were stringently selected as those with $fwerArea \leq 0.05$, defined as proportion of permutations with maximum bump area greater than the observed area for a DMR. `Minfi` was used to annotate each DMR with its nearest gene in base pairs, location relative to nearest gene, and location relative to nearest CpG island. Detailed description of each DMR gene available from National Center for Biotechnology Information were obtained with Biopython [150].

Gene set enrichment analysis

We restricted GSEA to genes differentially methylated at the promoter or gene body, given differential methylation at these regions has been shown to regulate gene expression [124]. To provide a qualitative picture of the biological processes impacted by differential methylation, DMR genes were tested for enrichment of gene ontology (GO) gene sets from the Molecular Signatures Database with the hypergeometric test [125]. The GO gene set totals 5,917, with 4,436 derived from biological process ontology, 580 from cellular component ontology, and 901 from molecular function ontology. Additionally, we included two gene sets consisting of genes shown to be differentially methylated or differentially expressed respectively, between SS cases and controls in LSG [104, 130]. We eliminated large gene sets numbering more than 100 genes for improved specificity GSEA results, retaining approximately 76% of gene sets. Since genes in the same pathway tend to be up or down-regulated together [127], GSEA was performed separately for hypermethylated and hypomethylated DMR genes. False discovery rate was controlled with the Benjamini-Hochberg procedure [126]. We report the top 10 enrichment results by statistical significance as sufficient to provide an overall biological picture, and avoid interpreting the rest of the results, given GSEA with the hypergeometric test makes unrealistic independence assumptions between genes [151].

4.3 Results

Identification of patient clusters

We identified patient clusters in two steps: (1) performing dimensionality reduction of methylation data onto a latent space using a VAE model, and (2) applying hierarchical clustering to latent variables to identify patient clusters. Following guidelines established by Way and Greene [148], the entire dataset was split into a 9:1 train validation ratio. Figure $C.2$ shows that both training and validation VAE loss converged after around 40 epochs.

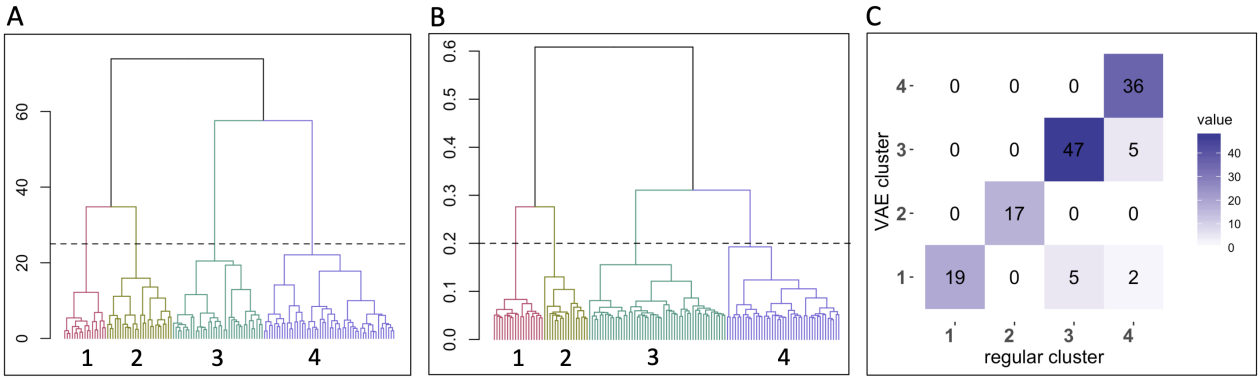


Figure 4.1: Clustering of patient DNA methylation profiles. (A) Dendrogram from hierarchical clustering of VAE-based latent variables, with cluster numbering in the bottom and dendrogram cut height indicated by horizontal dotted line. (B) Dendrogram of baseline hierarchical clustering of DNA methylation profiles (see Section 4.2), with same annotations as (A). (C) Confusion matrix showing clustering agreement between results from (A) compared to that of (B).

Hierarchical clustering identified four robust patient clusters with distinct DNA methylation profiles (Figure 4.1A), with high agreement with clusters from a baseline hierarchical clustering approach (Figure 4.1B-C). A significantly higher proportion of subjects in clusters 1 and 2 are cases compared to those in clusters 3 and 4 (Table 4.1; chi-square test of independence p-value = 2.43E−11). However, clusters 3 and 4 also contained a non-negligible proportion (i.e. 21.2% and 36.1% respectively) of cases. The dendrogram heights in Figure 4.1A suggest that clusters 1 and 2 are more distant to clusters 3 and 4, than between themselves (e.g. cluster 1 compared to cluster 2). Thus, the imperfect partitioning between cases and non-cases in clusters 1 and 2 versus clusters 3 and 4 reflects some disagreement with the 2016 ACR/EULAR classification criteria for SS.

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 3**
Case	23	17	11	13
Control	3	0	41	23

Table 4.1: Subject stratification into patient clusters. Clustering result of 131 study subjects from VAE-based clustering analysis of DNA methylation profiles. The 2016 ACR/EULAR SS classification criteria was used to establish case/non-case status [116].

We identified smoking, age of SS onset, and anticholinergic drug use in our clinical data as potentially having unwanted influence on clustering analysis [152–154]. Statistical analysis for history of cigarette use (p-value = 0.72), cigarettes smoked per day (p-value = 0.17),

self-reported age of SS onset (p-value = 0.29), and anticholinergic drug use (p-value = 0.18) showed no significant differences by patient clusters.

Clinical phenotype analysis by cluster and disease subgroup

Analysis of clinical phenotypes revealed that subjects in clusters 1 and 2 have a higher symptom burden on average compared to subjects in clusters 3 and 4 (Table 4.2). However, except for mouth pain, there were no significant differences in self-reported symptoms pertaining to dryness, fatigue, pain, anxiety, or depression between the clusters (Table C.2; Figure C.3). Patient cluster 2, which is entirely comprised of SS cases, has on average the most severe clinical phenotypes among the four clusters (Table 4.2). As a result, individuals in clusters 1 and 2 are at higher risk of lymphoma compared to those in clusters 3 and 4 [93]. Although no subjects have physician-confirmed lymphoma, clusters 1 and 2 accounted for all cases of germinal center-like formation, an indicator of lymphoma. However, there were no significant differences in prevalence of other extraglandular disorders (i.e. thyroid, liver, kidney, and other systemic disease) between the clusters [115].

Since clusters 1 and 2 have a higher overall symptom burden and a distinct DNA methylation profile from that of clusters 3 and 4, we investigated whether cases in clusters 1 and 2 ($n = 40$) and cases in clusters 3 and 4 ($n = 24$) constitute clinically distinct disease subgroups. Table 4.3 and Figure 4.2 show a significantly higher proportion of cases from clusters 1 and 2 are positive for anti-SS-A, anti-SS-B, and rheumatoid factor, germinal center-like formation test results. Cases from clusters 1 and 2 also tend to have more severe antinuclear antibody titer results, immunoglobulin G results, left eye ocular SICCA scores, and focus scores, at a $\alpha = 0.05$ significance level (Figure 4.2; Table 4.3). However, there was no evidence to suggest the disease subgroups have different severities in self-reported symptoms (Table C.3), nor were there significant differences in the prevalence of extraglandular disorders (Table 4.3). Together, this analysis suggests that SS cases in clusters 3 and 4 are clinically more similar to symptomatic non-cases than to cases in subgroups 1 and 2. We will refer to cases in clusters 1 and 2 as "severe cases" and the other cases as "mild cases".

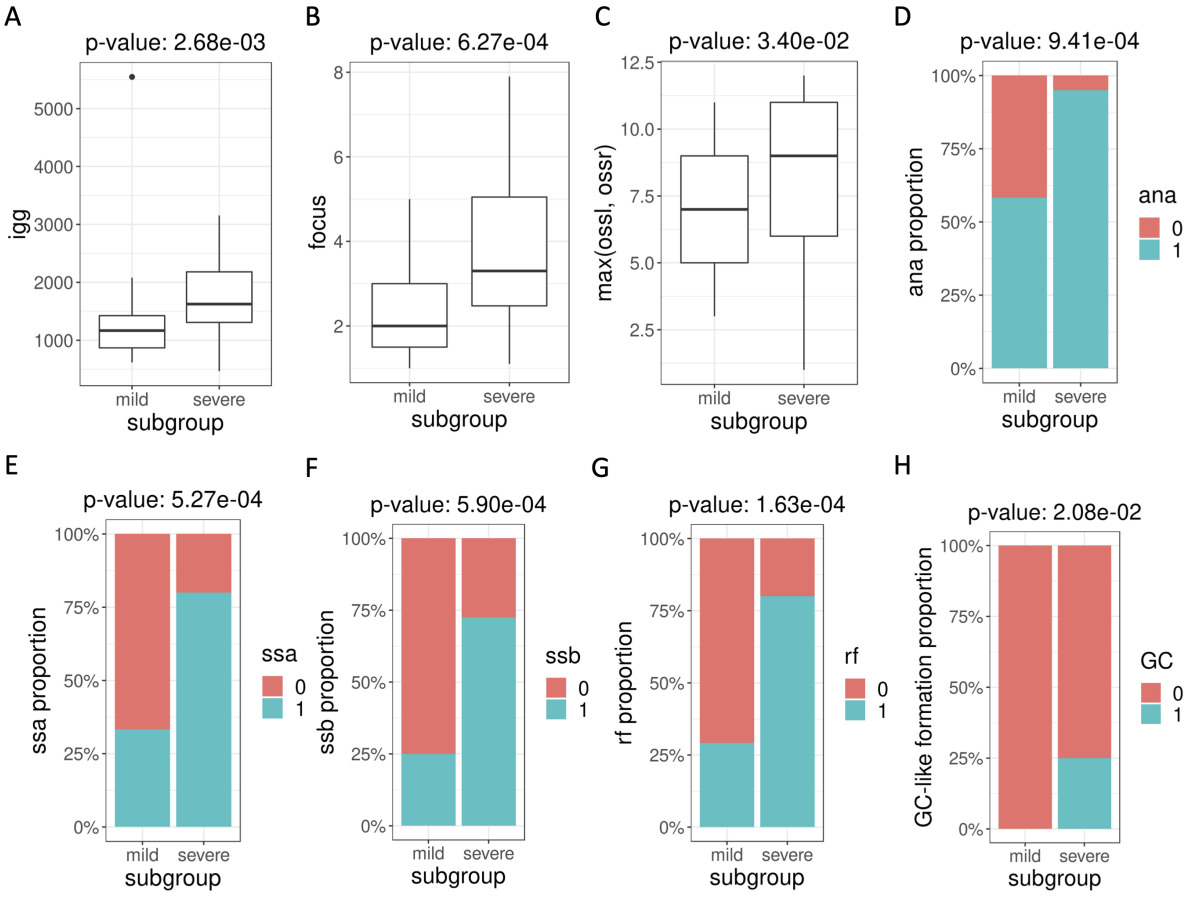


Figure 4.2: Clinical phenotype comparison between severe cases and mild cases. Plots for clinical phenotypes that are different at $\alpha = 0.05$ significance level between severe cases and mild cases. Severe cases are from clusters 1 and 2 and mild cases are from clusters 3 and 4. P-values from Wilcoxon rank sum test or chi-square test of independence shown above each subplot. (A) Immunoglobulin G box plot. (B) Focus score box plot. (C) Ocular SICCA socre (maximum of left and right eyes). (D) Detection of antinuclear antibody at 1:40 concentration level bar plot. (E) Anti-SS-A bar plot. (F) Anti-SS-B bar plot. (G) Rheumatoid factor bar plot. (H) Germinal center (GC)-like formation bar plot.

Clinical phenotype analysis highlighted areas of disagreement between the clustering analysis and the ACR/EULAR classification criteria. The ACR/EULAR criteria is based on the total score from meeting a list of phenotype requirements, where each requirement contributes a score [116]. For the four patient clusters, cluster 2 has the highest proportion of individuals meeting each phenotype requirement (Table 4.2). In cluster 1, which has a slightly lower proportion of cases, these proportions decrease the most for anti-SS-A and Schirmer's test, relative to proportions in cluster 2. Further decline in proportions was observed in

clusters 3 and 4. For disease subgroups, anti-SS-A is the most discriminative phenotype between severe and mild cases (p-value = 5.27E−4), and is the only one that is significant at the level $\alpha = 0.05$ (Table 4.3). Discriminatory power is followed by unstimulated whole saliva flow rate, ocular staining score, Schirmer's test, and focus score, in decreasing order. The focus score-based requirement is completely unable to distinguish between the severe and mild cases, since all cases have focus scores greater than 1. However, severe cases have significantly higher focus scores compared to mild cases (Table 4.3).

	Cluster 1	**Cluster 2**	**Cluster 3**	**Cluster 4**	**p-value**
ana	0.85	1.00	0.40	0.44	**2.02E−6**
igg	1370.31	2130.53	995.69	1222.33	**1.48E−8**
c3	117.65	118.71	122.06	128.42	1.89E−1
c4	24.42	22.00	26.92	27.89	**2.81E−2**
ssb	0.50	0.94	0.06	0.08	**6.38E−14**
rf	0.65	0.94	0.13	0.17	**2.33E−11**
tbul_time	4.88	4.53	7.71	6.11	**1.45E−4**
tbur_time	4.88	3.59	7.52	5.58	**7.27E−6**
uws	0.42	0.19	0.67	0.55	**3.80E−4**
focus	3.36	4.77	1.68	2.25	**5.51E−6**
max(ossl, ossr)	7.58	9.18	4.33	4.54	**1.46E−7**
rparenlg	0.08	0.18	0.04	0.00	5.42E−2
lparenlg	0.08	0.18	0.06	0.03	2.45E−1
drymouth	0.88	1.00	0.98	0.89	1.33E−1
liqmouth	0.85	0.94	0.65	0.69	5.75E−2
dryeye	0.92	0.94	0.88	0.92	8.80E−1
lymphoma	0.00	0.00	0.00	0.00	NA
GC_like_formation	0.19	0.29	0.00	0.00	**2.93E−5**
thyroid	0.27	0.12	0.18	0.11	3.83E−1
liver	0.04	0.00	0.06	0.00	3.79E−1
kidney	0.00	0.00	0.02	0.06	4.36E−1
othersys	0.00	0.00	0.02	0.00	6.68E−1
pSS	0.88	1.00	0.21	0.36	**2.43E−11**
Satisfies 2016 ACR/EULAR SS criteria [116]					
LSG with focal lymphocytic sialadenitis and focus score ≥ 1	0.88	1.00	0.23	0.42	**1.73E−10**

anti-SS-A +	0.62	0.94	0.06	0.14	**8.52E−14**
Ocular staining score ≥ 5 on at least one eye	0.88	0.94	0.52	0.47	**1.08E−4**
Schirmer ≤ 5 mm/5min on at least one eye	0.12	0.41	0.12	0.11	**1.74E−2**
Unstimulated whole saliva flow rate ≤ 0.1 ml/min	0.77	0.94	0.46	0.50	**7.32E−4**

Table 4.2: Clinical phenotype averages by patient cluster, determined from the VAE-based clustering analysis. P-values were computed using Kruskal-Wallis test for ordinal or continuous clinical phenotypes, and computed using chi-square test of independence for categorical or binary phenotypes. Significant p-values at $\alpha = 0.05$ are bolded. Refer to Table $C.1$ for key of clinical phenotype abbreviations. Note the average is equivalent to proportion for binary phenotypes.

	Mild cases	**Severe cases**	**p-value**
ana	0.58	0.95	**9.41E−4**
igg	1353.04	1716.55	**2.68E−3**
c3	131.33	118.10	8.29E−2
c4	25.63	23.30	5.70E−2
ssb	0.25	0.73	**5.90E−4**
rf	0.29	0.80	**1.63E−4**
tbul_time	6.04	4.60	9.94E−2
tbur_time	5.29	4.28	1.97E−1
uws	0.43	0.29	4.36E−1
focus	2.44	3.96	**6.27E−4**
max(ossl, ossr)	7.04	8.38	**3.40E−2**
rparenlg	0.04	0.13	5.06E−1
lparenlg	0.04	0.13	5.06E−1
drymouth	0.96	0.93	1.00E0
liqmouth	0.71	0.90	1.04E−1
dryeye	0.96	0.93	1.00E0
lymphoma	0.00	0.00	NA
GC_like_formation	0.00	0.25	**2.08E−2**
thyroid	0.13	0.20	6.69E−1

liver	0.04	0.03	1.00E0
kidney	0.08	0.00	2.66E−1
othersys	0.04	0.00	7.95E−1
Satisfies 2016 ACR/EULAR SS criteria [116]			
LSG with focal lymphocytic sialadenitis and focus score ≥ 1	1.00	1.00	NA
anti-SS-A +	0.33	0.80	**5.27E−4**
Ocular staining score ≥ 5 on at least one eye	0.79	0.93	2.42E−1
Schirmer ≤ 5 mm/5min on at least one eye	0.21	0.25	9.39E−1
Unstimulated whole saliva flow rate ≤ 0.1 ml/min	0.67	0.85	1.60E−1

Table 4.3: Clinical phenotype averages for severe cases and mild cases. Severe cases belong to clusters 1 and 2 and mild cases belong to clusters 3 and 4 from the VAE-based clustering analysis. P-values were computed using Wilcoxon rank sum test for ordinal or continuous clinical phenotypes, and computed using chi-square test of independence for categorical or binary phenotypes. Significant p-values at $\alpha = 0.05$ are bolded. Refer to Table $C.1$ for key of clinical phenotype abbreviations. Note the average is equivalent to proportion for binary phenotypes.

Differential genetic risk between patient clusters and disease subgroups

We investigated whether patient clusters and disease subgroups exhibit differential risk at established genetic risk loci for SS [99, 100, 155], where each loci was associated with SS at genome-wide significance level (p-value $<$ 5.0E−8) [156]. Of the SNPs in our genotype data, four (rs485497, rs9271573, rs3021302, rs9275572) demonstrated significant difference in risk allele frequency between the patient clusters, with higher frequencies in the high symptom burden clusters compared to those in low symptom burden clusters (Table 4.4). Additionally, three of the four SNPs reside in the major histocompatibility complex (MHC), tag the HLA genes *HLA-DRB1* and *HLA-DQA1*, and have p-values $<$ 0.01. Lastly, the ordering of risk allele frequencies follow the rough ordering of phenotype severity. Cluster 2 has the highest risk allele frequencies, followed by cluster 1, cluster 4, and cluster 3, in decreasing order. Comparison of disease subgroups show continued association for SNPs rs9271573 (OR = 3.31, p-value = 3.27E−3) and rs3021302 (OR = 4.33, 5.21E−3), treating severe cases as "cases" and mild cases as "controls" in logistic regression (Table 4.5). As expected, severe cases have higher risk allele frequencies. These analyses provide a genetic basis for patient clusters and disease subgroups.

Gene	**Chr**	**SNP**	**Ref**	**Cluster 1**	**Cluster 2**	**Cluster 3**	**Cluster 4**	**p-value**
STAT4	2	rs11889341	A	0.29	0.35	0.18	0.22	1.66E−1
STAT4	2	rs7574865	A	0.29	0.32	0.19	0.24	3.08E − 1
IL12A	3	rs485497	G	0.50	0.56	0.35	0.54	2.89E − 2
HLA-DRB1, HLA-DQA1	6	rs9271573	A	0.69	0.71	0.35	0.47	3.16E − 5
HLA-DQA1, HLA-DQB1	6	rs3021302	G	0.37	0.41	0.13	0.19	3.05E − 4
HLA-DQB1, HLA-DQA2	6	rs9275572	A	0.65	0.65	0.33	0.44	1.64E − 4
IRF5-TNPO3	7	rs3823536	A	0.54	0.59	0.49	0.39	1.80E − 1
IRF5-TNPO3	7	rs3807306	A	0.54	0.65	0.51	0.42	1.56E − 1
IRF5-TNPO3	7	rs59110799	A	0.14	0.29	0.18	0.10	7.30E − 2
OAS1	12	rs10774671	G	0.31	0.44	0.43	0.40	4.66E − 1

Table 4.4: Frequency of SS genetic risk alleles by patient cluster. Minor allele frequency of established SS genetic risk alleles [99, 100, 155], by patient cluster from VAE-based clustering analysis. Chr = chromosome; Ref = risk allele. P-values computed from chi-square test of independence and p-values significant at $\alpha = 0.05$ are bolded.

Gene	Chr	SNP	Ref	Mild cases	Severe cases	OR (95% CI)	p-value
STAT4	2	rs11889341	A	0.29	0.31	1.10 (0.52 – 2.33)	8.11E−1
STAT4	2	rs7574865	A	0.29	0.30	1.04 (0.48 – 2.26)	9.22E−1
IL12A	3	rs485497	G	0.44	0.53	1.58 (0.69 – 3.62)	2.82E−1
HLA-DRB1, HLA-DQA1	6	rs9271573	A	0.42	0.71	3.31 (1.49 – 7.34)	3.27E−3
HLA-DQA1, HLA-DQB1	6	rs3021302	G	0.19	0.40	4.33 (1.55 – 12.15)	5.21E−3
HLA-DQB1, HLA-DQA2	6	rs9275572	A	0.48	0.66	2.04 (0.98 – 4.23)	5.57E−2
IRF5-TNPO3	7	rs3823536	A	0.46	0.56	1.73 (0.75 – 4.00)	2.00E−1
IRF5-TNPO3	7	rs3807306	A	0.50	0.59	1.49 (0.69 – 3.24)	3.11E−1
IRF5-TNPO3	7	rs59110799	A	0.17	0.22	1.42 (0.54 – 3.69)	4.74E−1
OAS1	12	rs10774671	G	0.35	0.35	0.982 (0.47 – 2.07)	9.62E−1

Table 4.5: Association analysis of SS genetic risk loci with disease subgroups. Logistic regression evaluation of association between established SS genetic risk loci [99, 100, 155] and disease severity status of cases. The columns "Mild cases" and "Severe cases" contain minor allele frequencies for the mild and severe cases respectively. Chr = chromosome; OR (95% CI) = odd ratio (95% confidence interval); Ref = risk allele. P-values significant at $\alpha = 0.05$ are bolded.

Differentially methylated regions between disease subgroups

We identified DMRs that underlie methylation differences between severe SS cases and mild cases (Figure 4.1). The overall result is a general hypomethylation at the MHC and a general hypermethylation in other areas of the genome (Figure 4.3). Specifically, we identified a total of 207 significant DMRs from 826 candidate DMRs, with 41 hypomethylated regions and 166 hypermethylated regions, in severe cases relative to mild cases.

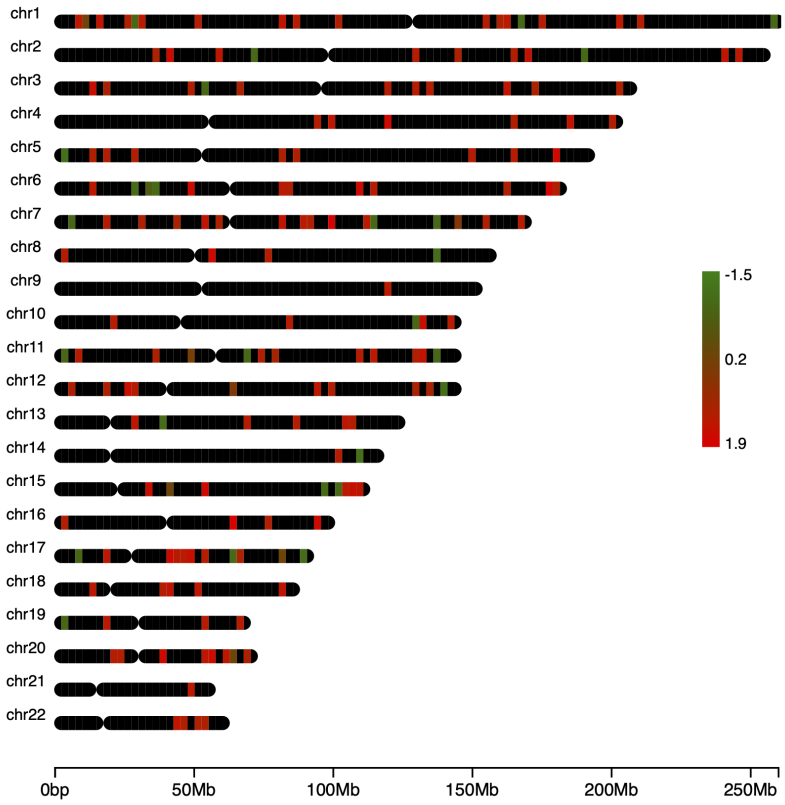


Figure 4.3: Chromosome heatmap of DMRs. Statistically significant DMRs ($fwerArea \leq 0.05$) between severe and mild cases on a chromosome heatmap, with red/green indicating hypermethylation/hypomethylation in severe cases relative to mild cases.

