Computational approaches for characterization and prioritization of human genetic variants

by

Qianqian Liang

BS, East China Universigy of Science and Technology, 2014

MS, Fudan University, 2017

Submitted to the Graduate Faculty of

the School of Public Health in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This dissertation was presented

by

Qianqian Liang

It was defended on

December 7, 2023

and approved by

Dennis Kostka, PhD, Associate Professor, Departments of Developmental Biology and Computational & Systems Biology, School of Medicine, University of Pittsburgh

Daniel E. Weeks, PhD, Professor, Departments of Human Genetics & Biostatistics, School of Public Health, University of Pittsburgh

Hyun Jung Park, PhD, Assistant Professor, Department of Human Genetics, School of Public Health, University of Pittsburgh

Cecilia W. Lo, PhD, Professor, Department of Developmental Biology, School of Medicine, University of Pittsburgh

Computational approaches for characterization and prioritization of human genetic variants

Qianqian Liang, PhD

University of Pittsburgh, 2024

Protein-coding and non-protein-coding genetic variants both play essential roles in contributing to human diseases. Therefore, better approaches for characterizing and prioritizing genetic variants can advance our understanding of the genetic causes of disease and contribute to the design of diagnostic and therapeutic strategies. In this dissertation, I explore both coding and non-coding genetic variants and report on new computational methods for their annotation and prioritization.

First, I developed a disease-specific approach for prioritizing non-coding variants. Integrating tissue-specific functional genomics data with non-coding disease-associated variants from the NHGRI-EBI GWAS catalog allowed me to design a model for disease-specific variant prioritization. This approach outperformed other variant-prioritization approaches, yielded interpretable and sensible associations between tissues and diseases, and enabled the calculation of disease similarities and the identification of biologically meaningful disease groups.

Next, I further improved this disease-specific approach by combining disease-associated variants across different disease terms, in order to enable information sharing. Through a systematic evaluation of all pairs of disease terms in the GWAS catalog, I discovered that combining variants from related diseases improved the performance of variant prioritization. Additionally, I found that suitable disease pairs for combination could be quickly identified using the disease similarity we derived previously.

Finally, I focused on a specific type of protein-coding variant that introduces a premature termination codon (PTC) and can lead to mRNA non-sense mediated decay (NMD). Since not all PTC-causing variants trigger NMD, I contributed to the development of a software tool called ”aenmd” that annotates whether such a variant is predicted to trigger NMD, or not (NMD escape). Applying aenmd to coding variants from the GWAS Catalog identified disease terms that were enriched with NMD-escaping and NMD-triggering variants, respectively.

Altogether, my thesis presents novel approaches for effectively characterizing and prioritizing protein-coding and non-protein-coding genetic