Gene set enrichment analysis (GSEA) of hypomethylated genes revealed an overall enrichment of immune biological processes (Table 4.6). The top result is the set of genes whose promoters were shown by Cole *et al.* to be differentially methylated in 13 SS cases and 13 symptomatic non-cases from the same SICCA study. This further supports the notion that mild SS cases have DNA methylation profiles more similar to that of symptomatic non-cases. Other enriched biological processes such as response to type I interferon and T cell migration, are known to be involved in the pathobiology of SS [94]. For GSEA of hypermethylated genes, many neurological processes appeared in the top 10 results (Table 4.7). Another top enriched gene set is the regulation of cell fate commitment, which could potentially reflect differences in proportion of immune cells that infiltrated the LSG in SS patients [143].

gene set	n	**overlap genes**	**p-value**	**adj. p-value**
SS DMP genes	6	*AIM2, CTSZ, PSMB8, TAP1, LCP2, ARHGAP25*	5.21E−13	2.35E−9
Response to Type I interferon	5	*HLA-E, XAF1, PSMB8, ISG20, IRF5*	2.04E−10	4.60E−7
Mast cell activation	3	*PIK3CD, LCP2, RHOH*	1.00E−8	1.50E−5
Cellular extravasation	3	*PIK3CD, TNF, ITGB2*	2.11E−8	2.38E−5
Positive regulation of monooxygenase activity	3	*TNF, GDNF, NPR3*	3.96E−8	3.03E−5
Tumor necrosis factor receptor binding	3	*TNFSF13B, TNF, TRAF3*	4.56E−8	3.03E−5
Receptor metabolic process	4	*GRB2, TNF, ITGB2, CD81*	4.72E−8	3.03E−5
Antigen processing and presentation of peptide antigen via MHC class I	4	*HLA-E, PSMB9, PSMB8, TAP1*	7.98E−8	4.48E−5
Spleen development	3	*PSMB9, PITX2, PKN1*	1.36E−7	6.78E−5
Regulation of immunoglobulin production	3	*HLA-E, TNF, PKN1*	2.92E−7	1.10E−4

Table 4.6: Top gene sets enriched for hypomethylated genes. Candidate gene sets include GO gene sets from the Molecular Signatures Database [125], a set of genes previously reported to harbor differentially methylated CpG sites between SS cases and non-cases (SS DMP genes) [104], and a set of genes previously reported to be differentially expressed between SS cases and healthy controls (SS DE genes) [130]. n = number of overlapping genes; adj. p-value = Benjamini-Hochberg adjusted p-value.

gene set	n	**overlap genes**	**p-value**	**adj. p-value**
Autonomic nervous system development	5	*EDNRB, TFAP2A, PHACTR4, HAND2, NAV2*	9.21E−9	4.14E−5
Odontogenesis of dentin containing tooth	5	*SOSTDC1, HAND2, HTRA1, GLI2, BMP7*	3.21E−7	7.21E−4
Neural crest cell migration	4	*EDNRB, PHACTR4, HAND2, SEMA3E*	1.15E−6	1.45E−3
Response to auditory stimulus	3	*CNTNAP2, FOXP2, TACR1*	1.35E−6	1.45E−3
Embryonic camera type eye morphogenesis	3	*TFAP2A, PHACTR4, BMP7*	1.61E−6	1.45E−3
Regulation of cell fate commitment	3	*DUSP6, SOSTDC1, NKX6-2*	2.25E−6	1.45E−3
Dynein binding	3	*RAB11FIP3, SNCA, RILP*	2.25E−6	1.45E−3
Peripheral nervous system development	4	*EDNRB, TFAP2A, HAND2, ERBB2*	5.24E−6	2.09E−3
Embryonic eye morphogenesis	3	*TFAP2A, PHACTR4, BMP7*	6.04E−6	2.09E−3
Ventral spinal cord interneuron specification	2	*NKX6-2, GLI2*	7.34E−6	2.09E−3

Table 4.7: Top gene sets enriched for hypermethylated genes. Candidate gene sets include GO gene sets from the Molecular Signatures Database [125], a set of genes previously reported to harbor differentially methylated CpG sites between SS cases and non-cases (SS DMP genes) [104], and a set of genes previously reported to be differentially expressed between SS cases and healthy controls (SS DE genes) [130]. n = number of overlapping genes; adj. p-value = Benjamini-Hochberg adjusted p-value.

Principal component analysis (PCA) of methylation latent variables revealed that the first principal component (PC1) separates the high symptom burden clusters 1 and 2 from low symptom burden clusters 3 and 4 (Figure 4.4A), with clusters ordered roughly according to phenotype severity on PC1. The observed ordering suggests that PC1 represents methylation variation in DMRs common to all clusters. PC2 separates cluster 4 from the rest of the clusters, and may represent differential methylation that is specific to cluster 4. Together, this suggests that in the majority of highly variable CpG sites, all four patient clusters have varying degrees of methylation that distinguish each other. Only some CpG sites are

uniquely differentially methylated in one cluster compared to the rest, which may be the case for cluster 4. We further support this interpretation by investigating CpG sites at the MHC that are differentially methylated between severe and mild cases. A dot plot of the average methylation levels indeed shows that each cluster has varying degrees of methylation, with cluster 2 experiencing the most severe hypomethylation, followed by cluster 1, cluster 4, and cluster 3, in decreasing order (Figure 4.4B). At the MHC, more severe hypomethylation generally corresponds with more severe clinical phenotypes.

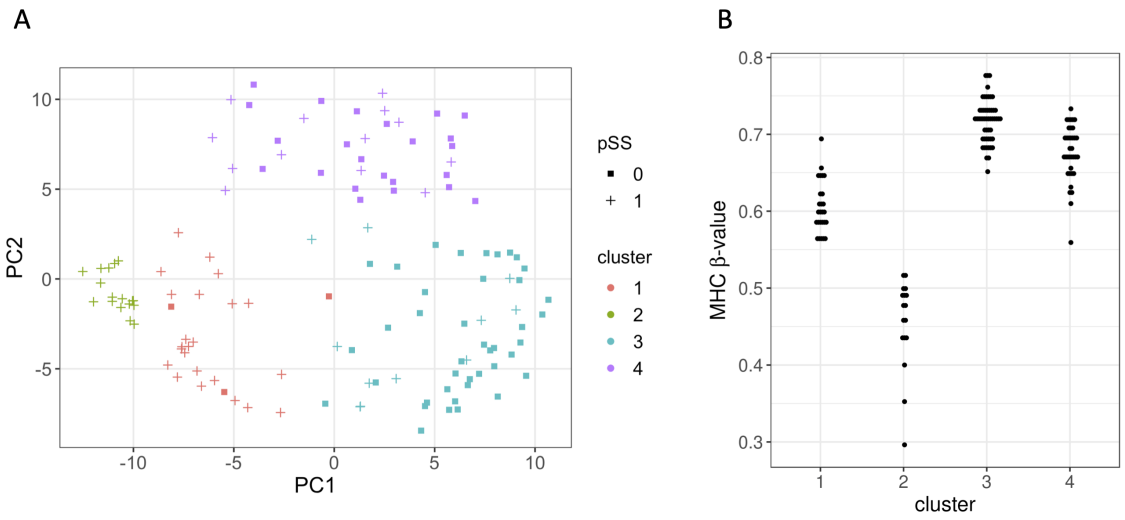


Figure 4.4: Analysis of differential methylation among patients. (A) PCA plot of VAE-based latent variables for all 131 study subjects, with patient cluster indicated by color and primary SS (pSS) status indicated by shape. (B) Dot plot of average β-value over DMR CpG sites at the MHC region, by patient cluster, where each dot represents an individual's average β-value at the MHC.

4.4 Discussion

In this study, we performed a cluster analysis on SS cases and symptomatic non-cases based on DNA methylation profiles of the LSG. The analysis yielded four robust patient clusters that partitioned cases into two disease subgroups. Differential methylation further revealed cluster-specific levels of methylation at the MHC. Together this demonstrates the effectiveness of DNA methylation in capturing disease variation at a high resolution. Specifically, DNA methylation was able to separate mild cases from severe cases while the 2016 ACR/EULAR classification criteria considers them as one group. These two disease subgroups differ significantly across many key clinical phenotypes and have a genetic contribution. It is possible that some mild cases are biologically closer to symptomatic non-cases.

We investigated the effectiveness of each ACR/EULAR phenotype criterion in separating mild from severe cases. Since every case had focus scores at or greater than 1, this threshold was unable to separate mild cases from severe cases. In contrast, a significantly higher proportion of severe cases have positive anti-SS-A test result compared to mild cases. The effectiveness of the remaining criteria falls somewhere in between. A potential revision to the classification criteria to isolate severe cases may involve increasing the criteria thresholds, and using the original criteria to distinguish healthy controls from cases. Besides revision to the current classification criteria, targeted DNA methylation assays aimed at well-defined DMRs may provide additional resolution. Commercial DNA methylation assays have already seen applications for a variety of cancers, such as cancers of the bladder, breast, liver, lung, etc [157].

It is possible that the high and low symptom burden patient groups identified by Tarn *et al.* corresponds to the severe and mild cases identified in this study respectively [144]. Although we did not observe disease subgroups with dryness or pain as dominant symptoms, this is possibly due to smaller sample sizes. By including symptomatic non-cases in our study, we were able to see that mild cases are in fact similar to symptomatic non-cases, instead of belonging to a patient cluster of its own.

A higher presence of antibodies appeared to co-occur with hypomethylation of the MHC and a higher frequency of genetic risk alleles at the MHC. MHC associations with autoantibody manifestations have also been demonstrated in European systemic lupus erythematosus [158], which is known to co-occur with secondary SS [93]. This further suggests a functional link between antibody concentration and the genetics and epigenetics of the MHC, although a formal causal study may be required. Additionally, without temporal DNA methylation measurements, it is difficult to determine whether varying DNA methylation levels is a cause of or a consequence of different phenotypic severities. It remains to be determined whether differential methylation is due to differences in cellular infiltration of the LSG, or differential methylation between the same cell types. The biological role of hypermethylation in the LSGs of severe compared to mild cases also remains unclear. GSEA of hypermethylated genes in severe cases suggests the presence of neurological complications (Table 4.7), which although well described in SS [132], does not have an established relationship with DNA methylation. Lastly, since differential methylation has also been reported between SS cases and controls in CD4+ T cells, CD19+ B cells, and whole blood [102, 103, 105, 110–112], the question remains whether their DNA methylation can provide a similar diagnostic resolution observed in LSG. However, biopsies of the LSG should be available from any diagnosis based on the formal diagnostic criteria [93].

To conclude, we show that DNA methylation profiles from LSG tissue alone can distinguish two SS disease subgroups with distinct clinical phenotype patterns and a genetic basis at the MHC. GSEA of hypomethylated genes implicated increased involvement of immune processes in severe cases compared to mild cases. Since the 2016 ACR/EUCLAR classification criteria does not distinguish between the two disease subgroups, this study provides a basis for potential revision to provide better diagnostic, disease management, and treatment subgroups and approaches.

Chapter 5

Bipartite graph-based approach for clustering of cell lines by gene expression-drug response associations

5.1 Introduction

One of the goals of precision cancer medicine is to identify predictive genomic features for drug response, which can be in a disease-specific or pan-cancer context [159]. Multiple pharmacogenomic datasets have been generated to this end, such as the NCI-60 drug sensitivity database [45], the Cancer Cell Line Encyclopedia (CCLE) [46], the Cancer Target Discovery and Development small molecule screening dataset [160], and the Genomics of Drug Sensitivity in Cancer dataset [47, 48]. A pharmacogenomic dataset typically comprises of drug response measurements for a panel of drugs and genomic measurements such as genotype data, copy number variation, and gene expression, for cell lines grouped by their tissue of origin.

Machine learning has become a popular tool for discovering drug-response associated (DRA) genomic features, or DRA biomarkers, and for predicting drug response in pharmacogenomic datasets [47, 161–166]. Aside from the classic approaches of elastic net and naïve Bayes, modern methods that focus on finding DRA biomarkers include multi-task learning of drug response [167] and the MERGE algorithm integrating multi-omic prior information [168]. Methods addressing the high-dimensional challenge imposed by genomic datasets include using drug-specific informative genes or pathway activity scores to model drug response [169, 170]. However, none of these approaches consider how the composition of the cell lines influences DRA biomarker discovery. In the absence of such consideration, these analyses are either performed only on cell lines known to originate from the same disease tissue, or on all available cell lines together. However, the disease-specific approach potentially sacrifices power by failing to include similar cell lines from other groups, and the pan-cancer approach reduces the disease-specificity of the discovered DRA biomarkers. The only approach we

are aware of to date, that explicitly considers the similarity between cell lines, models drug response as a weighted combination of responses from cell lines with correlated gene expression patterns and responses from drugs with correlated chemical structure features [171]. However, the approach by Zhang *et al.* is designed to predict drug response and does not explicitly identify DRA biomarkers, nor does it explicitly suggest which groups of cell lines are similar in terms of gene-drug associations.

We start by constructing a weighted undirected bipartite graph describing the associations between genes and drugs for each group of cell lines. The disjoint set of nodes represent the genes and drugs respectively, and gene-drug edge weights are assigned based on the direction and magnitude of association. The weights are derived from sparse canonical correlation analysis (SCCA), which solves for a linear combination of genes and drugs such that the Pearson correlation between the combination of genes and drugs is maximized [172]. SCCA shares the same advantage as the multi-task approach for drug response prediction in its ability to model the associations between multiple genes with multiple drugs simultaneously. We then introduce a nuclear norm-based dissimilarity measure to quantify the similarity between these graphs. Using this dissimilarity measure, we implement an agglomerative merging algorithm to successively combine cell line groups. Permutation-based p-values are generated to indicate the significance of each merge, to help differentiate inconsequential groupings from groupings of cell lines that share similar gene-drug association patterns.

We demonstrate our method on gene expression and drug sensitivity measurements from the CCLE dataset. We choose to work with gene expression because it has been shown to be the most predictive genomic data type for drug response compared to DNA methylation, cancer type, mutation, and copy number alteration [173]. First, we show that our method suggests significant merging between positive control groups expected to have similar gene-drug association patterns. Next, we show that our method ranks acute myeloid leukemia (AML) and chronic myeloid leukemia (CML) as the most similar groups compared to existing approaches that apply agglomerative hierarchical clustering. By combining AML and CML, SCCA was able to rank myeloid leukemia genes much higher compared to running SCCA on AML or CML alone. Finally, we demonstrate that our method was able to effectively infer the true hierarchy from simulation compared to existing agglomerative hierarchical clustering approaches.

5.2 Overview of proposed approach

In pharmacogenomic datasets such as CCLE, cell lines are grouped by their tissue of origin. Each group i with n_i cell lines has a genomic features data matrix $X^{[i]} \in \mathbb{R}^{n_i \times p}$ and a drug sensitivity data matrix $Y^{[i]} \in \mathbb{R}^{n_i \times d}$, all of which we can assume to be standardized. With G groups, the entire dataset is denoted $X = \{X^{[1]}, \ldots, X^{[G]}\}$ and $Y = \{Y^{[1]}, \ldots, Y^{[G]}\}$, and the goal is to identify groups that share similar gene-drug association patterns. Our approach begins by constructing a weighted bipartite graph to describe the associations between genes and drugs for each group. The edge weights of each graph are entries of the

outer product $B = a \otimes b$ between canonical vector $a \in \mathbb{R}^p$ for genes and canonical vector $b \in \mathbb{R}^d$ for drugs, which are solved for by SCCA. The vectors $a \in \mathbb{R}^p$ and $b \in \mathbb{R}^d$ specify sparse linear combinations of genomic features and drug responses to maximize Pearson correlation $Corr(Xa, Yb)$. The linear combinations Xa, Yb are sometimes referred to as canonical variates. We refer to the matrix $B \in \mathbb{R}^{p \times d}$ as the bipartite edge weight matrix (Figure 5.1B). We then introduce a nuclear norm-based dissimilarity measure to compare a given pair of bipartite edge weight matrices $B^{[u]}$ and $B^{[v]}$. With this dissimilarity measure, we can perform bottom-up merging of groups until all groups are merged. At each merge, a new graph is constructed using the merged cell lines. A hypothetical merging process involving datasets X, Y is depicted in Figure 5.1A.

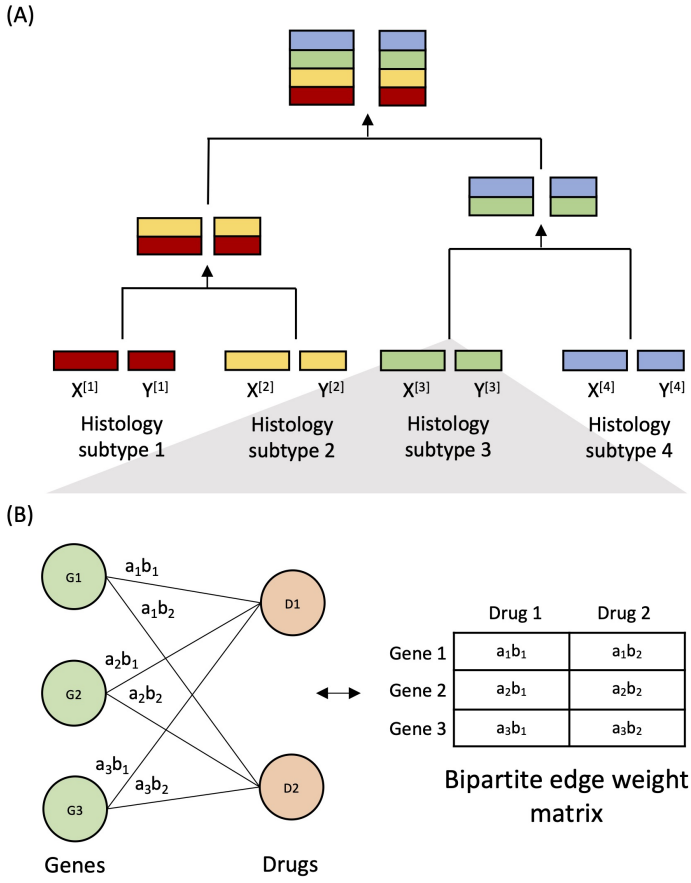


Figure 5.1: Overview of proposed approach. (A) Visualization hierarchical clustering applied to pharmacogenomic datasets X, Y. (B) For each cluster in the dendrogram, we represent gene-drug associations as a bipartite graph between genes and drugs, with edge weights between gene i and drug j as the product $a_i b_j$ of canonical vector entries. The canonical vectors are solved by SCCA, and the resulting bipartite edge weight matrices are compared using the nuclear norm-based dissimilarity measure.

5.3 Materials and methods

Review of sparse canonical correlation analysis

Motivated by the use of Pearson correlation to evaluate DRA biomarkers [174], we use SCCA to identify genomic features and drugs showing strong correlation. SCCA is a penalized extension of canonical correlation analysis (CCA) developed by Hotelling [175]. Since CCA is not scale invariant, assume each feature in X, Y is centered and scaled to variance one. In high throughput genomics data, p and sometimes d are typically much larger than n, and the subset of relevant biomarkers is often small. Hence, we impose sparsity on a, b by adopting the following diagonal penalized CCA criterion developed by Witten, Tibshirani, and Hastie [176], which treats sample covariance matrices $S_{XX} \in \mathbb{R}^{p\times p}$ and $S_{YY} \in \mathbb{R}^{d\times d}$ as diagonal and relaxes equality constraints for convexity

$$\begin{aligned} \max_{a\in\mathbb{R}^p, b\in\mathbb{R}^d} \; & a^\top X^\top Y b \\ & ||a||_2 \le 1, ||b||_2 \le 1 \\ & p_1(a) \le c_1, p_2(b) \le c_2, \end{aligned} \tag{5.1}$$

where p_1 and p_2 are convex penalty functions, and c_1 and c_2 are hyperparameters that control the degree of regularization. In our application, the ℓ_1 penalty is chosen as $p_1(\cdot) = p_2(\cdot) = ||\cdot||_1$ to induce sparse regularization [177], and c_1 and c_2 are selected based on k-fold cross validation. Thus, zero entries in a, b suggest that the corresponding genes and drugs are not associated with each other. Conversely, when the magnitudes of entries a_i, b_j for gene i and drug j respectively are large, then gene i and drug j are strong associated with each other. Genes with top canonical vector entries in absolute value are considered candidate DRA biomarkers. We adopt the modified NIPALS algorithm proposed by Lee *et al.* [172] to solve the above optimization, which is reported to have superior empirical performance than the algorithm proposed by Witten, Tibshirani, and Hastie [176].

Dissimilarity measure

We introduce a dissimilarity measure to compare bipartite edge weight matrices created from SCCA canonical vectors. Specifically, given canonical vectors $a \in \mathbb{R}^p, b \in \mathbb{R}^d$ solved by SCCA for a group of cell lines, we form the bipartite edge weight matrix $B \in \mathbb{R}^{p\times d}$ as the outer product $B = a \otimes b$ of the canonical vectors. Each entry $B[i, j] = a_i b_j$ contains information about the direction and magnitude of association between gene i and drug j, with negative values (e.g. $a_i > 0$ and $b_j < 0$) indicating negative association. We introduce a nuclear norm based dissimilarity measure between a pair of such edge weight matrices $B^{[u]}, B^{[v]}$ as

$$d(B^{[u]}, B^{[v]}) = \frac{\sum_i \sigma_i(B^{[u]} - B^{[v]})}{\sum_i \sigma_i(B^{[u]}) + \sum_i \sigma_i(B^{[v]})}, \tag{5.2}$$

where $d(\cdot,\cdot) \in [0,1]$ and $\sigma_i(A)$ denotes the i-th singular value of matrix A. The dissimilarity measure is designed based on the nuclear norm $||A||_* = \sum_{i=1}^{r} \sigma_i(A)$, which is the convex envelope of the rank function $Rank(A)$. Specifically, this means $||A||_*$ satisfies $Rank(A) \geq \frac{1}{M}||A||_*$ for all $A \in \{A \mid ||A|| \leq M\}$ [178]. If $B^{[u]}$ and $B^{[v]}$ are similar, then any meaningful matrix structure in $B^{[u]}$ and $B^{[v]}$ will become deficient in $B^{[u]} - B^{[v]}$, and the matrix difference may resemble a noise matrix. If we assume noise matrices tend to have small norm (e.g. Frobenius norm), then $||B^{[u]} - B^{[v]}||_*$ will tend to be small as well because $||A||_* \leq \sqrt{r}||A||_F$ holds for any $A \in \mathbb{R}^{m \times n}$ (see Appendix D.1 for proof). The denominator term in Equation (5.2) ensures $d(B^{[u]}, B^{[v]})$ is bounded between 0 and 1.

Agglomerative clustering of cell line groups

We implement an algorithm to perform bottom-up merging of the initial cell line groups, which is presented in Algorithm 5.1. In Algorithm 5.1, a new bipartite graph is constructed for each newly merged group and compared against existing bipartite graphs. The merging process continues until all groups are merged, yielding a dendrogram. See Appendix D.1 for how dendrogram height is calculated from merged dissimilarity measures.

Algorithm 5.1 Agglomerative hierarchical clustering

procedure HIERARCHICAL_CLUSTERING($(X^{[1]}, Y^{[1]}), \ldots, (X^{[G]}, Y^{[G]})$)
  Run `SCCA()` on $(X^{[i]}, Y^{[i]}) \Rightarrow B^{[i]}$ for $i = 1, \ldots, G$
  Construct $D \in \mathbb{R}^{G \times G}$ distance matrix
  while not all groups merged **do**
    Identify most similar groups $(X^{[1']}, Y^{[1']})$ and $(X^{[2']}, Y^{[2']})$
    Merge groups $\tilde{X} = \begin{bmatrix} X^{[1']} \\ X^{[2']} \end{bmatrix}$, $\tilde{Y} = \begin{bmatrix} Y^{[1']} \\ Y^{[2']} \end{bmatrix}$
    Run `SCCA()` on $(\tilde{X}, \tilde{Y}) \Rightarrow \tilde{B}$
    Update distance matrix
  Convert merged distances to dendrogram merge heights
  Output dendrogram

We also implement the option of subsampling cells to improve robustness of the resulting dendrogram. In this subsampling scheme, we replace running SCCA once to produce matrix B with running SCCA multiple times over repeatedly subsampled cell lines from a given group to produce the final element-wise average $\tilde{B}$. This subsampling procedure is summarized in Algorithm 5.2. This subsampling procedure is performed for every merged cluster, including the starting clusters.

Algorithm 5.2 Subsampling for robustness of hierarchical clustering.

procedure SUBSAMPLING((X, Y), m, p)
  Initialize $\tilde{B} = \mathbf{0} \in \mathbb{R}^{p \times q}$
  for i = 1 to m **do**
    Subsample p fraction of cells from (X, Y) to get $\hat{X}, \hat{Y}$
    $\hat{a}, \hat{b} = \texttt{SCCA}(\hat{X}, \hat{Y})$
    $\hat{B} = \hat{a} \otimes \hat{b}$
    $\tilde{B} := \tilde{B} + \hat{B}$
  $\tilde{B} := \frac{1}{m}\tilde{B}$
  Output $\tilde{B}$

We provide a statistical approach to quantify how similar the gene-drug associations are for each merge. Specifically, we compute a p-value for each merge by generating a null distribution of dissimilarities by permuting the ordering of cell lines between groups u and v. Our null and alternative hypotheses are

$$H_0 : \text{there is no shared gene-drug relationship between } u \text{ and } v. \tag{5.3}$$
$$H_1 : \text{there is shared gene-drug relationships between } u \text{ and } v. \tag{5.4}$$

This is implemented by keeping the rows of $Y^{[u]}$ and $Y^{[v]}$ fixed, and permuting the rows of $X^{[u]}$ and $X^{[v]}$ separately and independently of each other. Through permutation, any shared gene-drug relationships between groups u and v are broken. This procedure is summarized in Algorithm 5.3.

Algorithm 5.3 Permutation scheme for generating p-values.

procedure PERMUTATION($(X^{[u]}, Y^{[u]})$, $(X^{[v]}, Y^{[v]})$, n)
  Initialize empty $\texttt{D}[\cdot]$ of length n
  for i = 1 to n **do**
    Permute rows of $X^{[u]} \Rightarrow \tilde{X}^{[u]}$
    Permute rows of $X^{[v]} \Rightarrow \tilde{X}^{[v]}$
    $\tilde{a}^{[u]}, \tilde{b}^{[u]} = \texttt{SCCA}(\tilde{X}^{[u]}, Y^{[u]})$
    $\tilde{a}^{[v]}, \tilde{b}^{[v]} = \texttt{SCCA}(\tilde{X}^{[v]}, Y^{[v]})$
    $\tilde{B}^{[u]} = \tilde{a}^{[u]} \otimes \tilde{b}^{[u]}$
    $\tilde{B}^{[v]} = \tilde{a}^{[v]} \otimes \tilde{b}^{[v]}$
    $\texttt{D}[i] = d(\tilde{B}^{[u]}, \tilde{B}^{[v]})$
  Output $\texttt{D}[\cdot]$

With the generated null distribution of dissimilarities, the p-value is defined as the proportion of dissimilarity measures that is less than the observed dissimilarity. When H_0 is true, the p-value should be high, and when H_1 is true, the p-value should be small.

CCLE dataset

We apply our method to the mRNA expression and drug sensitivity datasets from CCLE [161], which contains pharmacologic profiling of 24 compounds across ~ 500 cell lines. Drug sensitivity is measured in terms of area over dose-response curve, as described by Barretina *et al.* [161]. Expression is measured in $\log_2 TPM$, where TPM stands for transcripts per million, a normalized unit of transcript expression. Cell lines and drugs were removed due to missing values to satisfy input requirement for $\texttt{SCCA}(\cdot)$. For computational efficiency, we followed the example in Barretina *et al.* [161] by selecting the most variable genomic features, reducing the number of genomic features from $\sim 50,000$ to ~ 800. The final dataset comprises of 391 cell lines grouped by 12 different primary sites, or tissue of origin, with 791 genomic features and 16 compounds (see Appendix D.1). Since many primary sites belong to different organ systems, we do not expect most merges to be biologically significant. Thus, we create positive control groups that are expected to be similar by randomly splitting the largest groups into two to test if they merge first. These groups are the lung tissue and hematopoietic and lymphoid tissue respectively. We also perform a separate analysis restricted to cell lines from the hematopoietic and lymphoid tissue, which has multiple tumor subtypes, to see if there are cell line groupings that lead to better DRA biomarker candidates.

Simulation

We compare the performance of our method against existing agglomerative merging approaches when a hierarchical relationship of gene-drug associations exists and is known, to see how well the resulting dendrograms resemble the true hierarchy. Additionally, we study how the merge p-values vary with the similarity between groups.

We simulate $X \in \mathbb{R}^{n \times p}$ and $Y \in \mathbb{R}^{n \times d}$ with $n = 700$, $p = 1,000$ and $d = 20$, comprising of seven groups of 100 cell lines each. We pre-select 200 genomic features as DRA biomarkers for each group, and generate drug response accordingly as a function of the selected genomic features. The simulation for a group of n_u cell lines is as follows:

1. Generate genomic covariance matrix $\Sigma^{[u]} \in \mathbb{R}^{p \times p}$ where only DRA biomarkers are highly correlated with each other. See Appendix D.1 for details on generating $\Sigma^{[u]}$.
2. Generate each cell line $X^{(i)} \sim \mathcal{N}(1, \Sigma^{[u]})$ for $i = 1, \ldots, n_u$.
3. Generate initial drug response as noise $Y_{ij} \sim \mathcal{N}(0, 0.1)$ for each cell line i and drug j.
4. Generate final drug response of cell line i to drug j as $Y_{ij} := Y_{ij} + (X^{(i)})^\top \beta^{[u]}$ where $\beta_k^{[u]}$ is drawn independently and uniformly from $\{-3, -2, 2, 3\}$ if k-th feature is a biomarker and $\beta_k^{[u]} = 0$ otherwise.

Each cell line group u is simulated from its own $\Sigma^{[u]}$ and $\beta^{[u]} \in \mathbb{R}^p$. The set of gene-dependent drug responses and the set of gene-independent drug responses is the same for each group. We set half of the drugs to be gene-dependent and the other half as gene-independent.

To simulate the hierarchy by varying the proportion of DRA biomarker entries in β that is shared between groups, where shared entries β_k have the same values. For instance, we can simulate two similar groups u and v by setting 90% of biomarker entries in $\beta^{[u]}$ and $\beta^{[v]}$ be identical. The hierarchical structure we simulate is presented in Figure 5.2.

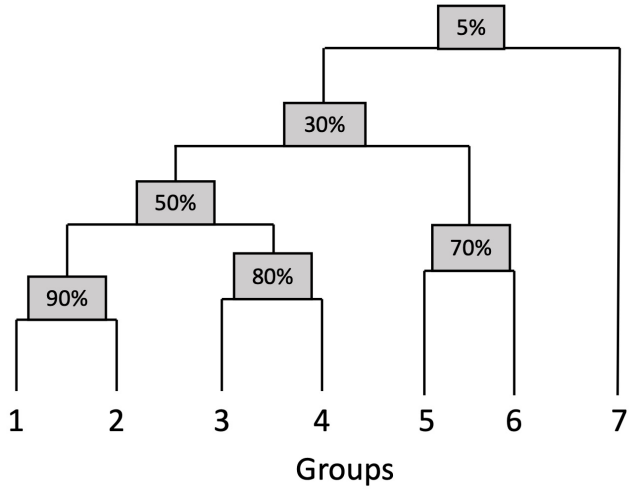


Figure 5.2: Simulated hierarchy with percentage at each non-leaf node indicating percentage biomarker overlap between any group from left child and any group from right child (height not necessarily proportion to dissimilarity).

5.4 Results

Agglomerative clustering identifies sensible clusters in CCLE data

On the CCLE dataset, our method merged groups expected to have similar gene-drug association patterns. We constructed positive control groups by randomly splitting cell lines from lung ($n = 88$) and hematopoietic and lymphoid tissue ($n = 68$) each into two groups, to see if these groups merge directly first with relatively low p-values. Figure 5.3 shows the dendrograms without and with the subsampling in Algorithm 5.2. From Figure 5.3A, we see cell lines from lung tissue and hematopoietic and lymphoid tissue were the first to merge with each other respectively, with both merges having the lowest p-values. Additionally, the p-value was lower for the lung control groups than that of the hematopoietic and lymphoid tissue control groups, which potentially reflects the greater heterogeneity present in hematopoietic and lymphoid tissue. Cell lines from the stomach and large intestine, which

are part of the digestive system, were also among the first groups to merge. However, the merge had a larger p-value of 0.48, potentially reflecting a larger difference between primary sites rather than within. To demonstrate stability of the merges, we ran the subsampling procedure presented in Algorithm 5.2, and the dendrogram remained unchanged (Figure 5.3B). The remaining p-values were closer to one and this supports the notion that most primary sites tend to be distant biologically.

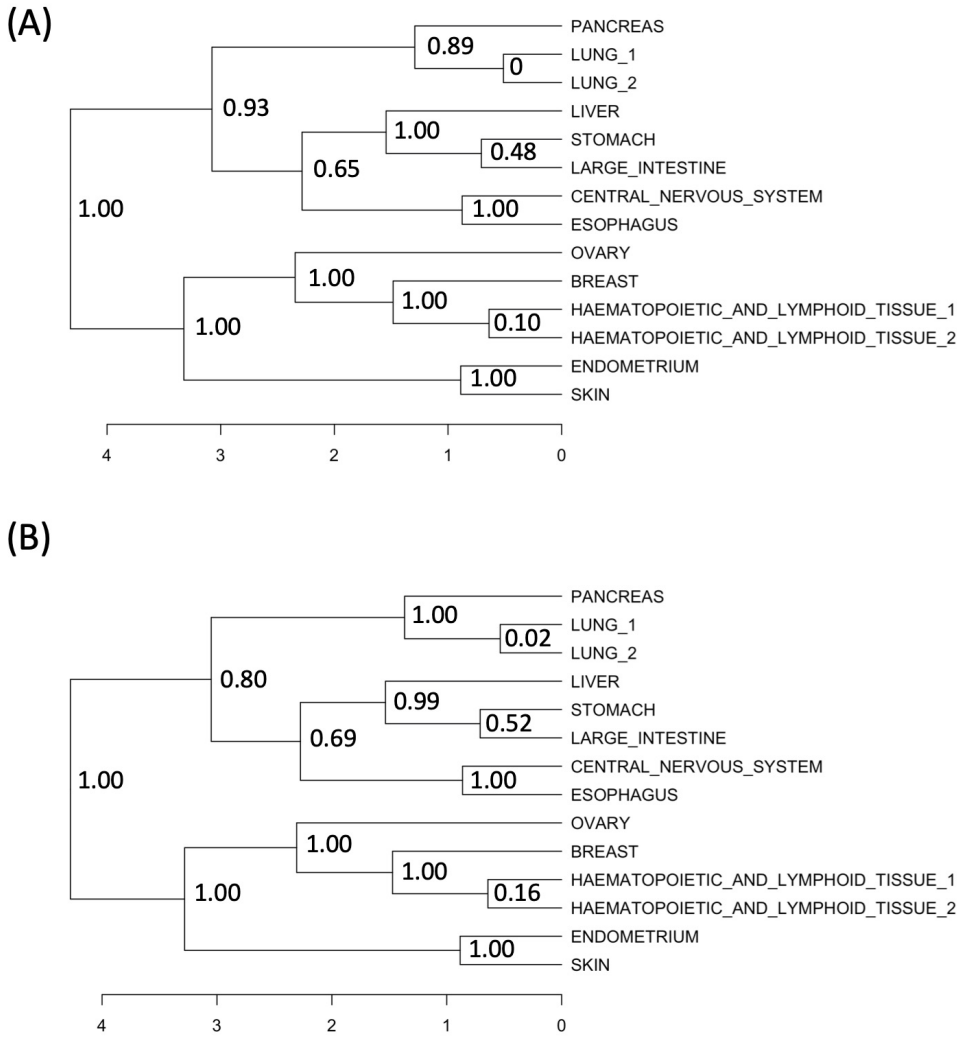


Figure 5.3: Hierarchical clustering dendrograms of CCLE dataset (A) without subsampling and (B) with subsampling ($n = 100, p = 80\%$). Permutation-based merge p-values are annotated to the right of each merge point, with 100 permutations.

Cell line composition influences discovery of candidate DRA biomarkers and drugs

We demonstrate that the appropriate grouping of tumor subtypes can lead to more disease-relevant DRA biomarkers in hematopoietic and lymphoid tissue. After filtering out subtypes with less than five cell lines, we ended with 60 cell lines, with five from chronic myeloid leukemia (CML), five from Burkitt lymphoma, six from T cell acute lymphoblastic leukemia (ALL), eight from diffuse large B cell lymphoma (DLBCL), nine from acute myeloid leukemia (AML), 13 from lymphoma unclassified (other), and 14 from multiple myeloma. Due to small sample sizes, we did not generate merge p-values.

We compared our clustering results against results from other hierarchical clustering-based approaches. The first baseline approach directly applied standard hierarchical clustering with Ward's minimum variance merging criterion to the combined dataset of mRNA expression and drug sensitivity $C = [X, Y] \in \mathbb{R}^{n\times(p+d)}$, with each feature centered and standardized. Hierarchical clustering was applied to group centroids for each starting group rather than individual cell lines. The second clustering approach is based on joint latent variables estimated from expression and drug sensitivity datasets by iCluster [179]. Latent variables $Z \in \mathbb{R}^{n\times k}$ were estimated by modeling gene expression and drug sensitivity with Gaussian distributions and setting $k = 6$ (one less than the number of starting groups, as recommended). Then, regular hierarchical clustering with Ward's minimum variance criterion was applied to the starting group centroids in the latent feature space. Both alternate approaches learn hierarchy from the joint feature space of genes and drug response instead of from the association patterns between genes and drugs.

Although the direct merge between AML and CML was observed in all approaches, this merge was prioritized as first in our approach (Figure 5.4). The similarity between AML and CML has a clear biological interpretation because myeloid leukemias originate from myeloid cells whereas the remaining disease subtypes develop from lymphoctyes. Myeloid cells reside in the bone marrow and develop into red blood cells, platelets, or white blood cells (except lymphocytes), whereas infection-fighting lymphocytes refer to B cells, T cells, or plasma cells.

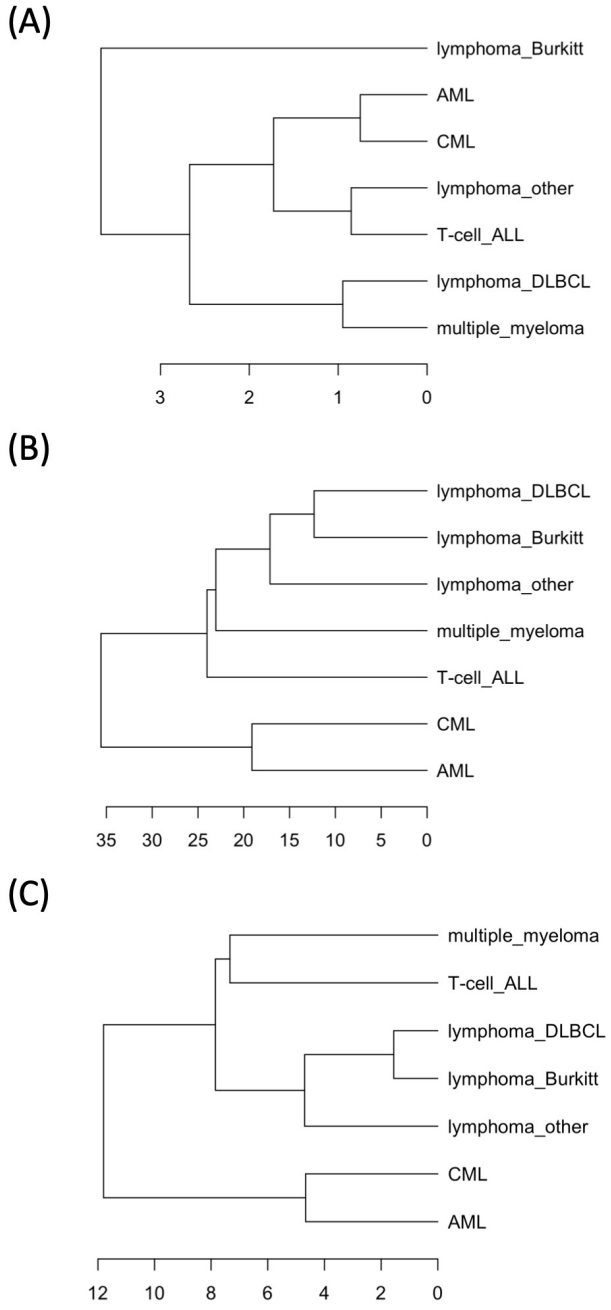


Figure 5.4: Comparison of agglomerative merging results of subtumors from hematopoietic and lymphoid tissue. (A) Bipartite graph-based clustering. P-values not computed due to small sample sizes. (B) Hierarchical clustering applied to starting group centroids of combined gene expression and drug sensitivity datasets. (C) Hierarchical clustering applied to starting group centroids of latent variables Z estimated from iCluster [179].

We found that both the candidate DRA biomarkers and associated drugs found by SCCA depend on cell line composition. To assess this, we ran SCCA separately on cell lines from AML alone, CML alone, AML combined with CML (AML+CML), and from all subtumors combined (hematopoietic and lymphoid tissue). In both AML and AML+CML, the compound Sorafenib emerged as the top compound. However, Sorafenib was ranked fifth in the hematopoietic and lymphoid group. The top five compounds in decreasing order for the AML+CML group were Sorafenib, TAE684, Topotecan, Nutlin-3, and AZD6244. In contrast, the top five compounds for the hematopoietic and lymphoid tissue group were ranked in decreasing order as PD-0325901, AZD6244, PF2341066, TAE684, and Sorafenib. Thus, the ordering of the compounds depended on the cell line composition.

We evaluated whether SCCA ranked myeloid leukemia-related genes higher in AML+CML compared to any other grouping of subtypes from hematopoietic and lymphoid tissue. Genes are ranked by coefficient magnitude in the canonical vectors. For comparison, we investigated how the top 10 genes for each group ranked in the other groups. Additionally, we labeled genes as myeloid leukemia related, other immune-related, or neither. Genes were labeled based on descriptions provided by the National Center for Biotechnology Information (NCBI) or published results from literature, and these are listed in Table 5.1.

Gene	**Label**	**Description**
FERMT3	myeloid leukemia	*FERMT3* (Kindlin-3) interacts with the ribosome and regulates *c-Myc* expression required for proliferation of CML cells [180].
CD70	myeloid leukemia	*CD70*/*CD27* signaling in AML cells activates stem cell gene expression programs, including the *Wnt* pathway, and promotes symmetric cell divisions and proliferation [181].
CD40	myeloid leukemia	*CD40* ligation reverses T cell tolerance in AML [182].
PPARG	myeloid leukemia	The *PPARG* receptor protein is expressed in primary myeloid and lymphoid leukemias and in lymphoma and myeloma cell lines. *PPARG* ligation alone and in combination with retinoids holds promise as novel therapy for leukemias by activating the transcriptional activity of target genes that control apoptosis and differentiation in leukemias [183].
YAP1	myeloid leukemia	In multiple myeloma (MM) and leukemias, *YAP* seems to exert a tumor suppressive function by regulating the *Abl1*-dependent DNA damage response, which leads to apoptosis in cancer cells. This explains why deletion or downregulation of *YAP*/*TAZ* are frequently observed in MM and leukemias [184].

PEG10	immune	Overexpression of this gene has been associated with several malignancies, such as hepatocellular carcinoma and B-cell lymphocytic leukemia.
LRMP	immune	The protein encoded by this gene is expressed in a developmentally regulated manner in lymphoid cell lines and tissues. The protein is localized to the cytoplasmic face of the endoplasmic reticulum.
CXCL8	immune	The protein encoded by this gene is a member of the CXC chemokine family and is a major mediator of the inflammatory response. The encoded protein is secreted primarily by neutrophils, where it serves as a chemotactic factor by guiding the neutrophils to the site of infection. This chemokine is also a potent angiogenic factor.
LYZ	immune	This gene encodes human lysozyme, which is one of the antimicrobial agents found in human milk, and is also present in spleen, lung, kidney, white blood cells, plasma, saliva, and tears.
SRGN	immune	This gene encodes a protein best known as a hematopoietic cell granule proteoglycan. Proteoglycans stored in the secretory granules of many hematopoietic cells also contain a protease-resistant peptide core, which may be important for neutralizing hydrolytic enzymes.
F3	immune	This gene encodes coagulation factor III which is a cell surface glycoprotein.
ETS1	immune	Copy number analysis on marginal zone B cell lymphomas of the gastrointestinal tract revealed amplification of the *ETS1* gene along with some flanking genes in the more aggressive large cell variants of these tumors. *ETS1* expression levels are a poor prognostic marker for diffuse large B cell lymphoma [185].

Table 5.1: Top DRA biomarkers with either myeloid leukemia or other immune-related functions. Each gene is ranked top 10 in either CML and AML combined, CML, AML, or all cell lines from hematopoietic and lymphoid tissue by SCCA. Unless cited, descriptions are provided by NCBI.

Figure 5.5 shows how the ranking of these genes changed depending on the grouping of cell lines. Although AML and CML are both myeloid leukemias, majority of myeloid leukemia genes were not highly ranked when SCCA was run separately on AML and CML cell lines respectively. By combining cell lines from AML and CML, most myeloid leukemia

genes surfaced among the top 10 ranked genes, demonstrating improved power. We observed that when we run SCCA on all 60 cell lines from hematopoietic and lymphoid tissue, the disease-specificity of DRA biomarkers decreased. Instead, the DRA biomarkers tended to have more general immune functions (Table 5.1).

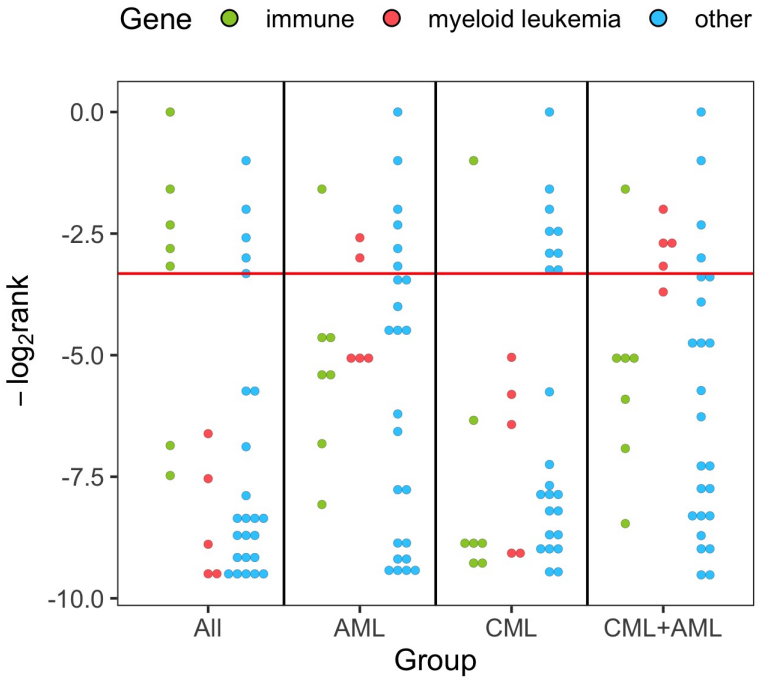


Figure 5.5: Dotplot of SCCA rankings of myeloid leukemia, immune, and genes with other functions by grouping of cell lines from hematopoietic and lymphoid tissue. The groups are hematopoietic and lymphoid tissue (All), AML, CML, and AML with CML (AML+CML). Ranking is transformed to negative log scale, so higher ranked genes are at the top. The red horizontal line corresponds to 10-th ranking.

Hierarchy inference in simulated data

We compared our approach against the other agglomerative clustering approaches introduced previously in inferring the hierarchy of clusters in the simulated dataset, where a hierarchy of gene-drug associations exists and is known (Figure 5.2). The resulting dendrograms are shown in Figure 5.6. Our bipartite graph-based approach was the only method to correctly infer the true hierarchy presented in Figure 5.2. Both alternate approaches fail to merge even the most similar groups. Using a p-value of 0.05 as a cutoff to choose clusters, we have groups 1, 2, 3, and 4 as one cluster, groups 5 and 6 as another cluster, and group 7 as its own cluster. According to ground truth, groups within each cluster have at least 50% of DRA

biomarkers in common. As desired, the p-values correlated negatively with the number of shared biomarkers, with $p = 0.13$ between groups sharing 30% biomarkers and $p = 0.72$ between groups that share only 5% of biomarkers.

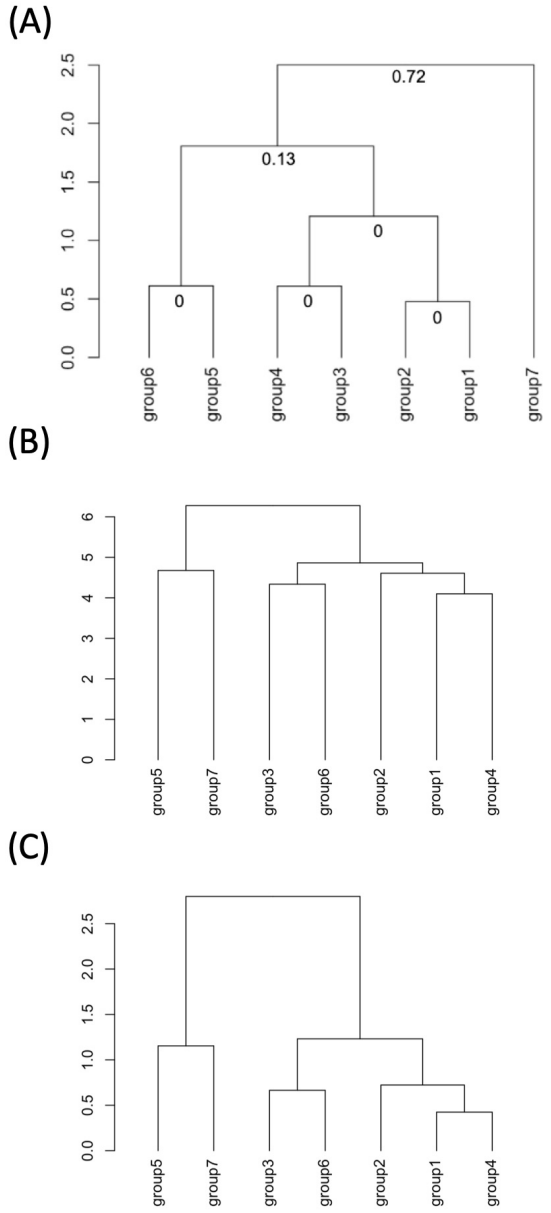


Figure 5.6: Comparison of hierarchical clustering approaches in simulated dataset. (A) Bipartite hierarchical clustering, with p-values after 100 permutations annotated below each merge point. (B) Baseline method of applying hierarchical clustering to starting group centroids of combined gene expression and drug sensitivity datasets. (C) Hierarchical clustering to starting group centroids of latent variables Z estimated from iCluster [179].

5.5 Discussion

Given the importance of tissue of origin for pharmacogenomic models of drug response [46, 161, 186], we developed a bipartite graph-based approach to describe gene-drug associations where agglomerative clustering can be applied to identify cell line groups with similar gene-drug associations. This could be helpful in large pharmacogenomic profilings of cell lines grouped by subtypes of a disease, where it is not obvious how the disease subtypes are related. To compare edge weight matrices, we introduce a dissimilarity measure based on the nuclear norm. To merge cell line groups successively, we apply agglomerative clustering using the nuclear norm-based dissimilarity measure. A subsampling procedure is introduced to improve dendrogram stability, and significant clusters can be selected with the help of p-values.

Applying our method to the CCLE dataset [161], we first illustrate that our method correctly merges cell lines from the same tissue of origin as positive control groups, with p-values that could be considered significant in the classical statistics setting. In hematopoietic and lymphoid tissue, our method identified AML and CML as the pair of disease subtypes with the most similar gene-drug association patterns. The rankings of candidate myeloid leukemia DRA biomarkers improved significantly when SCCA was run on CML and AML together as opposed to each group separately. When SCCA was run on hematopoietic and lymphoid tissue as a group, the rankings of disease-specific genes decreased, and the top associated drugs changed. Together this highlights the strong influence of cell line selection on DRA biomarker discovery. When the true hierarchical structure arising from gene-drug relationships is known from simulation, our method was the only one to uncover the true hierarchy completely. P-values appeared to grow with the degree with dissimilarity between merged groups. In our simulation, a p-value cutoff of 0.05 produced groups with at least 50% biomarkers in common. We remark though, that although p-value reflects the degree of shared gene-drug associations between two groups, the lack of statistical significance (e.g. p-value ≤ 0.05) does not necessarily imply lack of significant biological relatedness. In practice, we recommend using p-values along with prior biology knowledge and examination of candidate DRA biomarkers to decide on which groups to combine or not combine.

Bipartite agglomerative clustering can more generally be applied to identify clusters based on relationship between any two sets of variables, which in the pharmacogenomic setting could include genotype, DNA methylation, copy number variation, and other phenotype variations. In pharmacogenomic studies, our method could be used as the primary means of identifying groups of cell lines to select for further analysis, or itself be used to identify candidate DRA biomarkers for a subset of anti-cancer drugs in a subset of cancer cell lines.

The total runtime is mainly contributed by SCCA, the subsampling procedure described in Algorithm 5.2, and p-value generation described in Algorithm 5.3. Since the subsampling and permutation steps are independent processes, our implementation provides the option of parallelizing these steps using the `parallel` R package for improved runtime. We ran our clustering method with parallelization on the preprocessed CCLE dataset with 381 cell lines across 12 primary sites, 791 genomic features, and 16 compounds, using a MacBook

Pro with 2.4 GHz Quad-Core Intel Core i5 processor with 8 GB of RAM. P-values were computed ($n = 100$), and the runtimes ranged from less than 1 hour without subsampling to less than 2 hours with subsampling ($m = 100$). While the agglomerative merging algorithm we implemented merges until one group remains, more time-efficient implementations could consider early stopping of the merging process for some groups once the corresponding p-value exceeds a predefined threshold.

Multiple other approaches to the CCA problem exist which could be adapted to our method. To avoid prohibitively long runtimes from the SCCA implementation by Lee *et al.* [172], genomic features are restricted to a smaller subset of features on the order of thousands. Solari, Brown, and Bickel [187] recently proposed a two-step algorithm which first infers sparsity before solving for canonical vectors, an approach which reduces the search space to offer greater computational efficiency. Other CCA approaches serving various purposes include Bayesian CCA [188], deep neural network-based CCA [189], and kernel CCA [190], which could substitute the `SCCA`$(\cdot)$ procedure in Algorithm 5.1.

Chapter 6

HLA Allele Imputation with Deep Convolutional Neural Network

6.1 Introduction

The major histocompatibility complex (MHC) harbors the human leukocyte antigen (HLA) system on chromosome 6p21.3. HLA genes encode cell-surface proteins that present antigen peptides for recognition by T cells of the host immune system, and are thus among the most polymorphic genes in the human genome [191]. These genes are of strong epidemiological interest due to their large effect sizes in autoimmune diseases, infectious diseases, severe drug reactions, and transplant medicine [192–195].

Direct typing of HLA alleles include sequence specific oligonucleotide hybridization, capillary sequencing, and next-generation sequencing, but these approaches are labor-intensive, time-consuming, and expensive [196]. Thus, multiple approaches to impute HLA alleles from single nucleotide polymorphism (SNP) data were developed. These methods include HLA Genotype Imputation with Attribute Bagging (HIBAG), HLA*IMP:02, and SNP2HLA [71, 197, 198]. A comparison of HLA imputation programs concluded that HIBAG and SNP2HLA have higher concordance rates than HLA*IMP:02 in European Americans and African Americans [199]. However, HIBAG performs imputation for each locus independently and thus cannot impute HLA haplotypes, which may limit its applicability for haplotype association studies [200, 201]. Since both HIBAG and SNP2HLA combine the phasing and training stages, usage may incur more time than is necessary when the SNP genotype data are already phased. SNP2HLA combines all stages of phasing, training, and imputing into one stage so that entire process has to be repeated for each new SNP genotype dataset to impute.

We present a deep learning approach to HLA imputation using convolutional neural networks (CNN). Deep learning is characterized by its high model capacity to fit arbitrarily complex functions [202]. CNNs have already been successfully applied to a variety of genetic sequence modeling problems, such as DNA-protein binding or chromatin accessibility prediction [203, 204]. CNNs are effective at modeling genetic sequences by learning to de-

tect motifs with its convolutional filters, each of which can be thought of as learning some position weight matrix for a motif [205]. Our multiple input, multiple output CNN accepts phased haplotypes ±250kb flanking each HLA locus and outputs a probability distribution over alleles for each locus. In other words, it maps a haplotype of genotypes to a haplotype of HLA alleles. In this manner, the CNN is able to learn from long range linkage disequilibrium patterns across HLA loci. The embedding and convolutional layers are shared across loci for extraction of higher-order features from flanking genotypes, and then CNN branches off as separate fully-connected layers for each locus. We train and test our model on individuals of European ancestry from the Type 1 Diabetes Genetics Consortium (T1DGC) [206]. We report that our CNN has improved imputation accuracy over SNP2HLA while having comparable performance with HIBAG, but can take considerably less time when the data has already been phased.

6.2 Materials and Methods

The T1DGC dataset comprises of 5,225 unrelated individuals of European ancestry, with phased genotype and HLA allele data. We start with 5,698 SNPs in the genotype data at the MHC region assayed with the Illumina 550K platform and extract SNPs flanking ±250kb from each HLA locus as predictive SNPs. HLA alleles were typed for *HLA-A*, *-B*, *-C*, *-DQA1*, *-DQB1*, *-DPA1*, *-DPB1* and *−DRB1* at four-digit resolution, totaling 296 distinct alleles. A total of 109 individuals were removed for not having information for both alleles per HLA locus, resulting in 5,116 individuals. Details of this dataset are described elsewhere [71, 199, 206].

We tokenize, or split, a haplotype sequence of SNPs corresponding to each HLA locus into k-mers, with $k = 5$. For example, tokenization of the haplotype sequence AGTCGATAGCAT with $k = 5$ is the process AGTCGATAGCAT $\rightarrow$ [AGTCG, ATAGC], with the remaining SNPs that cannot form a complete k-mer at the right end omitted. Let $x_l^{(i)} \in \mathbb{R}^{n_l}$ denote the sequence of n_l k-mers corresponding to HLA locus l from haplotype i. Then haplotype i across all eight HLA loci is denoted

$$x^{(i)} = [x_1^{(i)}, x_2^{(i)}, \ldots, x_8^{(i)}], \tag{6.1}$$

with corresponding HLA alleles $y^{(i)} \in \mathbb{R}^8$ at the four-digit resolution. With m individuals, the dimensions of the dataset are $X \in \mathbb{R}^{2m \times (n_1 + \cdots + n_8)}$ and $Y \in \mathbb{R}^{2m \times 8}$, with each individual contributing two haplotypes per HLA locus. The goal of HLA imputation is to find f for the mapping $f : X \rightarrow Y$.

Data pre-processing

To encode the input haplotypes, one-hot encoding is applied to n_{kmer} distinct k-mers present in the training dataset, with the zero vector serving as a placeholder for unobserved k-mers. One-hot encoding involves assigning a 1-to-1 mapping between each k-mer to a vector of

length n_{kmer}, with one at the index corresponding to the k-mer and zero elsewhere. For instance, when $k = 1$, there are 4 possible k-mers (i.e. A, T, C, G), and the one-hot encoding for the third k-mer is $\begin{bmatrix}0 & 0 & 1 & 0\end{bmatrix}$.

Since the embedding and convolutional layers are shared between HLA loci, each haplotype at a HLA locus $x_l^{(i)}$ is post-appended with zero vectors to the maximum sequence length $n_{\max} = \max_l n_l$, such that each HLA locus has the same input feature length of $n_{\max}$. Thus, the one-hot encoding of a haplotype at HLA locus l has dimensions $x_l^{(i)} \in \mathbb{R}^{n_{\max} \times n_{kmer}}$, where $n_{\max}$ is the maximum number of k-mers across HLA loci and n_{kmer} is the k-mer encoding length. Each haplotype i has multiple inputs $x_1^{(i)}, \ldots, x_8^{(i)}$ into the CNN, which imputes HLA alleles $y_1^{(i)}, \ldots, y_8^{(i)}$.

Network architecture

The CNN architecture is organized into an embedding layer followed by two convolutional layers that is shared between HLA loci. After the convolutional layers, the CNN architecture branches out into separate fully-connected layers for each HLA locus. Overview of the multiple input, multiple output CNN architecture is presented in Figure 6.1 and details are outlined below

1. Embedding layer of dimension $d = 8$.
2. 1D convolution with 64 filters with window size $h = 4$, stride step size $s = 1$, ReLU, 1D max-pool over window of size 4.
3. Batch normalization with momentum 0.8, 1D convolution with 64 filters with window size $h = 8$, stride step size $s = 1$, ReLU, 1D max-pool over window of size 4.
4. Flatten activation feature maps, batch normalization with momentum 0.8, and dropout with probability $p = 0.5$.
5. Concatenate activation feature maps between neighboring loci.
6. Fully-connected layer with 32 units, ReLU, and dropout with probability $p = 0.5$.
7. Softmax output layer.

The first layer of the CNN is an embedding layer for learning a representation of each k-mer such that k-mers associated with same HLA allele have similar representation. Each k-mer is represented by a real-valued vector of dimensions that is typically much less than that for one-hot encoded vectors, which have dimension equal to the number of possible k-mers. The purpose of the embedding layer is to transform sparse high-dimensional input vectors into dense low-dimensional vectors that may encode similarities between k-mers for improved learning [207]. This concept has been applied in natural language processing tasks [208, 209] as well as prediction task for genetic sequences [204]. The embedding layer is a

learnable matrix of dimension $(n_{kmer}+1) \times d$, where d is the size of the embedding dimension and $n_{kmer}+1$ is the number of distinct k-mers plus one for the placeholder zero vector. Each row $r_i \in \mathbb{R}^d$ of the embedding matrix can be thought of as the learned representation for the i-th k-mer.

In our CNN architecture, the embedding layer is followed by two convolutional layers. Each convolutional layer comprises at least the following sequential operations.

1. 1D convolution
2. ReLU nonlinearity
3. Max pooling

The second convolutional layer is preceded by batch normalization. Batch normalization involves normalizing to each input feature independently to have mean zero and variance one, where mean and variance estimates are estimated from mini-batches of data used for stochastic gradient training. To improve representation power of the network, batch normalization also introduces a pair of learnable parameters for scaling and shifting each feature following normalization. The main purpose of batch normalization is to lead to faster training that is less sensitive to parameter initialization. Additionally, batch normalization is also known to have a slight regularization effect [210].

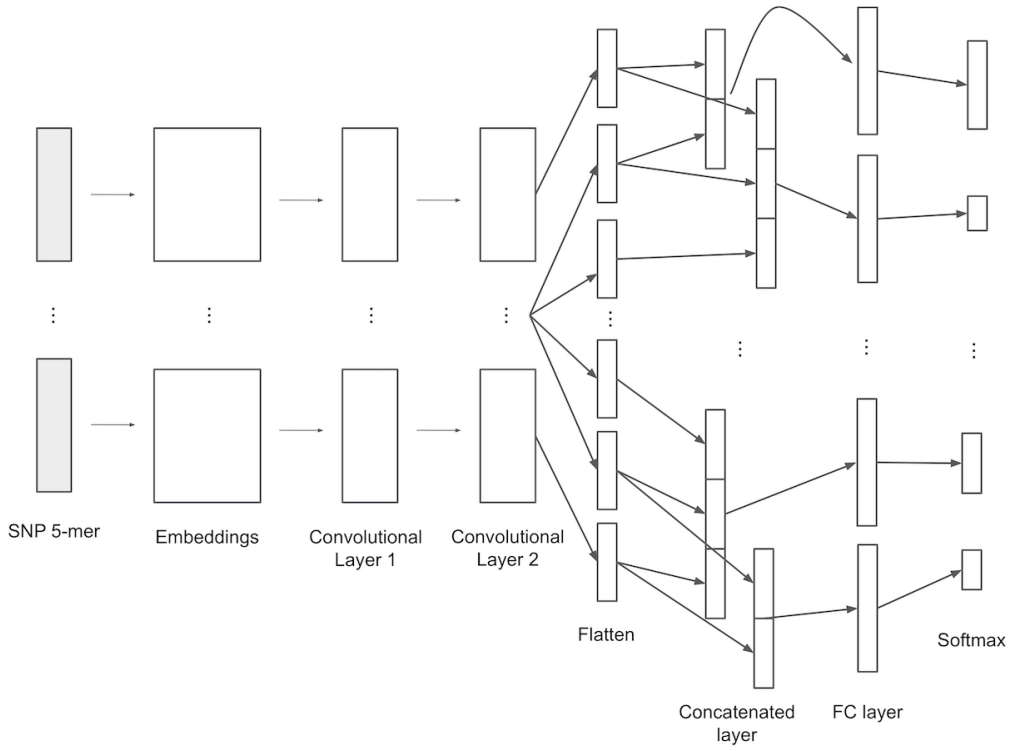


Figure 6.1: Graphical illustration of CNN architecture for HLA imputation. The k-mer sequences are first passed through the shared embedding and convolutional layers to compute intermediate activation feature maps. Activations from neighboring loci are jointly used for imputation by a full-connected network corresponding to each HLA locus. FC = fully-connected.

For haplotype i at HLA locus l, omit i, l for clarity and let $x_j \in \mathbb{R}^d$ be the j-th k-mer vector for haplotype i. The 1D convolution operation involves applying a trainable filter $w \in \mathbb{R}^{h \times d}$ to a window of h k-mer vectors to produce a new feature $z_j = \langle w, x_{j:j+h-1} \rangle_F + b$. Here $b \in \mathbb{R}$ is a learnable bias term and $\langle A, B \rangle_F$ denotes the Frobenius inner product between matrices A, B. The filter w is applied against the next window of h k-mers in a sliding window fashion with a stride step size s. Thus, for haplotype $x = [x_{1:h}, x_{2:h+1}, \ldots, x_{n_{\max}-h+1:n}]$, 1D convolution outputs the intermediate feature vector

$$z = [z_1, z_2, \ldots, z_{n_{\max}-h+1}] \tag{6.2}$$

with $z \in \mathbb{R}^{n_{\max}-h+1}$. The non-linear ReLU activation function $f : \mathbb{R} \to \mathbb{R}$ is defined as

$$f(z) = \max(0, z) \tag{6.3}$$

and is applied to each z_i to produce an activation feature a_i. Applying 1D convolution on a window of h k-mers followed by ReLU nonlinearity can be succinctly expressed as

$$a_i = f(\langle w, x_{i:i+h-1} \rangle_F + b) \tag{6.4}$$

which for $n_{\max}$ input k-mers produces the activation feature vector

$$a = [a_1, a_2, \ldots, a_{n_{\max}-h+1}] \tag{6.5}$$

with $a \in \mathbb{R}^{n_{\max}-h+1}$. The ReLU activation function is chosen over other non-linearities such as $f(x) = \tanh(x)$ or $f(x) = 1/(1+e^{-x})$ because it offers faster training time due to ReLU's non-saturating property [211]. Finally, max pooling is applied by retaining the maximum activation value over non-overlapping windows of size p, reducing the dimension of a by p and retaining the most important activation feature (i.e. one with largest value in window h). Refer to Kim [209] for additional details on the 1D convolution layer.

During training, a filter w can be thought of as learning to detect a particular motif across $n_{\max}$ k-mers. A high activation value a_i from filter w is indicative of the presence of a motif in window i. A convolutional layer learning to detect f different motifs would involve training f different filters. With f filters, the first convolutional layer would have input and output dimensions $x \in \mathbb{R}^{n_{\max} \times d} \rightarrow a \in \mathbb{R}^{\lfloor \frac{n_{\max}-h+1}{p} \rfloor \times f}$. We call this output the activation feature map. In a deep neural network, multiple such convolutional layers can be stacked, with the output activation feature map from the previous layer serving as input to the next layer.

The final portion of our CNN architecture is a fully-connected (FC) layer for final prediction. Following convolution, activation feature maps corresponding each HLA loci $a_1, \ldots a_8$ are first flattened, then neighboring activations are concatenated according to

$$a'_i := \begin{cases} [a_{i-1}, a_i, a_{i+1}] & i \in [2, 7] \\ [a_i, a_{i+1}] & i = 1 \\ [a_{i-1}, a_i] & i = 8 \end{cases} \tag{6.6}$$

so that our CNN can learn from long-range disequilibrium patterns between neighboring loci. For example, since *HLA-DRB1*15:01* and *HLA-DQB1*06:02* are strongly linked in European populations, presence of *HLA-DRB1*15:01* is indicative of the presence of *HLA-DQB1*06:02* [212]. The output layer for each HLA locus is a softmax layer that outputs a probability distribution over alleles. For example, for L possible alleles at a HLA locus, the softmax function for probability of allele j given input feature activations $a \in \mathbb{R}^L$ is

$$P(y = j \mid a) = \frac{e^{a_j}}{\sum_{l=1}^{L} e^{a_l}}. \tag{6.7}$$

Let $q_j^{(i)}$ be the predicted probability of the true allele at locus j of the i-th haplotype, then the target loss function we optimize over is the categorical cross entropy

$$\mathcal{L}(q) = -\frac{1}{8m} \sum_{j=1}^{8} \sum_{i=1}^{m} \log q_j^{(i)}, \tag{6.8}$$

where m is number of haplotypes and $q = \{q^{(1)}, \ldots, q^{(m)}\}$ with $q^{(i)} \in \mathbb{R}^8$.

Regularization

Dropout is applied to the respective activation feature maps feeding into the FC and softmax layers to prevent overfitting[213]. Dropout involves retaining activation values with some fixed probability p and zero otherwise, independent of other activation values, with each forward pass during training. With vector inputs, dropout is implemented with a binary mask vector $u \in \mathbb{R}^d$ that is multiplied element-wise with the input

$$z = W \cdot (a \circ u) + b, \tag{6.9}$$

where $\circ$ denotes element-wise multiplication, $W \in \mathbb{R}^{k \times d}$ is the trainable weight matrix of a layer, and $b \in \mathbb{R}$ is a trainable bias term. Applying dropout during training can be interpreted as updating weight parameters of a sampled neural networks within the full neural network. By approximating the process of combining exponentially many different neural network architectures, dropout prevents overfitting [213]. At test time, each activation value is multiplied by dropout probability $a := pa$ so that the activation value has the same expected output in test and training time.

Hyperparameter optimization and training

We partition 70% of individuals for training and model development, and the remaining individuals for final evaluation. We perform random search over 100 random samples from the hyperparameters embedding dimension, batch size, number of convolutional filters, filter stride step size, number of hidden units for the FC layer, max-pooling window, and dropout probability. Random search is more computationally efficient than exhaustive grid search and can outperform grid search when only a small number of hyperparameters affect final model performance [214]. The set of hyperparameters is optimized over 20% of the training dataset set aside as the validation dataset.

We choose the Adam optimizer for stochastic optimization [215] with a learning rate 0.001 and batch size of 512 haplotypes. During training we apply early stopping when the loss over the development dataset ($10\% - 15\%$ of the training dataset) does not improve for two epochs, or two passes over the entire training dataset.

6.3 Results and discussion

HLA imputation performance

HLA imputation accuracy on the test dataset was comparable to the state of the art performance achieved with SNP2HLA and HIBAG, which are the HLA imputation programs publicly available to date. Figure 6.2 and Table 6.1 summarize the imputation accuracies

by imputation program and HLA locus. The overall accuracies for CNN, HIBAG, and SNP2HLA were 97.6%, 97.5%, and 95.8% respectively. Thus, CNN and HIBAG had the best and most comparable performance. Either CNN or HIBAG had superior accuracy over SNP2HLA for every HLA loci. HIBAG out-performed CNN for *HLA-A*, *HLA-B*, *HLA-DQA1*, and *HLA-DQB1*. CNN out-performed HIBAG for *HLA-C*, *HLA-DPA1*, *HLA-DPB1*, and *HLA-DRB1*.

Table 6.1: Comparison of test imputation accuracy by HLA locus between HLA imputation methods.

	CNN	**SNP2HLA**	**HIBAG**
HLA-A	0.971	0.952	**0.978**
HLA-B	0.949	0.915	**0.958**
HLA-C	**0.993**	0.992	0.992
HLA-DPA1	**0.993**	0.973	0.989
HLA-DPB1	**0.960**	0.935	0.955
HLA-DQA1	0.996	0.993	**0.997**
HLA-DQB1	0.992	0.986	**0.993**
HLA-DRB1	**0.951**	0.919	0.938
Overall	**0.976**	0.958	0.975

HIBAG out-performed CNN the most at locus *HLA-B* (by about 1%), which is the most polymorphic gene in our training dataset with 96 alleles. It is possible that HIBAG is effective for polymorphic genes due to its design as an ensemble classifier that employs both bootstrap aggregation on individuals and feature bagging on SNPs. However, this does not necessarily imply that deep learning is less effective for imputation of high polymorphic genes, since the size of the T1DGC dataset is relatively small compared that in other deep learning applications.

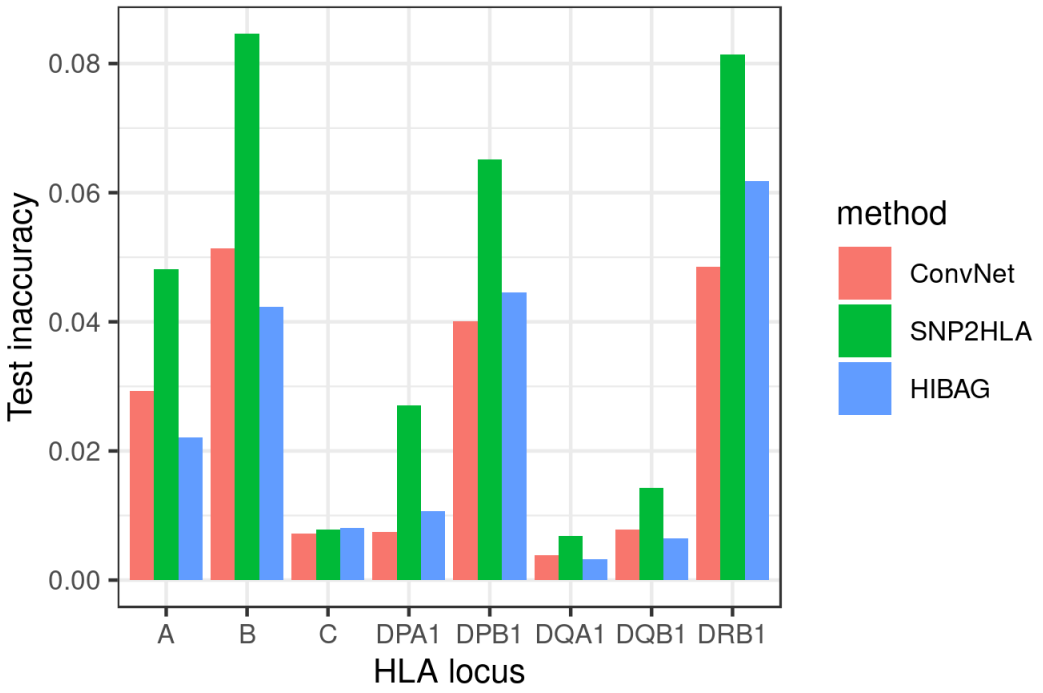


Figure 6.2: Comparison of test imputation performance by HLA locus between HLA imputation methods. Performance is reported with inaccuracy.

Our CNN presents an advantage in training time over that required by SNP2HLA and HIBAG when the genotypes are already phased. For HIBAG, 10 classifiers are trained in parallel over 7 CPU cores for each HLA locus. All training was performed by a Linux server with four 10-core Intel(R) Xeon(R) CPU E7-4860 processors (running at 2.8 GHz, with 640KiB/2560KiB/24MiB L1/L2/L3 cache, and using a 64-bit architecture) and a total of 128 GB RAM. The runtimes are presented in Table 6.2. Both SNP2HLA and HIBAG perform phasing regardless of whether the input is phased or not, which account for the longer times required for imputation. The runtime for HIBAG to impute all eight HLA loci is the longest because imputation is performed separately and independently for each locus. As a consequence, a limitation of HIBAG is its inability to impute HLA haplotypes. It should be noted however, that the runtime of HIBAG could be reduced if imputation for each HLA locus could be performed in parallel. Compared to CNN and HIBAG, which separates the training and testing procedures, the program SNP2HLA requires repeating the entire imputation procedure with the reference panel for each test imputation, which could amount to additional runtime in practice.

Table 6.2: Comparison of imputation program runtimes. HIBAG was trained with 10 classifiers over 7 CPU cores per HLA locus and runtime is summed over HLA loci. SNP2HLA runtime includes test imputation since it combines training and imputation into one procedure. *Since the runtimes of SNP2HLA and HIBAG include phasing, we also report the runtime for CNN that includes phasing performed by BEAGLE in brackets [9].

Imputation program	**Training time (hours)**
CNN	0.25(9.6)*
SNP2HLA	10.5
HIBAG	32.3

We computed approximate error bars for CNN test accuracy per locus to get an estimate of the variation in imputation accuracy due to variation in test allele frequencies. This is accomplished by creating $B = 1,000$ bootstrap samples (with replacement) from the original test dataset and computing the test accuracy from the trained CNN for each bootstrap sample. The distribution of test accuracies are shown in Figure 6.3.

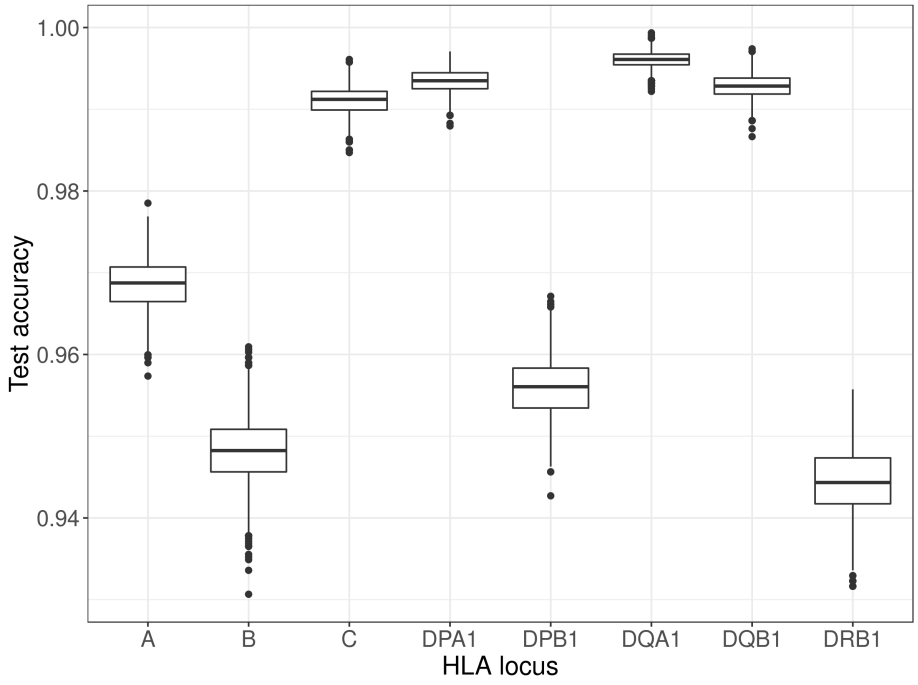


Figure 6.3: Boxplot of bootstrap test accuracies by HLA locus. The upper whisker extends from the upper hinge (3rd quartile) to the largest value no further than 1.5 the inter-quartile range (IQR) from the upper hinge. The lower whisker extends from the lower hinge (1st quartile) to the smallest value at most 1.5 the IQR of the lower hinge.

The variability in per locus accuracy is correlated with the degree of corresponding polymorphism. The number of HLA alleles in the training dataset is 96 for *HLA-B*, 51 for *HLA-DRB1*, 50 for *HLA-A*, 33 for *HLA-DPB1*, 31 for *HLA-C*, 18 for *HLA-DQB1*, 8 for *HLA-DQA1*, and 7 for *HLA-DPA1*.

Occlusion Analysis

One way to assess what SNPs the CNN might be using to learn the mapping between phased genotypes to HLA alleles is to perform occlusion analysis, which involves independently "masking", or occluding, a section of SNPs and observing how the probability of the true HLA allele from the trained model subsequently decreases. The larger the decrease in probability, the more important the "masked" SNPs are for imputation. Specifically for a given HLA locus, we successively set 30 k-mers corresponding to the locus to zero at a time, which removes information from the selected 30 k-mers.

Given we know the tag SNPs for many HLA alleles [216], occlusion analysis can help us determine to what extent the CNN is using the tag SNPs to impute the correct allele. For this analysis we selected the HLA alleles *HLA-DRB1*15:01*, *HLA-DRB1*04:01*, *HLA-DQB1*03:02*, *HLA-DQA1*01:02*, *HLA-B*08:01*, and *HLA-C*16:01*, many for their association with autoimmune diseases [217]. The allele *HLA-DRB1*15:01* is associated with multiple sclerosis, *HLA-DRB1*04:01* is associated with rheumatoid arthritis, *HLA-DQB1*03:02* is associated with celiac disease, *HLA-DQA1*01:02* is associated with systemic lupus erythematosus, and *HLA-B*08:01* is associated with plasma beta-2 microglobulin [217, 218]. Figure 6.4 shows the resulting probability heatmap from occlusion analysis. There is evidence that the CNN learned to impute the alleles *HLA-B*08:01*, *HLA-DQB1*03:02*, *HLA-DQA1*01:02*, and *HLA-DRB1*15:01* based more so on tag SNPs as opposed to other SNPs, since the drops in probability were largest around the respective tag SNPs. This is especially true for *HLA-DRB1*15:01*, where the probability dropped by nearly half when SNP k-mers in the neighborhood of tag SNP rs3135388 were "masked".

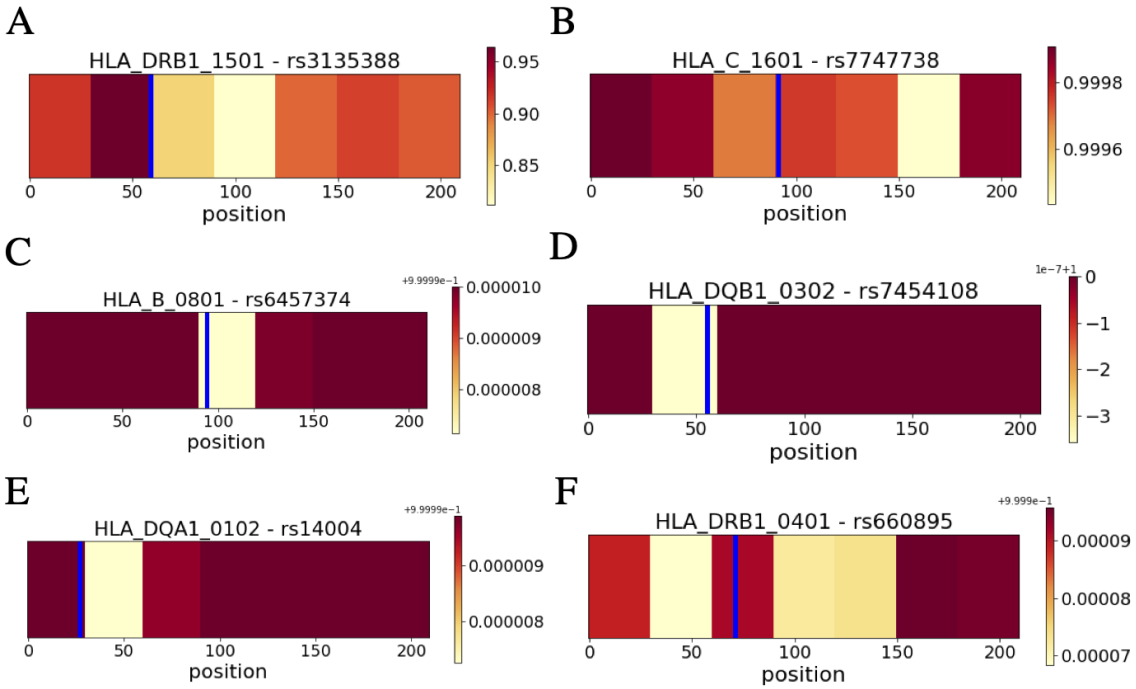


Figure 6.4: Occlusion sensitivity analysis for HLA alleles (A) *HLA-DRB1*15:01*, (B) *HLA-DRB1*04:01*, (C) *HLA-DQB1*03:02*, (D) *HLA-DQA1*01:02*, (E) *HLA-B*08:01*, and (F) *HLA-C*16:01*. Each panel is a heatmap of true allele probability after blocks of 30 k-mers were "masked" to zero for a HLA allele, with the rs number of tag SNPs in the title. The position of the tag SNP is marked with a blue vertical line.

Except for *HLA-DRB1*15:01*, the drops in probability were mostly minuscule, which shows that the CNN learned the imputation mapping based also on genetic information other than what is provided by the tag SNP neighborhood. This suggests that imputation performance by the CNN is robust against genotying errors. For alleles *HLA-C*16:01* and *HLA-DRB1*04:01*, "masking" the neighborhood of tag SNPs did not result in the largest decrease in probability. Together, this demonstrates that the CNN learned to impute based on known associations between SNPs and HLA alleles, but that this was not necessarily the case for all HLA alleles.

Sensitivity analysis

We perform sensitivity analysis to check the robustness of our CNN architecture against some hyperparameters, since the hyperparameter tuning process itself is noisy due to random weight initialization, stochastic optimization, etc. We chose to focus on the hyperparameters k-mer length k, embedding dimension d, and stride step size s for all convolutional filters f.

According to Figure 6.5, performance of the CNN is not sensitive to the choices of embedding dimension d and k-mer length k. We tested the embedding dimensions $d = 5, 10, 15, 20$, which controls the model complexity due to the embedding layer. We varied the k-mer length for $k = 3, 4, 5, 6, 7$, which controls the input dimension to the CNN.

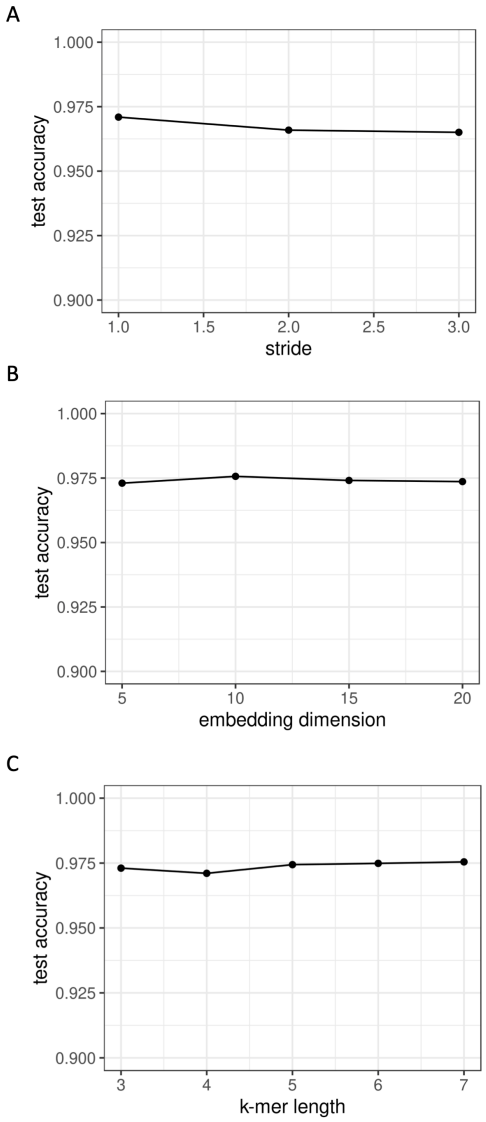


Figure 6.5: Sensitivity analysis of hyperparameters (A) embedding dimension d, (B) convolutional filter f stride step size s, and (C) k-mer length k, performed on the test dataset.

However, performance decreased slightly with stride step size s, where step sizes $s = 1, 2, 3$ were tested. This could be due to the fact that a larger step size s reduces the dimension of

the activation feature maps more aggressively between layers of the CNN and discards more information. Thus, the recommended step size for the CNN is $s = 1$.

6.4 Discussion

We introduce a 1D CNN that simultaneously imputes alleles at the four-digit resolution for HLA loci *HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DPB1* and *-DRB1* from phased genotype data flanking each locus. Genotype data corresponding to each HLA locus are first tokenized into units of k-mers and one-hot encoded before serving as input to the CNN. The CNN architecture starts with an embedding layer that learns dense low-dimensional vectors to represent k-mers from high dimensional one-hot encodings. Following the embedding layer are two convolutional layers that learn to detect genotype motifs for HLA imputation. The activation feature maps learned for each locus are concatenated with feature maps of neighboring loci for final allele imputation using fully-connected networks. The concatenation of neighboring activations allows genotype information corresponding to a HLA locus to influence imputation of neighboring loci when there exists long-range disequilibrium between neighboring loci.

Our CNN shares the ability of SNP2HLA to impute HLA allele haplotypes while having high imputation accuracy comparable to that from HIBAG. Additionally, separation of the stages of phasing, training, and imputation reduces unnecessary runtime that results from combining these stages into one. We demonstrate via occlusion sensitivity analysis that CNN can indeed learn to impute HLA alleles based on known associations between tag SNPs and HLA alleles, but also that the learned mapping between genotype and HLA allele is dependent on more than these known associations. The architecture we propose is relatively robust against the hyperparameters embedding dimension d, stride step size s, and k-mer length k. We anticipate that the performance of deep learning for HLA imputation will improve as more SNP and HLA datasets become available.

Although we demonstrate the effectiveness of CNN for imputation of HLA alleles at the four digit resolution, our CNN can in principle also impute amino acid polymorphisms in HLA proteins from SNP genotype data. A extension of our work is to explore how effective deep learning could be for HLA imputation in admixed populations, which are populations with more than one ancestral populations. A study found that HIBAG was able to achieve reasonable HLA imputation accuracy in admixed Brazilian population using the 1000 Genomes HLA and SNP dataset [73]. However, a decrease in HLA imputation accuracy is generally observed in admixed populations compared to that in European populations [199]. Part of this is due to a partial mismatch between the ancestries present in the reference population and the population to be imputed, but another could be due to insufficient conditioning on population substructure and differing linkage disequilibrium patterns at the MHC between populations. It is known that many HLA alleles are population-specific [219]. One potential approach to take genetic ancestry into consideration is via multi-task deep learning that learns to map SNP genotype data to both HLA alleles and genetic ancestry.

A successful application of deep learning for HLA imputation in admixed populations would likely require careful consideration of network architecture as well as how to take into account genetic ancestry.

Chapter 7

Conclusion

This dissertation illustrates both the application and development of statistical and computational methods in epidemiological and pharmacogenomic studies. Studies in Chapters 2 - 4 apply methods to answer genetic epidemiological questions, and Chapters 5 - 6 contain studies that develop methods for pharmacogenomic and epidemiological studies.

In Chapter 2, we used multi-dimensional scaling to separate admixed samples into African Americans, Hispanics, and Asian Americans to perform admixture mapping study for each group. Using a linear-chain conditional random field implemented by RFMix, we inferred local ancestry from a reference panel to (1) compare local admixture proportions between multiple sclerosis (MS) cases and controls, and to (2) study known risk alleles of different ancestry. In African Americans, cases had increased European ancestry at the class I and *MICB-LST1* regions of the major histocompatibility complex (MHC) compared to controls. In Asian Americans, cases had decreased European ancestry at the *HLA-DQB1* and *HLA-DRB1* loci at the MHC compared to controls. Logistic regression analysis of the prominent MS risk allele *HLA-DRB1*15:01* revealed that the European haplotype conferred three times the disease risk compared to that on the African haplotype in African Americans. Lastly, admixture mapping identified a candidate risk locus for MS by revealing a genomic region near the *ZNF596* gene on chromosome 8 where Hispanic cases had significantly higher proportion of European ancestry compared to controls. All association tests controlled for global admixture differences between cases and controls.

In Chapters 3, we studied the causal relationship between genetic variation, DNA methylation, and Sjögren's syndrome (SS) status in labial salivary glands (LSG) of 64 cases and 67 symptomatic non-cases. Specifically, genome-wide DNA methylation profiling was performed on LSG biopsy samples obtained from 131 female members of the Sjögren's International Collaborative Clinical Alliance (SICCA) registry. *Bumphunter* was used to first identify differential methylated regions (DMRs), then the causal inference test (CIT) was applied to find methylation quantitative trait loci (MeQTL) whose effect on SS is mediated by nearby methylation. DMR analysis yielded 215 DMRs, with the majority located in the MHC on chromosome 6p21.3. Consistent with what is known, regions hypomethylated in cases were enriched for gene sets associated with immune processes. Under the CIT, we discovered a

total of 19 DMR-MeQTL pairs that significantly exhibited a causal mediation relationship. Close to half of these DMRs were within the region spanning the *HLA-DQA1*, *HLA-DQB1*, and *HLA-DQA2* loci at the MHC. The risk conferred by these MeQTLs at the MHC is further substantiated by a previous large genome-wide association study. Our findings are significantly relevant to the potential development of targeted epigenetic therapies for SS, which requires an understanding of the causal relationships between DNA methylation, its influencing factors, and its implications for SS.

In Chapter 4, we report clinically distinct patient clusters in 64 cases and 67 symptomatic non-cases from cluster analysis of genome-wide DNA methylation data from LSG tissue. This was performed by applying a variational autoencoder to find a low dimensional projection of methylation data, followed by hierarchical clustering. We identified four robust patient clusters that partitioned cases into phenotypically mild and severe subgroups. Compared to mild cases, severe cases have higher genetic risk at the major histocompatibility complex and tend to experience hypomethylation at genes implicated in immune processes such as type I interferon response and T cell migration. However, for most phenotypic requirements from the SS classification criteria, the proportions of satisfying patients are not significantly different between severe and mild cases. These results highlight the effectiveness of LSG methylation at capturing disease variation and provide a basis for revision of the SS classification criteria.

In Chapter 5, we present an approach that constructs and compares bipartite graphs describing gene-drug associations for each group of cell line in a pharmacogenomic dataset. Specifically, edges between genes and drugs are weighted by the direction and magnitude of association, which are quantified using sparse canonical correlation analysis (SCCA). We introduce a nuclear norm-based dissimilarity measure to compare graphs from different groups. Agglomerative clustering is implemented to merge groups from different tissue of origin. To suggest meaningful clusters, we generate permutation-based p-values for each merge. The genes highlighted by SCCA serve as candidate drug-response associated biomarkers for associated drugs. We demonstrate our method in hematopoietic and lymphoid malignancies from the CCLE dataset that by combining cell lines from acute myeloid leukemia and chronic myeloid leukemia, more myeloid leukemia-relevant genes surface as top genes associated with drug response. Additionally, when the hierarchy of relationships is known from simulation, our method out-performs existing hierarchical clustering-based approaches in correctly inferring true hierarchy.

In Chapter 6, we present deep learning using convolutional neural network (CNN) as a novel approach to the HLA allele imputation problem at the MHC. In this approach, we use phased SNP genotype data flanking $\pm$250 kb from each HLA locus to simultaneously impute HLA allele haplotyes across loci *HLA-A, -B, -C, -DQA1, -DQB1, -DPA1, -DPB1* and *-DRB1*. We split genotype haplotypes into k-mers and learn low-dimensional representations for each k-mer using an embedding layer of the CNN. Detection of genotype motifs is learned with convolutional layers and genotype information from neighboring loci are jointly used for imputation of a given locus to allow learning of long-range disequilibrium across loci. We show that the CNN learned from known associations between HLA alleles and their tag

SNPs and the model was robust against a selection of hyperparameters. On the T1DGC dataset, we show that our CNN achieved 97.6% imputation accuracy, which is comparable with the best performance achieved with existing HLA imputation methods. By separating the training and imputation stages, our imputation program can involve less runtime than existing imputation programs.

Appendix A

Supplementary Materials for "Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry"

A.1 Supplementary Materials and Methods

Statistical Analysis

Let $p_{d,l}^{(i)}(k)$ denote proportion ancestry k for case individual i at locus l and $p_{c,l}^{(i)}(k)$ the same for control individual i. Then from definition for n cases and m controls

$$z_{l,d}(k) - z_{c,l}(k) = \frac{p_{d,l}^{(1)} + \cdots + p_{d,l}^{(n)}}{n} - \frac{p_{c,l}^{(1)} + \cdots + p_{c,l}^{(m)}}{m} \tag{A.1}$$

Under assumptions that $p_{d,l}^{(1)}, \ldots, p_{d,l}^{(n)}$ and $p_{c,l}^{(1)}, \ldots, p_{c,l}^{(m)}$ are each identically and independently distributed, since related individuals were removed by identity-by-state, the variance $Var(\bar{z}_{l,d}(k) - \bar{z}_{l,c}(k))$ can be expressed as

$$\begin{aligned} Var(\bar{z}_{d,l}(k) - \bar{z}_{c,l}(k)) &= Var\left(\frac{p_{d,l}^{1} + \cdots + p_{d,l}^{n}}{n}\right) + Var\left(\frac{p_{d,l}^{1} + \cdots + p_{d,l}^{m}}{m}\right) \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var(p_{d,l}) + \frac{1}{m^2}\sum_{i=1}^{m} Var(p_{c,l}) \\ &= \frac{Var(p_{d,l})}{n} + \frac{Var(p_{c,l})}{m} \end{aligned} \tag{A.2}$$

For computing the test statistic, $Var(\bar{z}_{d,l}(k) - \bar{z}_{c,l}(k))$ was calculated with empirical estimates of $Var(p_{d,l})$ and $Var(p_{c,l})$.

A.2 Supplementary Results

Risk of MS between European and African HLA alleles in African Americans

HLA Allele	OR	P-value
*DRB1*15:01*	3.01 (1.90-4.75)	2.49E−6
*DRB1*03:01*	0.64 (0.43-0.96)	3.03E−2
*A*02:01*	0.91 (0.63-1.31)	5.94E−1
*DRB1*14:01*	0.40 (0.09-1.74)	2.24E−1
*B*07:02*	1.66 (1.12-2.47)	1.18E−2
*A*03:01*	1.54 (1.04-2.29)	2.97E−2
*C*08:02*	0.66 (0.29-1.54)	3.37E−1
*C*04:01*	1.07 (0.72-1.59)	7.41E−1
*C*07:02*	1.12 (0.75-1.67)	5.82E−1

Table A.1: Odds ratio (OR) of European HLA allele to African HLA allele as determined from logistic regression for African American MS-associated alleles, adjusting for first 3 MDS components. OR are shown with 95% confidence interval and corresponding p-values. HLA alleles with sample size less than 50 or with predominant ancestry greater than 90% are excluded from the analysis. Furthermore, alleles not inferred to be completely European or African are excluded, and only alleles from individuals with one copy are included.

DRB1-DQB1 Haplotye	Case (n)	Control (n)	
EUR *DRB1*15:01-DQB1*06:02*	99	142	241
EUR *DRB1*15:01-DQB1*X*	30	49	79
	129	191	320

Table A.2: *HLA-DRB1*15:01* haplotypes in African Americans. Two-by-two table of counts of *DRB1*15:01-DQB1* haplotypes where *DRB1*15:01* is European and the identity of the *DQB1* allele on the haplotype is summarized. All HLA alleles had allele frequency greater than 0.005, and only *DRB1*15:01* alleles that were completely European were considered. *DQB1*X* denotes any *DQB1* allele that is not *DQB1*06:02*, and that there is no restriction on the ancestry of the *DQB1* allele. Note: *DQB1* alleles did not pass imputation quality cutoff of $R^2 = 0.80$.

DRB1-DQB1 Haplotye	Case (n)	Control (n)	
*DRB1*X*-EUR *DQB1*06:02*	10	38	48
*DRB1*X*-AFR *DQB1*06:02*	137	252	389
	147	290	437

Table A.3: European and African *HLA-DRB1*06:02* haplotypes in African Americans. Two-by-two table of counts of *DRB1*X–DQB1*06:02* haplotypes where *DRB1*X* denotes any allele other than *DRB1*15:01*. All HLA alleles had allele frequency greater than 0.005, and only *DQB1* alleles that were either completely European or African are considered. There is no restriction on the ancestry of the *DRB1* allele. Note: *DQB1* alleles did not pass imputation quality cutoff of $R^2 = 0.80$.

Appendix B

Supplementary Materials for "Hypomethylation of immune genes mediates methylation quantitative trait loci at the major histocompatibility complex in Sjögren's Syndrome"

B.1 Supplementary Materials and Methods

Methylotyping and preprocessing

The two primary measures of DNA methylation of each CpG site are β-values and M-values. A β-value is a ratio of the methylated probe intensity to the sum of methylated and unmethylated probe intensities, which ranges from 0 to 1, and reflects the proportion of methylation at a CpG site, and is more interpretable. The M-value can be derived from a β-value as $\log_2 \frac{\beta}{1-\beta}$, ranges from $-\infty$ and ∞, and has been the ideal measure to use for identifying differentially methylated CpG sites due to less severe heteroscedasticity [27].

Removing unwanted DNA methylation variation

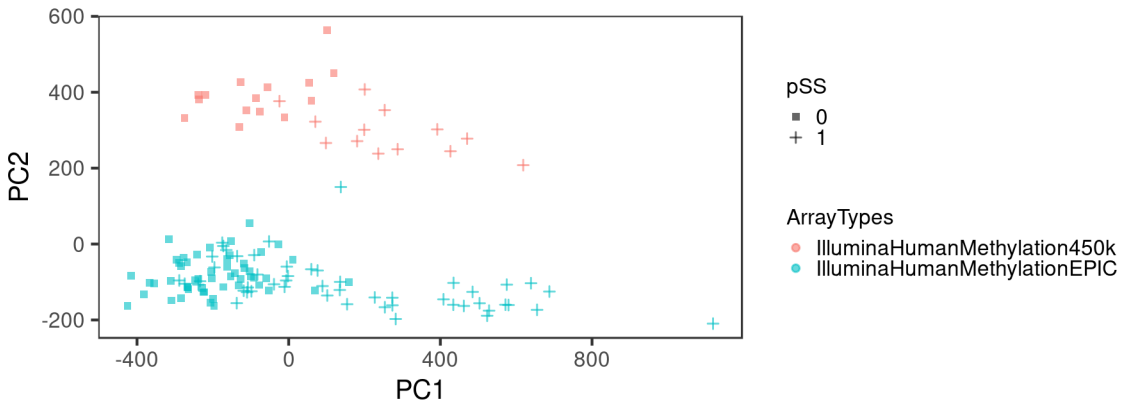


Figure B.1: PCA of preprocessed, β-values prior to batch-correction with `ComBat`. The array type (450K or EPIC) for methylotyping is indicated by color. The array types 450K and EPIC show strong separation on PC2.

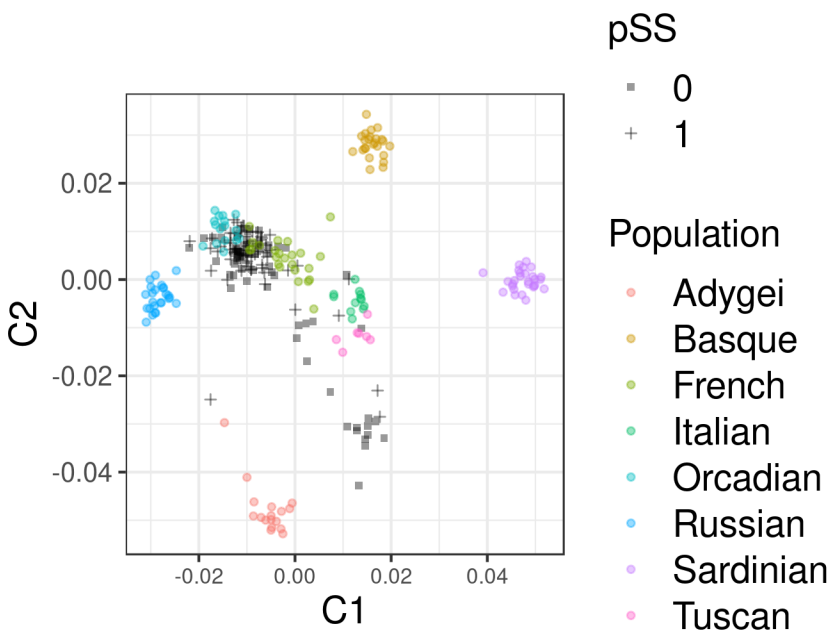


Figure B.2: Multidimensional scaling analysis of 131 SICCA study subjects with HGDP reference European samples.

Parametric adjustment was used for `ComBat` because the densities of the additive and multiplicative batch parameters were neither highly skewed nor bimodal (Figure B.3). Missing methylation values were mean imputed per CpG site before applying `ComBat`, then missingness restored after adjustment.

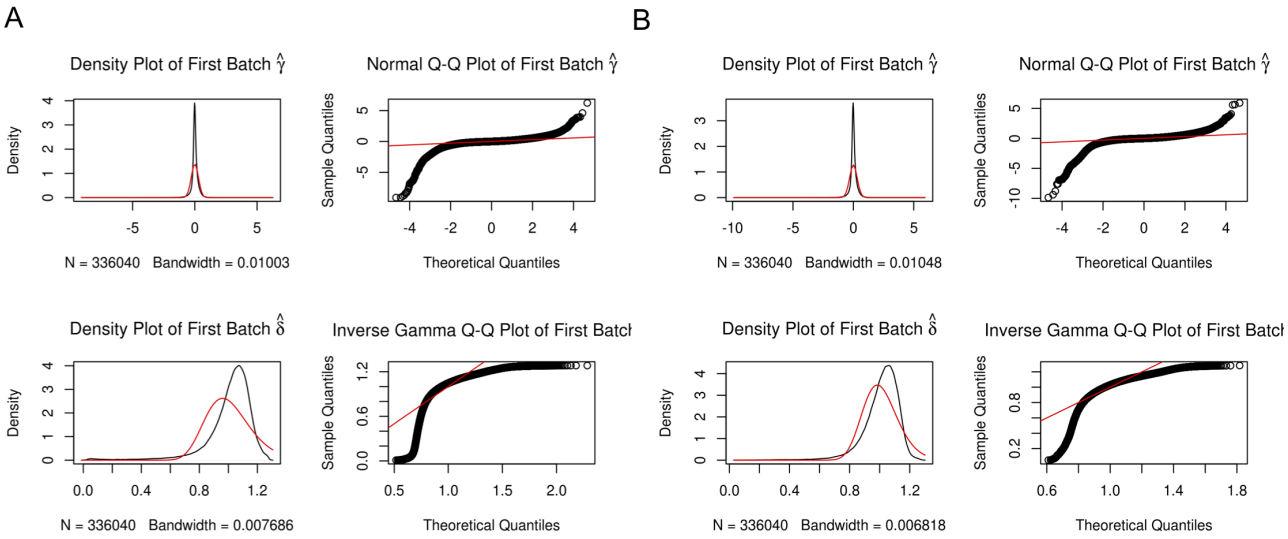


Figure B.3: Prior plot of kernel estimate of batch effect (black) and parametric estimate of batch effect (red) from batch correction using `ComBat` for (A) β-values and (B) M-values.

Dimensionality reduction

Let n and p be the numbers of subjects and CpG sites respectively. Principal component analysis (PCA) was performed on the centered and scaled β-value matrix $X \in \mathbb{R}^{n \times p}$. Missing values were replaced with per CpG site average before PCA. The symmetric matrix $X^\top X \in \mathbb{R}^{p \times p}$ has the eigendecomposition $X^\top X = V \Sigma V^\top$, where $V \in \mathbb{R}^{p \times p}$ is an orthogonal matrix. In PCA, the columns $v_1, \ldots, v_p$ of V specify the optimal orthogonal directions to project samples onto to preserve variability in the data. Principal component i (PCi) refers to projections of the n samples onto v_i, which we refer to as the i-th principal axis. For example, PC1 is computed as $\text{PC1} = Xv_1$, which can be seen here as a linear combination of the CpG site methylation levels in X. The first principal axis v_1 thus contains coefficients for each CpG site, with larger magnitudes indicating greater contribution to PC1. We refer to the absolute value of entries in v_1 as the loadings for PC1, which we analyzed to determine which CpG sites contributed most to PC1.

Identification of differentially methylated regions

Bootstrap resampling was run with option `nullMethod = bootstrap` so that adjustment covariates were controlled for. One of the most important hyperparameters of bumphunter is the effect size cutoff for defining a candidate differentially methylated region (DMR), where effect size is the estimated expected change in methylation from one group to the other. In this study, a cutoff of 1.0 was chosen to achieve a balance between effect size, number of bumps found, and bump sizes in terms of number of CpG sites (Figure $B.4$).

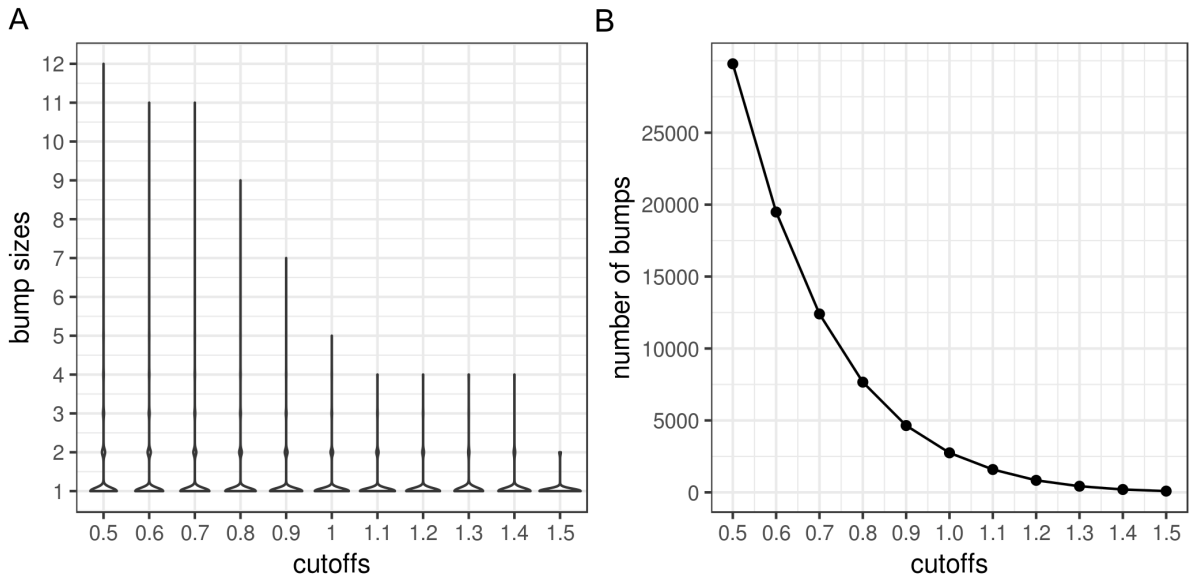


Figure B.4: Number of bumps found for SS and their sizes at different *bumphunter* coefficient cutoffs. (A) Violin plot of bump sizes at each cutoff (B) Number of bumps discovered at each cutoff.

`Minfi` was used to annotate each DMR with its nearest gene in base pairs, location relative to nearest gene, and location relative to nearest CpG island. The DMR location relative to nearest CpG island was set as the majority location of all CpGs that comprise the DMR. Detailed descriptions of each DMR gene were obtained from the National Center for Biotechnology Information.

Gene set enrichment analysis

The gene ontology (GO) gene sets total 5,917, with 4,436 derived from biological process ontology, 580 derived from cellular component ontology, and 901 derived from molecular function ontology. Additionally, we included two gene sets consisting of genes shown to be differentially methylated or differentially expressed respectively, between SS cases and controls in labial salivary gland (LSG) [104, 130]. We eliminated large gene sets numbering more than 100 genes, retaining approximately 76% of gene sets.

Mediation analysis with causal inference test

The causal inference test (CIT) consists of statistical tests evaluating the following necessary and sufficient conditions for the causal mediation model involving genotype "G", methylation "M", and case status "S",

1. $S \sim G$
2. $G \sim M|S$
3. $M \sim S|G$
4. $S \perp\!\!\!\perp G|M$,

where "$\sim$" denotes associated with and "$\perp\!\!\!\perp$" denotes independent of. Condition 1 is tested with the likelihood ratio test arising from logistic regression in Equation B.1

$$logit(S_i) \sim \beta_0 + \beta_1 G_{i1} + \beta_2 G_{i2} + \epsilon_{i1}, \tag{B.1}$$

with the null and alternative hypotheses as $H_0 : \{\beta_1 = 0, \beta_2 = 0\}$, $H_1 : \{\beta_1 \neq 0, \beta_2 \neq 0\}$. Condition 2 is tested with the F-test arising from linear regression in Equation B.2

$$M_i \sim \beta_0 + \beta_1 S_i + \beta_2 G_{i1} + \beta_3 G_{i2} + \epsilon_{i2}, \tag{B.2}$$

with the null and alternative hypotheses as $H_0 : \{\beta_2 = 0, \beta_3 = 0\}$, $H_1 : \{\beta_2 \neq 0, \beta_3 \neq 0\}$. Condition 3 is tested with the likelihood ratio test arising from logistic regression in Equation B.3

$$logit(S_i) \sim \beta_0 + \beta_1 M_i + \beta_2 G_{i1} + \beta_3 G_{i2} + \epsilon_{i3}, \tag{B.3}$$

with the null and alternative hypotheses as $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$. For condition 4, a proper hypothesis test requires estimating a null distribution under the independence model, which is defined by the conditions

1. G is causal for M
2. G is causal for S
3. $S \perp\!\!\!\perp M|G$

and illustrated in Figure B.5.

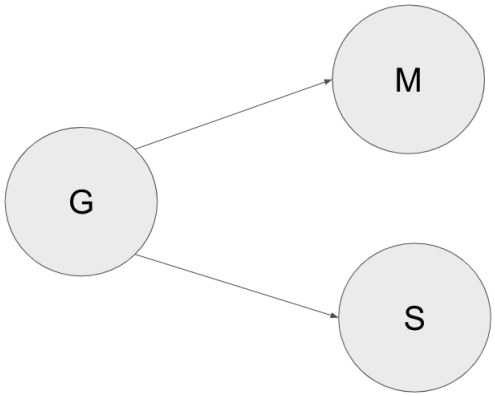


Figure B.5: Independence model. G = genotype; M = methylation; S = Sjögren's syndrome case status.

The test statistic null distribution is generated in the following manner.

1. Simulate M^* according to marginal effect of G on M, thereby breaking residual dependence between S and $M|G$.
   a) Regression: $M_i = \beta_0 + \beta_1 G_{i1} + \beta_2 G_{i2} + \epsilon_i$
   b) Residuals are randomly permuted to obtain ϵ_i^*, and set $M_i^* := \beta_0 + \beta_1 G_{i1} + \beta_2 G_{i2} + \epsilon_i^*$
2. Fit the following logistic regressions and conduct likelihood ratio test to obtain test statistic T^*
   - $logit(S_i) = \beta_0 + \beta_1 M_i^* + \beta_2 G_{i1} + \beta_3 G_{i2} + \epsilon$
   - $logit(S_i) = \beta_0 + \beta_1 M_i^* + \epsilon_i$
3. Repeat steps 1 - 2 B times to obtain empirical distribution of T^* under the null independence model.
4. Observed test statistic T from logistic regressions with observed M_i against empirical distribution of T^* to obtain p-value.

The CIT was run with default settings for pairs of meQTLs and DMRs whose association was determined to be significant after multiple hypothesis testing adjustment. The q-value, or false discovery rate, for a CIT was estimated based on simulating the null CIT outcome by permuting the relevant variable for each statistical test comprising the CIT, which was done with `n.perm = 100` permutations, as recommended to be sufficient by Millsten *et al.*[129] The permutations were specified the same across all tests to account for dependencies among the tests.

B.2 Supplementary Results

DNA methylation mediates the effect of MeQTL on SS at the MHC

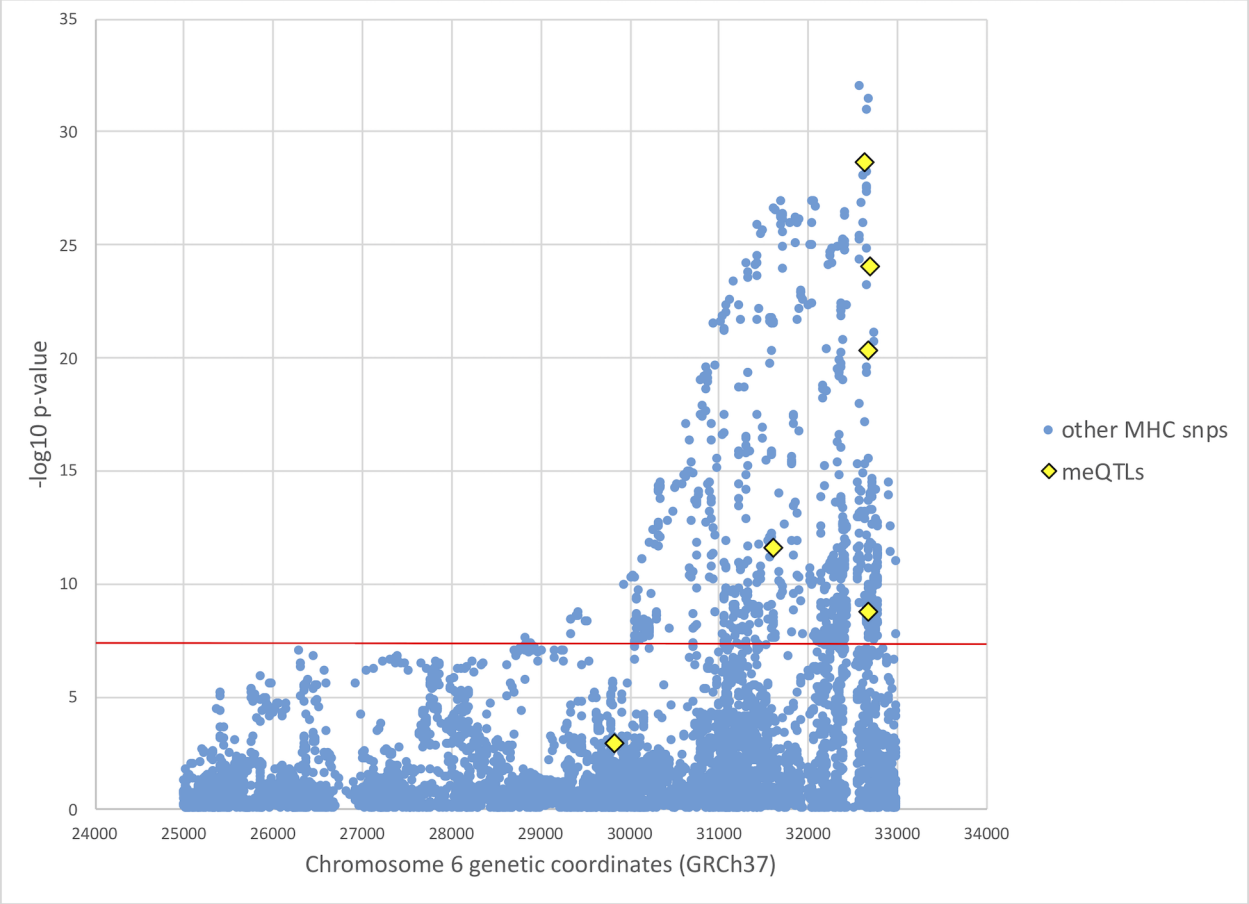


Figure B.6: Manhattan plot of genome-wide association study results at the MHC for SS. Genome-wide association study from Taylor *et al.*[99], with mediating meQTL p-values from this study colored in yellow. The red horizontal line indicates genome-wide significance level from Taylor *et al.*[99]

Model	OR	OR Std Err	95% Confidence Interval	p-value
rs2734985	0.952	0.071	0.823 - 1.101	0.505
rs2261033	0.875	0.065	0.756 - 1.01	0.072
rs3021302	1.688	0.153	1.413 - 2.015	0.000
rs9275224	1.257	0.117	1.048 - 1.509	0.014
rs9275374	1.066	0.105	0.878 - 1.294	0.519
rs2858332	1.304	0.112	1.102 - 1.544	0.002
constant	0.030	0.006	0.020 - 0.046	0.000

Table B.1: Logistic regression of SS case status against putative MeQTLs at the MHC. Logistic regression results for all six putative MeQTLs within the MHC, adjusting for first two European principal components, sex, and smoking status in the European GWAS dataset from Taylor et al.[99] OR = odds ratio; Std Err = standard error.

B.3 Supplementary Discussion

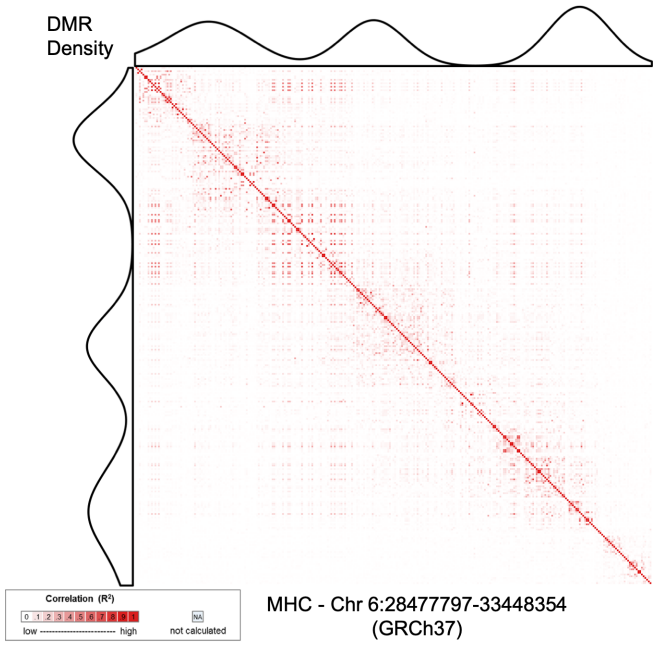


Figure B.7: Linkage disequilibrium (R^2) heatmap of SNPs at the MHC in European populations from the 1000 Genomes Project [220], with corresponding density plot of DMR locations placed at the top and left margins.

Appendix C

Supplementary Materials for "Epigenetic stratification identifies clinically-relevant disease subgroups in Sjögren's syndrome with differential genetic risk at the major histocompatibility complex"

C.1 Supplementary Materials and Methods

Study subjects and clinical evaluation

Variable name	Variable description
ana	Detection of antinuclear antibody at the 1:40 concentration level
igg	Immunoglobulin G (IgG) result
c3	Complement component 3 (C3) result
c4	Complement component 4 (C4) result
ssb	Anti-SS-B result
rf	Rheumatoid factor result
tbul_time	Tear break-up time left eye if less than 10 seconds
tbur_time	Tear break-up time right eye if less than 10 seconds
uws	Unstimulated whole salivary flow rate
focus	Focus score
ossr	Ocular SICCA score right eye
ossl	Ocular SICCA score left eye
rparenlg	Right parotid gland enlargement
lparenlg	Left parotid gland enlargement
drymouth	Dry mouth symptoms
liqmouth	Need liquids for swallowing
dryeye	Dry eye symptoms
lymphoma	Physician confirmed lymphoma
GC_like_formation	Presence of germinal center-like formation tested with H&E staining; only tested in individuals with focal or focal/sclerosing lymphocytic sialendenitis
thyroid	Physician confirmed thyroid disease
liver	Physician confirmed liver disease
kidney	Physician confirmed kidney disease
othersys	Physician confirmed other systemic disease
pSS	Primary Sjögren's syndrome case status
bq29_tiredLowEnergy	Feeling tired
bq21_painWork	During the past 4 weeks, how much did pain interfere with your normal work (outside and inside home)?
bq69g_painEyes	Pain or burning in the middle of the night or upon waking in the morning
bq22_calmPeaceful	Calm and peaceful

b_phq9_deprSeverity	PHQ-9 depression severity
bq27_downDepr	Feeling down, depressed
bq24_downheartedDepr	Down hearted and depressed
bq38_dryvaginal	Have you had significant vaginal dryness
bq56_painMouth	In the past year, have you avoided eating certain foods you wanted because they made your mouth hurt?

Table C.1: Clinical phenotype data key.

Removing unwanted DNA methylation variation

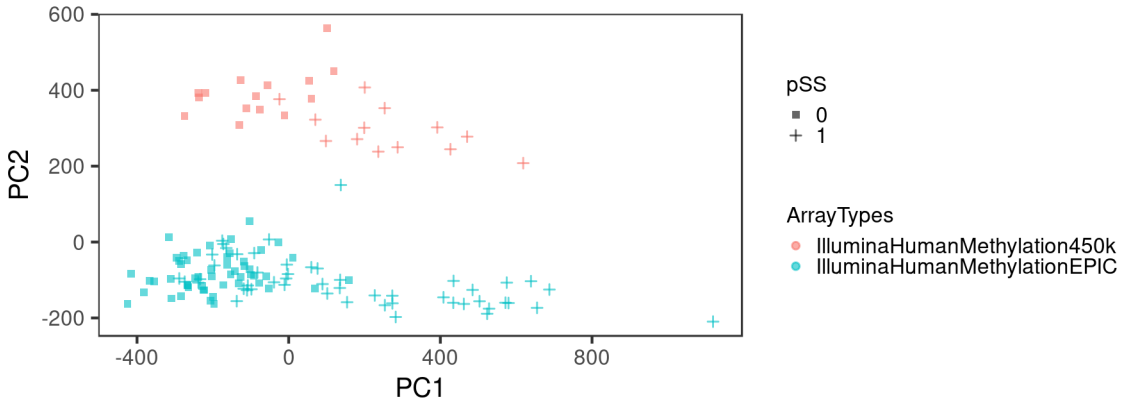


Figure C.1: PCA of β-values without batch-correction. The methylotyping array (i.e. 450K or EPIC) is indicated by color. PC2 captures variation in DNA methylation explained by array type.

C.2 Supplementary Results

Identification of patient clusters

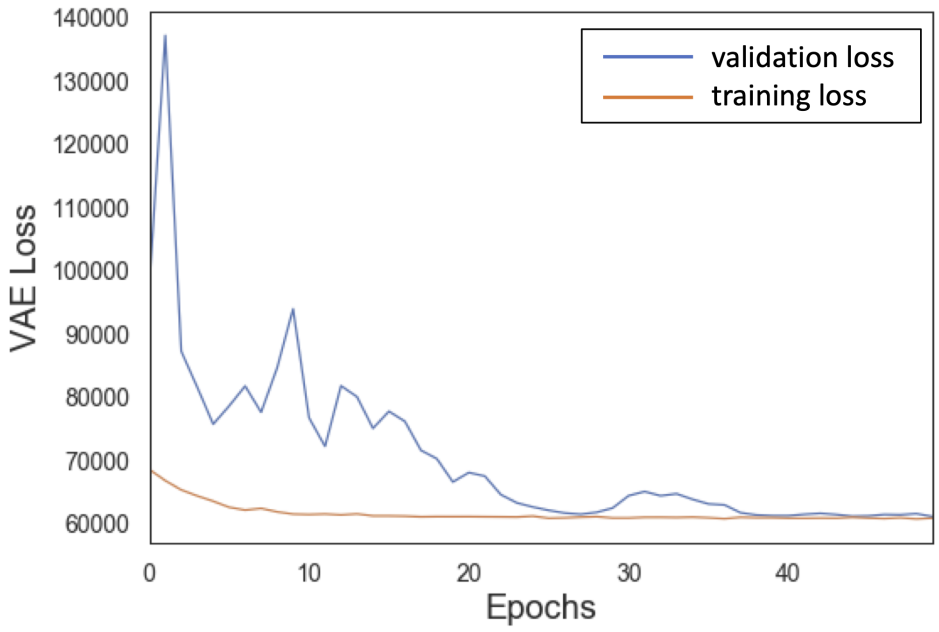


Figure C.2: VAE training and validation loss.

Clinical phenotype analysis by cluster and disease subgroup

	Cluster 1	Cluster 2	Cluster 3	Cluster 4	p-value
bq29_tiredLowEnergy	2.81	3.00	3.10	2.69	0.446
bq21_painWork	2.85	2.41	2.94	2.69	0.510
bq69g_painEyes	2.19	2.35	2.37	2.28	0.980
bq22_calmPeaceful	3.08	2.76	2.83	2.86	0.522
b_phq9_deprSeverity	2.15	2.18	2.37	2.03	0.604
bq27_downDepr	1.58	1.53	1.85	1.58	0.523
bq24_downheartedDepr	2.31	2.00	2.48	2.11	0.292
bq38_dryvaginal	0.73	0.71	0.63	0.86	0.140
bq56_painMouth	0.46	0.71	0.33	0.61	0.012

Table C.2: Averages of self-reported SS symptoms, by patient cluster, determined from VAE-based clustering analysis. P-values were computed using Kruskal-Wallis test for ordinal or continuous clinical phenotypes, and computed using chi-square test of independence for categorical or binary phenotypes. Refer to Table C.1 for key of clinical phenotype abbreviations. Note the average is equivalent to proportion for binary phenotypes.

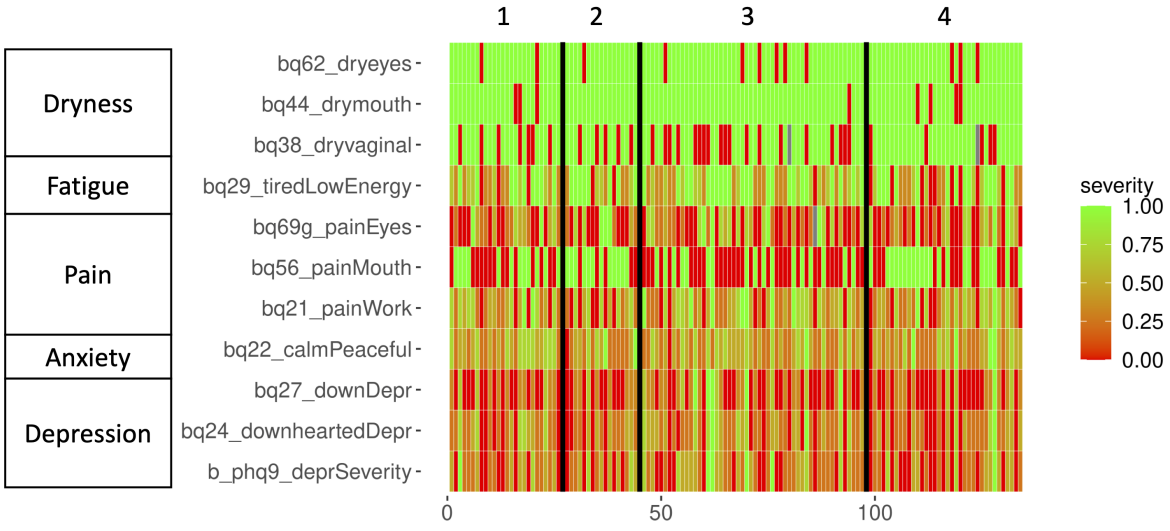


Figure C.3: Heatmap of self-reported SS symptoms. All phenotypes are either ordinal or binary, and normalized between 0 and 1, with larger values indicative of greater severity. Clinical phenotypes are grouped by general categories of dryness, fatigue, pain, anxiety, and depression. Each column represents a patient and all 131 subjects are grouped by patient clusters. Gray indicates missingness. See Table C.1 for clinical phenotype key.

	Mild cases	**Severe cases**	**p-value**
bq29_tiredLowEnergy	2.75	2.88	0.65
bq21_painWork	2.67	2.68	0.92
bq69g_painEyes	2.78	2.23	0.15
bq22_calmPeaceful	2.79	2.90	0.67
b_phq9_deprSeverity	2.00	2.15	0.52
bq27_downDepr	1.54	1.50	1.00
bq24_downheartedDepr	2.25	2.15	0.76
bq38_dryvaginal	0.71	0.73	1.00
bq56_painMouth	0.58	0.58	1.00

Table C.3: Analysis of self-reported SS symptoms, by disease subgroup. Averages of self-reported SS symptoms for severe cases and mild cases. Severe cases belong to clusters 1 and 2 and mild cases belong to clusters 3 and 4 from the VAE-based clustering analysis. P-values were computed using Wilcoxon rank sum test for ordinal or continuous clinical phenotypes, and computed using chi-square test of independence for categorical or binary phenotypes. Refer to Table C.1 for key of clinical phenotype abbreviations. Note the average is equivalent to proportion for binary phenotypes.

Appendix D

Supplementary Materials for "Bipartite graph-based approach for clustering of cell lines by gene expression-drug response associations"

D.1 Supplementary Materials and Methods

Dissimilarity measure

The nuclear norm for a matrix A of rank r is defined as $||A||_* = \sum_{i=1}^{r} \sigma_i(A)$, where σ_i is the i-th singular value of A. If we let $\sigma(A) \in \mathbb{R}^r$ denote the vector of singular values, then $||A||_* = ||\sigma(A)||_1$ since $\sigma_i(A) \geq 0$ for $i = 1, \ldots, r$. Then an upper bound on the nuclear norm is

$$||A||_* = ||\sigma(A)||_1 \leq \sqrt{r}||\sigma(A)||_2 = \sqrt{r}||A||_F, \tag{D.1}$$

where $||A||_F$ denotes the Frobenius norm of A. The inequality in Equation D.1 comes from applying Cauchy–Schwarz inequality

$$||\sigma(A)||_1 = \mathbf{1}_r^\top \sigma(A) \leq \sqrt{r}||\sigma(A)||_2, \tag{D.2}$$

where $\mathbf{1}_r \in \mathbb{R}^r$ is a vector of ones. The second equality is a consequence of singular value decomposition (SVD) and commutative property of trace

$$\begin{aligned} ||A||_F &= \sqrt{tr(A^\top A)} \\ &= \sqrt{tr(V^\top \Sigma^2 V)} \\ &= \sqrt{tr(\Sigma^2 V^\top V)} \\ &= \sqrt{tr(\Sigma^2)} \\ &= ||\sigma(A)||_2, \end{aligned} \tag{D.3}$$

where A has the SVD $A = U\Sigma V^\top$.

Agglomerative merging of cell line groups

Empirically, merge dissimilarity generally increases monotonically, but inversions where a parent group is merged at a lower dissimilarity than its children can still occur. To ensure height monotonicity for the resulting dendrogram, we define each merge height as the merge dissimilarity plus maximum height of the two groups. Specifically for two groups u and v with heights h_u and h_v respectively, their merge height h_{uv} is set as

$$h_{uv} = \max\{h_u, h_v\} + d(B^{[u]}, B^{[v]}), \tag{D.4}$$

CCLE dataset

The data preprocessing details are:

1. Removed 8 drugs due to missing values across cell lines
2. Removed 5 cell lines with missing values across drugs
3. Removed 31 cell lines without drug sensitivity measurements
4. Removed genomic features with variance less than 3
5. Removed 77 cell lines from tissue with less than 15 cell lines

Simulation

Additional details on simulating genomic features matrix $X \in \mathbb{R}^{n\times p}$ and drug responses matrix $Y \in \mathbb{R}^{n\times d}$. The genomic covariance matrix $\Sigma \in \mathbb{R}^{p\times p}$ for a given group is generated as follows:

1. Generate random orthonormal matrix $A \in \mathbb{R}^{n\times n}$ using `randortho`$(\cdot)$ function from `pracma` R package.

2. Simulate $\tilde{X} \in \mathbb{R}^{n \times p}$ as affine combinations of 10 randomly selected columns from A. Thus, columns of $\tilde{X}$ are highly correlated.

3. Construct initial covariance matrix $\Sigma = \tilde{X}^\top \tilde{X}$.

4. Reduce cross-covariance terms between biomarker genes and non-biomarker genes, and between any pair of non-biomarker genes by a factor $k = 10$. Thus, only biomarker genes have high covariance terms with each other.

Bibliography

[1] Daniel Shriner. "Overview of admixture mapping". In: *Current protocols in human genetics* 94.1 (2017), pp. 1–23.

[2] Ariel Darvasi and Sagiv Shifman. "The beauty of admixture". In: *Nature Genetics* 37.2 (2005), pp. 118–119.

[3] Calvin Chi et al. "Admixture mapping reveals evidence of differential multiple sclerosis risk by genetic ancestry". In: *PLoS genetics* 15.1 (2019), e1007808.

[4] Alkes L Price et al. "Sensitive detection of chromosomal segments of distinct ancestry in admixed populations". In: *PLoS genetics* 5.6 (2009).

[5] Bogdan Paşaniuc et al. "Inference of locus-specific ancestry in closely related populations". In: *Bioinformatics* 25.12 (2009), pp. i213–i221.

[6] Yael Baran et al. "Fast and accurate inference of local ancestry in Latino populations". In: *Bioinformatics* 28.10 (2012), pp. 1359–1367.

[7] Yongtao Guan. "Detecting structure of haplotypes and local ancestry". In: *Genetics* 196.3 (2014), pp. 625–642.

[8] Brian K Maples et al. "RFMix: a discriminative modeling approach for rapid and robust local-ancestry inference". In: *The American Journal of Human Genetics* 93.2 (2013), pp. 278–288.

[9] Sharon R Browning and Brian L Browning. "Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering". In: *The American Journal of Human Genetics* 81.5 (2007), pp. 1084–1097.

[10] Susana Eyheramendy et al. "Genetic structure characterization of Chileans reflects historical immigration patterns". In: *Nature communications* 6 (2015), p. 6472.

[11] Maxim VC Greenberg and Deborah Bourc'his. "The diverse roles of DNA methylation in mammalian development and disease". In: *Nature reviews Molecular cell biology* (2019), pp. 1–18.

[12] Michael B Stadler et al. "DNA-binding factors shape the mouse methylome at distal regulatory regions". In: *Nature* 480.7378 (2011), pp. 490–495.

[13] Ryan Lister et al. "Human DNA methylomes at base resolution show widespread epigenomic differences". In: *nature* 462.7271 (2009), pp. 315–322.

[14] Christina M Bender et al. "Roles of cell division and gene transcription in the methylation of CpG islands". In: *Molecular and cellular biology* 19.10 (1999), pp. 6690–6698.

[15] Katherine E Varley et al. "Dynamic DNA methylation across diverse human cell lines and tissues". In: *Genome research* 23.3 (2013), pp. 555–567.

[16] En Li, Timothy H Bestor, and Rudolf Jaenisch. "Targeted mutation of the DNA methyltransferase gene results in embryonic lethality". In: *Cell* 69.6 (1992), pp. 915–926.

[17] Masaki Okano et al. "DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development". In: *Cell* 99.3 (1999), pp. 247–257.

[18] Alika K Maunakea et al. "Conserved role of intragenic DNA methylation in regulating alternative promoters". In: *Nature* 466.7303 (2010), pp. 253–257.

[19] Stephen B Baylin and Peter A Jones. "Epigenetic determinants of cancer". In: *Cold Spring Harbor perspectives in biology* 8.9 (2016), a019505.

[20] Lucas T Husquin et al. "Exploring the genetic basis of human population differences in DNA methylation and their causal impact on immune gene regulation". In: *Genome biology* 19.1 (2018), pp. 1–17.

[21] Qiaoling Li et al. "Folate deficiency and aberrant DNA methylation and expression of FHIT gene were associated with cervical pathogenesis". In: *Oncology letters* 15.2 (2018), pp. 1963–1972.

[22] Carrie V Breton et al. "Prenatal tobacco smoke exposure affects global and gene-specific DNA methylation". In: *American journal of respiratory and critical care medicine* 180.5 (2009), pp. 462–467.

[23] Sonja Zeilinger et al. "Tobacco smoking leads to extensive genome-wide changes in DNA methylation". In: *PloS one* 8.5 (2013).

[24] Michael A Pereira et al. "Prevention by methionine of dichloroacetic acid–induced liver cancer and DNA hypomethylation in mice". In: *Toxicological Sciences* 77.2 (2004), pp. 243–248.

[25] Hyeran Jang, Joel B Mason, and Sang-Woon Choi. "Genetic and epigenetic interactions between folate and aging in carcinogenesis". In: *The Journal of nutrition* 135.12 (2005), 2967S–2971S.

[26] J Richard Pilsner et al. "Genomic methylation of peripheral blood leukocyte DNA: influences of arsenic and folate in Bangladeshi adults". In: *The American journal of clinical nutrition* 86.4 (2007), pp. 1179–1186.

[27] Pan Du et al. "Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis". In: *BMC bioinformatics* 11.1 (2010), p. 587.

[28] Florian Eckhardt et al. "DNA methylation profiling of human chromosomes 6, 20 and 22". In: *Nature genetics* 38.12 (2006), pp. 1378–1385.

[29] Batbayar Khulan et al. "Comparative isoschizomer profiling of cytosine methylation: the HELP assay". In: *Genome research* 16.8 (2006), pp. 1046–1055.

[30] Eiko Kitamura et al. "Analysis of tissue-specific differentially methylated regions (TDMs) in humans". In: *Genomics* 89.3 (2007), pp. 326–337.

[31] Robert Illingworth et al. "A novel CpG island set identifies tissue-specific methylation at developmental gene loci". In: *PLoS biology* 6.1 (2008).

[32] Wikimedia Commons. *Work flow of Illumina Methylation Assay using the Infinium II platform*. 2009. URL: `https://commons.wikimedia.org/wiki/File:Illuminamethylationworkflow.png`.

[33] Yi-an Chen et al. "Discovery of cross-reactive probes and polymorphic CpGs in the Illumina Infinium HumanMethylation450 microarray". In: *Epigenetics* 8.2 (2013), pp. 203–209.

[34] Daniel L McCartney et al. "Identification of polymorphic and off-target probe binding sites on the Illumina Infinium MethylationEPIC BeadChip". In: *Genomics data* 9 (2016), pp. 22–24.

[35] Ruth Pidsley et al. "Critical evaluation of the Illumina MethylationEPIC BeadChip microarray for whole-genome DNA methylation profiling". In: *Genome biology* 17.1 (2016), p. 208.

[36] Patrycja Daca-Roszak et al. "Impact of SNPs on methylation readouts by Illumina Infinium HumanMethylation450 BeadChip Array: implications for comparative population studies". In: *BMC genomics* 16.1 (2015), p. 1003.

[37] Adiv A Johnson et al. "The role of DNA methylation in aging, rejuvenation, and age-related disease". In: *Rejuvenation research* 15.5 (2012), pp. 483–494.

[38] Lutz Philipp Breitling et al. "Smoking, F2RL3 methylation, and prognosis in stable coronary heart disease". In: *European heart journal* 33.22 (2012), pp. 2841–2848.

[39] Bonnie R Joubert et al. "450K epigenome-wide scan identifies differential DNA methylation in newborns related to maternal smoking during pregnancy". In: *Environmental health perspectives* 120.10 (2012), pp. 1425–1431.

[40] Robert A Philibert, Steven RH Beach, and Gene H Brody. "Demethylation of the aryl hydrocarbon receptor repressor as a biomarker for nascent smokers". In: *Epigenetics* 7.11 (2012), pp. 1331–1338.

[41] Natalie S Shenker et al. "Epigenome-wide association study in the European Prospective Investigation into Cancer and Nutrition (EPIC-Turin) identifies novel genetic loci associated with smoking". In: *Human molecular genetics* 22.5 (2013), pp. 843–851.

[42] Ronald M Adkins et al. "Racial differences in gene-specific DNA methylation levels are present at birth". In: *Birth Defects Research Part A: Clinical and Molecular Teratology* 91.8 (2011), pp. 728–736.

[43] Jean-Pierre Gillet, Sudhir Varma, and Michael M Gottesman. "The clinical relevance of cancer cell lines". In: *Journal of the National Cancer Institute* 105.7 (2013), pp. 452–458.

[44] Jimmy Caroli, Martina Dori, and Silvio Bicciato. "Computational methods for the integrative analysis of genomics and pharmacological data". In: *Frontiers in Oncology* 10 (2020).

[45] Robert H Shoemaker. "The NCI60 human tumour cell line anticancer drug screen". In: *Nature Reviews Cancer* 6.10 (2006), pp. 813–823.

[46] Mahmoud Ghandi et al. "Next-generation characterization of the cancer cell line encyclopedia". In: *Nature* 569.7757 (2019), pp. 503–508.

[47] Mathew J Garnett et al. "Systematic identification of genomic markers of drug sensitivity in cancer cells". In: *Nature* 483.7391 (2012), pp. 570–575.

[48] Wanjuan Yang et al. "Genomics of Drug Sensitivity in Cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells". In: *Nucleic acids research* 41.D1 (2012), pp. D955–D961.

[49] Justin Lamb et al. "The Connectivity Map: using gene-expression signatures to connect small molecules, genes, and disease". In: *science* 313.5795 (2006), pp. 1929–1935.

[50] Benjamin Haibe-Kains et al. "Inconsistency in large pharmacogenomic studies". In: *Nature* 504.7480 (2013), pp. 389–393.

[51] Michael AA Cox and Trevor F Cox. "Multidimensional scaling". In: *Handbook of data visualization*. Springer, 2008, pp. 315–347.

[52] Daniel Falush, Matthew Stephens, and Jonathan K Pritchard. "Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies". In: *Genetics* 164.4 (2003), pp. 1567–1587.

[53] Judea Pearl et al. "Causal inference in statistics: An overview". In: *Statistics surveys* 3 (2009), pp. 96–146.

[54] Brooke Rhead et al. "Mendelian randomization shows a causal effect of low vitamin D on multiple sclerosis risk". In: *Neurology Genetics* 2.5 (2016), e97.

[55] Gareth James et al. *An introduction to statistical learning*. Vol. 112. Springer, 2013.

[56] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. `http://www.deeplearningbook.org`. MIT Press, 2016.

[57] Tomas Olsson, Lisa F Barcellos, and Lars Alfredsson. "Interactions between genetic, lifestyle and environmental risk factors for multiple sclerosis". In: *Nature Reviews Neurology* 13.1 (2017), p. 25.

[58] Sergio E Baranzini and Jorge R Oksenberg. "The genetics of multiple sclerosis: from 0 to 200 in 50 years". In: *Trends in Genetics* 33.12 (2017), pp. 960–970.

[59] Loukas Moutsianas et al. "Class II HLA interactions modulate genetic risk for multiple sclerosis". In: *Nature genetics* 47.10 (2015), p. 1107.

[60] VV Bashinskaya et al. "A review of genome-wide association studies for multiple sclerosis: classical and hypothesis-driven approaches". In: *Human genetics* 134.11-12 (2015), pp. 1143–1162.

[61] Antonio Alcina et al. "Multiple sclerosis risk variant HLA-DRB1* 1501 associates with high expression of DRB1 gene in different human populations". In: *PloS one* 7.1 (2012).

[62] Alessandro Didonna and Jorge R Oksenberg. "Genetic determinants of risk and progression in multiple sclerosis". In: *Clinica chimica acta* 449 (2015), pp. 16–22.

[63] Lekha Pandit et al. "European multiple sclerosis risk variants in the south Asian population". In: *Multiple Sclerosis Journal* 22.12 (2016), pp. 1536–1540.

[64] Noriko Isobe et al. "An ImmunoChip study of multiple sclerosis risk in African Americans". In: *Brain* 138.6 (2015), pp. 1518–1530.

[65] Noriko Isobe et al. "Genetic risk variants in African Americans with multiple sclerosis". In: *Neurology* 81.3 (2013), pp. 219–227.

[66] Jennifer P Rubin and Nancy L Kuntz. "Diagnostic criteria for pediatric multiple sclerosis". In: *Current neurology and neuroscience reports* 13.6 (2013), p. 354.

[67] W Ian McDonald et al. "Recommended diagnostic criteria for multiple sclerosis: guidelines from the International Panel on the diagnosis of multiple sclerosis". In: *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society* 50.1 (2001), pp. 121–127.

[68] Mark N Kvale et al. "Genotyping informatics and quality control for 100,000 subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohort". In: *Genetics* 200.4 (2015), pp. 1051–1060.

[69] Ewan Birney and Nicole Soranzo. "Human genomics: The end of the start for population sequencing". In: *Nature* 526.7571 (2015), pp. 52–53.

[70] Bryan Howie et al. "Fast and accurate genotype imputation in genome-wide association studies through pre-phasing". In: *Nature genetics* 44.8 (2012), pp. 955–959.

[71] Xiaoming Jia et al. "Imputing amino acid polymorphisms in human leukocyte antigens". In: *PloS one* 8.6 (2013).

[72] Pierre-Antoine Gourraud et al. "HLA diversity in the 1000 genomes dataset". In: *PloS one* 9.7 (2014).

[73] Kelly Nunes et al. "HLA imputation in an admixed population: An assessment of the 1000 Genomes data as a training set". In: *Human immunology* 77.3 (2016), pp. 307–312.

[74] Shaun Purcell et al. "PLINK: a tool set for whole-genome association and population-based linkage analyses". In: *The American journal of human genetics* 81.3 (2007), pp. 559–575.

[75] Anil Raj, Matthew Stephens, and Jonathan K Pritchard. "fastSTRUCTURE: variational inference of population structure in large SNP data sets". In: *Genetics* 197.2 (2014), pp. 573–589.

[76] Howard M Cann et al. "A human genome diversity cell line panel". In: *Science* 296.5566 (2002), pp. 261–262.

[77] Katarzyna Bryc et al. "The genetic ancestry of african americans, latinos, and european Americans across the United States". In: *The American Journal of Human Genetics* 96.1 (2015), pp. 37–53.

[78] Giovanni Montana and Jonathan K Pritchard. "Statistical tests for admixture mapping with case-control and cases-only data". In: *The American Journal of Human Genetics* 75.5 (2004), pp. 771–789.

[79] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[80] Andrew D Skol et al. "Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies". In: *Nature genetics* 38.2 (2006), pp. 209–213.

[81] Piyameth Dilokthornsakul et al. "Multiple sclerosis prevalence in the United States commercially insured population". In: *Neurology* 86.11 (2016), pp. 1014–1021.

[82] Nikolaos A Patsopoulos et al. "Fine-mapping the genetic association of the major histocompatibility complex in multiple sclerosis: HLA and non-HLA effects". In: *PLoS Genet* 9.11 (2013), e1003926.

[83] Bruce AC Cree et al. "Modification of multiple sclerosis phenotypes by African ancestry at HLA". In: *Archives of neurology* 66.2 (2009), pp. 226–233.

[84] Paul IW De Bakker et al. "A high-resolution HLA and SNP haplotype map for disease association studies in the extended human MHC". In: *Nature genetics* 38.10 (2006), pp. 1166–1172.

[85] David Reich et al. "A whole-genome admixture scan finds a candidate locus for multiple sclerosis susceptibility". In: *Nature genetics* 37.10 (2005), pp. 1113–1118.

[86] Tamar Sofer et al. "Admixture mapping in the Hispanic Community Health Study/Study of Latinos reveals regions of genetic associations with blood pressure traits". In: *PLoS One* 12.11 (2017).

[87] Graciela Ordoñez et al. "Genomewide admixture study in Mexican Mestizos with multiple sclerosis". In: *Clinical neurology and neurosurgery* 130 (2015), pp. 55–60.

[88] Steven J Mack et al. "High resolution HLA analysis reveals independent class I haplotypes and amino-acid motifs protective for multiple sclerosis". In: *Genes & Immunity* 20.4 (2019), pp. 308–326.

[89] Jorge R Oksenberg et al. "Mapping multiple sclerosis susceptibility to the HLA-DR locus in African Americans". In: *The American Journal of Human Genetics* 74.1 (2004), pp. 160–167.

[90] Donna Maglott et al. "Entrez Gene: gene-centered information at NCBI". In: *Nucleic acids research* 33.suppl_1 (2005), pp. D54–D58.

[91] Nikolaos Grigoriadis and Vincent Van Pesch. "A basic overview of multiple sclerosis immunopathology". In: *European journal of neurology* 22 (2015), pp. 3–13.

[92] Rachel E Ventura et al. "Hispanic Americans and African Americans with multiple sclerosis have more severe disease course than Caucasian Americans". In: *Multiple Sclerosis Journal* 23.11 (2017), pp. 1554–1557.

[93] Xavier Mariette and Lindsey A Criswell. "Primary Sjögren's syndrome". In: *New England Journal of Medicine* 378.10 (2018), pp. 931–939.

[94] Jisha J Nair and Tejas P Singh. "Sjogren's syndrome: review of the aetiology, pathophysiology & potential therapeutic interventions". In: *Journal of clinical and experimental dentistry* 9.4 (2017), e584.

[95] Ann Igoe and R Hal Scofield. "Autoimmunity and infection in Sjögren's syndrome". In: *Current opinion in rheumatology* 25.4 (2013), p. 480.

[96] Dimitris Karaiskos et al. "Stress, coping strategies and social support in patients with primary Sjögren's syndrome prior to disease onset: a retrospective case–control study". In: *Annals of the rheumatic diseases* 68.1 (2009), pp. 40–46.

[97] S Ferraro et al. "Air particulate matter exacerbates lung response on Sjögren's Syndrome animals". In: *Experimental and Toxicologic Pathology* 67.2 (2015), pp. 125–131.

[98] Bruce Freundlich et al. "A profile of symptomatic patients with silicone breast implants: a Sjögrens-like syndrome". In: *Seminars in arthritis and rheumatism*. Vol. 24. 1. Elsevier. 1994, pp. 44–53.

[99] Kimberly E Taylor et al. "Genome-wide association analysis reveals genetic heterogeneity of Sjögren's syndrome according to ancestry". In: *Arthritis & Rheumatology* 69.6 (2017), pp. 1294–1305.

[100] Christopher J Lessard et al. "Variants at multiple loci implicated in both innate and adaptive immune responses are associated with Sjögren's syndrome". In: *Nature genetics* 45.11 (2013), pp. 1284–1292.

[101] Yongzhe Li et al. "A genome-wide association study in Han Chinese identifies a susceptibility locus for primary Sjögren's syndrome at 7q11. 23". In: *Nature genetics* 45.11 (2013), pp. 1361–1365.

[102] Heng Yin et al. "Hypomethylation and overexpression of CD70 (TNFSF7) in CD4+ T cells of patients with primary Sjögren's syndrome". In: *Journal of dermatological science* 59.3 (2010), pp. 198–203.

[103] Xinhai Yu et al. "DNA hypermethylation leads to lower FOXP3 expression in CD4+ T cells of patients with primary Sjögren's syndrome." In: *Clinical immunology (Orlando, Fla.)* 148.2 (2013), p. 254.

[104] Michael B Cole et al. "Epigenetic signatures of salivary gland inflammation in Sjögren's syndrome". In: *Arthritis & Rheumatology* 68.12 (2016), pp. 2936–2944.

[105] Nicolas Gestermann et al. "Methylation profile of the promoter region of IRF5 in primary Sjögren's syndrome". In: *European cytokine network* 23.4 (2012), pp. 166–172.

[106] Yosra Thabet et al. "Epigenetic dysregulation in salivary glands from patients with primary Sjögren's syndrome may be ascribed to infiltrating B cells". In: *Journal of autoimmunity* 41 (2013), pp. 175–181.

[107] OD Konsta et al. "Defective DNA methylation in salivary gland epithelial acini from patients with Sjögren's syndrome is associated with SSB gene expression, anti-SSB/LA detection, and lymphocyte infiltration". In: *Journal of Autoimmunity* 68 (2016), pp. 30–38.

[108] Clio P Mavragani et al. "Defective regulation of L1 endogenous retroelements in primary Sjogren's syndrome and systemic lupus erythematosus: role of methylating enzymes". In: *Journal of autoimmunity* 88 (2018), pp. 75–82.

[109] Sergio González et al. "Alterations in type I hemidesmosome components suggestive of epigenetic control in the salivary glands of patients with Sjögren's syndrome". In: *Arthritis & Rheumatism* 63.4 (2011), pp. 1106–1115.

[110] Nezam Altorok et al. "Genome-wide DNA methylation patterns in naive CD4+ T cells from patients with primary Sjögren's syndrome". In: *Arthritis & rheumatology* 66.3 (2014), pp. 731–739.

[111] Juliana Imgenberg-Kreuz et al. "Genome-wide DNA methylation analysis in multiple tissues in primary Sjögren's syndrome reveals regulatory effects at interferon-induced genes". In: *Annals of the rheumatic diseases* 75.11 (2016), pp. 2029–2036.

[112] Corinne Miceli-Richard et al. "Overlap between differentially methylated DNA regions in blood B lymphocytes and genetic at-risk loci in primary Sjögren's syndrome". In: *Annals of the rheumatic diseases* 75.5 (2016), pp. 933–940.

[113] Joshua Millstein et al. "Disentangling molecular relationships with a causal inference test". In: *BMC genetics* 10.1 (2009), p. 23.

[114] Stefan H Stricker, Anna Köferle, and Stephan Beck. "From profiles to function in epigenomics". In: *Nature Reviews Genetics* 18.1 (2017), p. 51.

[115] Arundathi S Malladi et al. "Primary Sjögren's syndrome as a systemic disease: a study of participants enrolled in an international Sjögren's syndrome registry". In: *Arthritis care & research* 64.6 (2012), pp. 911–918.

[116] Caroline H Shiboski et al. "2016 American College of Rheumatology/European League Against Rheumatism classification criteria for primary Sjögren's syndrome: a consensus and data-driven methodology involving three international patient cohorts". In: *Arthritis & Rheumatology* 69.1 (2017), pp. 35–45.

[117] Martin J Aryee et al. "Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays". In: *Bioinformatics* 30.10 (2014), pp. 1363–1369.

[118] Timothy J Triche Jr et al. "Low-level processing of Illumina Infinium DNA methylation beadarrays". In: *Nucleic acids research* 41.7 (2013), e90–e90.

[119] Nizar Touleimat and Jörg Tost. "Complete pipeline for Infinium® Human Methylation 450K BeadChip data processing using subset quantile normalization for accurate DNA methylation estimation". In: *Epigenomics* 4.3 (2012), pp. 325–341.

[120] W Evan Johnson, Cheng Li, and Ariel Rabinovic. "Adjusting batch effects in microarray expression data using empirical Bayes methods". In: *Biostatistics* 8.1 (2007), pp. 118–127.

[121] Jeffrey T Leek et al. "The sva package for removing batch effects and other unwanted variation in high-throughput experiments". In: *Bioinformatics* 28.6 (2012), pp. 882–883.

[122] Christopher C Chang et al. "Second-generation PLINK: rising to the challenge of larger and richer datasets". In: *Gigascience* 4.1 (2015), s13742–015.

[123] Andrew E Jaffe et al. "Bump hunting to identify differentially methylated regions in epigenetic epidemiology studies". In: *International journal of epidemiology* 41.1 (2012), pp. 200–209.

[124] Peter A Jones. "Functions of DNA methylation: islands, start sites, gene bodies and beyond". In: *Nature Reviews Genetics* 13.7 (2012), pp. 484–492.

[125] Arthur Liberzon et al. "The molecular signatures database hallmark gene set collection". In: *Cell systems* 1.6 (2015), pp. 417–425.

[126] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the Royal statistical society: series B (Methodological)* 57.1 (1995), pp. 289–300.

[127] Guini Hong et al. "Separate enrichment analysis of pathways for up-and downregulated genes". In: *Journal of the Royal Society Interface* 11.92 (2014), p. 20130950.

[128] Joshua Millstein, Gary K Chen, and Carrie V Breton. "Cit: hypothesis testing software for mediation analysis in genomic applications". In: *Bioinformatics* 32.15 (2016), pp. 2364–2365.

[129] Joshua Millstein and Dmitri Volfson. "Computationally efficient permutation-based confidence interval estimation for tail-area FDR". In: *Frontiers in genetics* 4 (2013), p. 179.

[130] Trond Ove R Hjelmervik et al. "Gene expression profiling of minor salivary glands clearly distinguishes primary Sjögren's syndrome patients from healthy control subjects". In: *Arthritis & Rheumatism* 52.5 (2005), pp. 1534–1544.

[131] Deborah A Ferrington and Dale S Gregerson. "Immunoproteasomes: structure, function, and antigen presentation". In: *Progress in molecular biology and translational science*. Vol. 109. Elsevier, 2012, pp. 75–112.

[132] Joanna Perzyńska-Mazan, Maria Maślińska, and Robert Gasik. "Neurological manifestations of primary Sjögren's syndrome". In: *Reumatologia* 56.2 (2018), p. 99.

[133] Alicia K Smith et al. "Methylation quantitative trait loci (meQTLs) are consistently detected across ancestry, developmental stage, and tissue type". In: *BMC genomics* 15.1 (2014), p. 145.

[134] Lara Kular et al. "DNA methylation as a mediator of HLA-DRB1* 15: 01 and a protective variant in multiple sclerosis". In: *Nature communications* 9.1 (2018), pp. 1–15.

[135] Fusheng Zhou et al. "Epigenome-wide association data implicates DNA methylation-mediated genetic risk in psoriasis". In: *Clinical epigenetics* 8.1 (2016), p. 131.

[136] Yun Liu et al. "Epigenome-wide association data implicate DNA methylation as an intermediary of genetic risk in rheumatoid arthritis". In: *Nature biotechnology* 31.2 (2013), p. 142.

[137] Charlotte Ling and Tina Rönn. "Epigenetic adaptation to regular exercise in humans". In: *Drug discovery today* 19.7 (2014), pp. 1015–1018.

[138] Christiaan H Vinkers et al. "Traumatic stress and human DNA methylation: a critical review". In: *Epigenomics* 7.4 (2015), pp. 593–608.

[139] Steve Horvath. "DNA methylation age of human tissues and cell types". In: *Genome biology* 14.10 (2013), p. 3156.

[140] Shih-Kai Chu and Hsin-Chou Yang. "Interethnic DNA methylation difference and its implications in pharmacoepigenetics". In: *Epigenomics* 9.11 (2017), pp. 1437–1454.

[141] Peter A Jones, Jean-Pierre J Issa, and Stephen Baylin. "Targeting the cancer epigenome for therapy". In: *Nature Reviews Genetics* 17.10 (2016), p. 630.

[142] Aleksandra Majchrzak-Celińska and Wanda Baer-Dubowska. "Pharmacoepigenetics: Basic Principles for Personalized Medicine". In: *Pharmacoepigenetics*. Elsevier, 2019, pp. 101–112.

[143] Juliana Imgenberg-Kreuz, Johanna K Sandling, and Gunnel Nordmark. "Epigenetic alterations in primary Sjögren's syndrome–an overview". In: *Clinical Immunology* 196 (2018), pp. 12–20.

[144] Jessica R Tarn et al. "Symptom-based stratification of patients with primary Sjögren's syndrome: multi-dimensional characterisation of international observational cohorts and reanalyses of randomised clinical trials". In: *The Lancet Rheumatology* 1.2 (2019), e85–e94.

[145] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[146] Irina Higgins et al. "Early visual concept learning with unsupervised deep learning". In: *arXiv preprint arXiv:1606.05579* (2016).

[147] Joshua J Levy et al. "MethylNet: an automated and modular deep learning approach for DNA methylation analysis". In: *BMC bioinformatics* 21.1 (2020), pp. 1–15.

[148] Gregory P Way and Casey S Greene. "Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders". In: *BioRxiv* (2017), p. 174474.

[149] Joe H Ward Jr. "Hierarchical grouping to optimize an objective function". In: *Journal of the American statistical association* 58.301 (1963), pp. 236–244.

[150] Peter JA Cock et al. "Biopython: freely available Python tools for computational molecular biology and bioinformatics". In: *Bioinformatics* 25.11 (2009), pp. 1422–1423.

[151] Pablo Tamayo et al. "The limitations of simple gene set enrichment analysis assuming gene independence". In: *Statistical methods in medical research* 25.1 (2016), pp. 472–487.

[152] John M Sedivy, Gowrishankar Banumathy, and Peter D Adams. "Aging by epigenetics—a consequence of chromatin damage?" In: *Experimental cell research* 314.9 (2008), pp. 1909–1917.

[153] Ken WK Lee and Zdenka Pausova. "Cigarette smoking and DNA methylation". In: *Frontiers in genetics* 4 (2013), p. 132.

[154] Dolly Mahna, Sanjeev Puri, and Shweta Sharma. "DNA methylation signatures: biomarkers of drug and alcohol abuse". In: *Mutation Research/Reviews in Mutation Research* 777 (2018), pp. 19–28.

[155] He Li et al. "Identification of a Sjögren's syndrome susceptibility locus at OAS1 that influences isoform switching, protein expression, and responsiveness to type I interferons". In: *PLoS genetics* 13.6 (2017), e1006820.

[156] Juliana Imgenberg-Kreuz et al. "Genetics and epigenetics in primary Sjögren's syndrome". In: *Rheumatology* (2019).

[157] Warwick J Locke et al. "DNA methylation cancer biomarkers: translation to the clinic". In: *Frontiers in Genetics* 10 (2019).

[158] DL Morris et al. "MHC associations with clinical and autoantibody manifestations in European SLE". In: *Genes & Immunity* 15.4 (2014), pp. 210–217.

[159] Howard L. McLeod. "Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed". In: *Science* 339.6127 (2013), pp. 1563–1566. ISSN: 0036-8075. DOI: 10.1126/science.1234139. eprint: https://science.sciencemag.org/content/339/6127/1563.full.pdf. URL: https://science.sciencemag.org/content/339/6127/1563.

[160] Amrita Basu et al. "An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules". In: *Cell* 154.5 (2013), pp. 1151–1161.

[161] Jordi Barretina et al. "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity". In: *Nature* 483.7391 (2012), pp. 603–607.

[162] James C Costello et al. "A community effort to assess and improve drug sensitivity prediction algorithms". In: *Nature biotechnology* 32.12 (2014), p. 1202.

[163] Anneleen Daemen et al. "Modeling precision treatment of breast cancer". In: *Genome biology* 14.10 (2013), R110.

[164] Michael P Menden et al. "Machine learning prediction of cancer cell sensitivity to drugs based on genomic and chemical properties". In: *PLoS one* 8.4 (2013).

[165] Jane E Staunton et al. "Chemosensitivity prediction by transcriptional profiling". In: *Proceedings of the National Academy of Sciences* 98.19 (2001), pp. 10787–10792.

[166] David L Masica and Rachel Karchin. "Collections of simultaneously altered genes as biomarkers of cancer cell drug response". In: *Cancer research* 73.6 (2013), pp. 1699–1708.

[167] Han Yuan et al. "Multitask learning improves prediction of cancer drug sensitivity". In: *Scientific reports* 6 (2016), p. 31619.

[168] Su-In Lee et al. "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia". In: *Nature communications* 9.1 (2018), pp. 1–13.

[169] Luca Parca et al. "Modeling cancer drug response through drug-specific informative genes". In: *Scientific reports* 9.1 (2019), pp. 1–11.

[170] Xuewei Wang et al. "Predict drug sensitivity of cancer cells with pathway activity inference". In: *BMC medical genomics* 12.1 (2019), p. 15.

[171] Naiqian Zhang et al. "Predicting anticancer drug responses using a dual-layer integrated cell line-drug network model". In: *PLoS computational biology* 11.9 (2015).

[172] Woojoo Lee et al. "Sparse canonical covariance analysis for high-throughput data". In: *Statistical applications in genetics and molecular biology* 10.1 (2011).

[173] Nanne Aben et al. "TANDEM: a two-stage approach to maximize interpretability of drug response models based on multiple molecular data types". In: *Bioinformatics* 32.17 (2016), pp. i413–i420.

[174] Matthew G Rees et al. "Correlating chemical sensitivity and basal gene expression reveals mechanism of action". In: *Nature chemical biology* 12.2 (2016), p. 109.

[175] Hotelling Harold. "Relations between two sets of variates". In: *Biometrika* 28.3/4 (1936), pp. 321–377.

[176] Daniela M Witten, Robert Tibshirani, and Trevor Hastie. "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis". In: *Biostatistics* 10.3 (2009), pp. 515–534.

[177] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[178] Maryam Fazel, Haitham Hindi, and Stephen P Boyd. "A rank minimization heuristic with application to minimum order system approximation". In: *Proceedings of the 2001 American Control Conference.(Cat. No. 01CH37148)*. Vol. 6. IEEE. 2001, pp. 4734–4739.

[179] Qianxing Mo et al. "Pattern discovery and cancer gene identification in integrated cancer genomic data". In: *Proceedings of the National Academy of Sciences* 110.11 (2013), pp. 4245–4250.

[180] Jing Qu et al. "Kindlin-3 interacts with the ribosome and regulates c-Myc expression required for proliferation of chronic myeloid leukemia cells". In: *Scientific reports* 5 (2015), p. 18491.

[181] Carsten Riether et al. "CD70/CD27 signaling promotes blast stemness and is a viable therapeutic target in acute myeloid leukemia". In: *Journal of Experimental Medicine* 214.2 (2017), pp. 359–380.

[182] Long Zhang et al. "CD40 ligation reverses T cell tolerance in acute myeloid leukemia". In: *The Journal of clinical investigation* 123.5 (2013), pp. 1999–2010.

[183] Marina Konopleva et al. "Peroxisome proliferator-activated receptor γ and retinoid X receptor ligands are potent inducers of differentiation and apoptosis in leukemias". In: *Molecular Cancer Therapeutics* 3.10 (2004), pp. 1249–1262.

[184] Francesca Cottini et al. "Rescue of Hippo coactivator YAP1 triggers DNA damage–induced apoptosis in hematological cancers". In: *Nature medicine* 20.6 (2014), p. 599.

[185] Lee Ann Garrett-Sinha. "Review of Ets1 structure, function, and roles in immunity". In: *Cellular and molecular life sciences* 70.18 (2013), pp. 3375–3390.

[186] Douglas T Ross et al. "Systematic variation in gene expression patterns in human cancer cell lines". In: *Nature genetics* 24.3 (2000), pp. 227–235.

[187] Omid S Solari, James B Brown, and Peter J Bickel. "Sparse Canonical Correlation Analysis via Concave Minimization". In: *arXiv preprint arXiv:1909.07947* (2019).

[188] Arto Klami, Seppo Virtanen, and Samuel Kaski. "Bayesian exponential family projections for coupled data sources". In: *arXiv preprint arXiv:1203.3489* (2012).

[189] Galen Andrew et al. "Deep canonical correlation analysis". In: *International conference on machine learning*. 2013, pp. 1247–1255.

[190] Nicholas B Larson et al. "Kernel canonical correlation analysis for assessing gene–gene interactions and application to ovarian cancer". In: *European Journal of Human Genetics* 22.1 (2014), pp. 126–131.

[191] Roger Horton et al. "Gene map of the extended human MHC". In: *Nature Reviews Genetics* 5.12 (2004), pp. 889–899.

[192] Mandvi Bharadwaj et al. "Drug hypersensitivity and human leukocyte antigens of the major histocompatibility complex". In: *Annual review of pharmacology and toxicology* 52 (2012), pp. 401–431.

[193] Satoko Morishima et al. "Impact of highly conserved HLA haplotype on acute graft-versus-host disease". In: *Blood, The Journal of the American Society of Hematology* 115.23 (2010), pp. 4664–4670.

[194] Michelle MA Fernando et al. "Defining the role of the MHC in autoimmunity: a review and pooled analysis". In: *PLoS genetics* 4.4 (2008).

[195] Mary Carrington and Stephen J O'Brien. "The influence of HLA genotype on AIDS". In: *Annual review of medicine* 54.1 (2003), pp. 535–551.

[196] H Erlich. "HLA DNA typing: past, present, and future". In: *Tissue antigens* 80.1 (2012), pp. 1–11.

[197] Xiuwen Zheng et al. "HIBAG—HLA genotype imputation with attribute bagging". In: *The pharmacogenomics journal* 14.2 (2014), pp. 192–200.

[198] Alexander T Dilthey et al. "HLA* IMP—an integrated framework for imputing classical HLA alleles from SNP genotypes". In: *Bioinformatics* 27.7 (2011), pp. 968–972.

[199] Jason H Karnes et al. "Comparison of HLA allelic imputation programs". In: *PloS one* 12.2 (2017).

[200] Margaret M Madeleine et al. "Comprehensive analysis of HLA-A, HLA-B, HLA-C, HLA-DRB1, and HLA-DQB1 loci and squamous cell cervical cancer risk". In: *Cancer research* 68.9 (2008), pp. 3532–3539.

[201] J Yu Kelly et al. "Association of human leukocyte antigens with nasopharyngeal carcinoma in high-risk multiplex families in Taiwan". In: *Human immunology* 70.11 (2009), pp. 910–914.

[202] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[203] Haoyang Zeng et al. "Convolutional neural network architectures for predicting DNA–protein binding". In: *Bioinformatics* 32.12 (2016), pp. i121–i127.

[204] Xu Min et al. "Chromatin accessibility prediction via convolutional long short-term memory networks with k-mer embedding". In: *Bioinformatics* 33.14 (2017), pp. i92–i101.

[205] Peter K Koo and Sean R Eddy. "Representation learning of genomic sequence motifs with convolutional neural networks". In: *PLoS Computational Biology* 15.12 (2019).

[206] SS Rich et al. "Overview of the type I diabetes genetics consortium". In: *Genes & Immunity* 10.1 (2009), S1–S4.

[207] Yoav Goldberg. "Neural network methods for natural language processing". In: *Synthesis Lectures on Human Language Technologies* 10.1 (2017), pp. 1–309.

[208] Tomas Mikolov et al. "Distributed representations of words and phrases and their compositionality". In: *Advances in neural information processing systems*. 2013, pp. 3111–3119.

[209] Yoon Kim. "Convolutional neural networks for sentence classification". In: *arXiv preprint arXiv:1408.5882* (2014).

[210] Sergey Ioffe and Christian Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *arXiv preprint arXiv:1502.03167* (2015).

[211] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems*. 2012, pp. 1097–1105.

[212] Tariq Ahmad et al. "Haplotype-specific linkage disequilibrium patterns define the genetic topography of the human MHC". In: *Human molecular genetics* 12.6 (2003), pp. 647–656.

[213] Nitish Srivastava et al. "Dropout: a simple way to prevent neural networks from overfitting". In: *The journal of machine learning research* 15.1 (2014), pp. 1929–1958.

[214] James Bergstra and Yoshua Bengio. "Random search for hyper-parameter optimization". In: *Journal of machine learning research* 13.Feb (2012), pp. 281–305.

[215] Diederik P Kingma and Jimmy Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[216] Michael Cariaso and Greg Lennon. "SNPedia: a wiki supporting personal genome annotation, interpretation and analysis". In: *Nucleic acids research* 40.D1 (2012), pp. D1308–D1312.

[217] Juan-Manuel Anaya et al. *Autoimmunity: from bench to bedside*. El Rosario University Press, 2013.

[218] Adrienne Tin et al. "Genome-wide association study identified the human leukocyte antigen region as a novel locus for plasma beta-2 microglobulin". In: *Human genetics* 132.6 (2013), pp. 619–627.

[219] Marcelo A Fernandez Vina et al. "Tracking human migrations by the analysis of the distribution of HLA alleles, lineages and haplotypes in closed and open populations". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 367.1590 (2012), pp. 820–829.

[220] Mitchell J Machiela and Stephen J Chanock. "LDlink: a web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants". In: *Bioinformatics* 31.21 (2015), pp. 3555–3557.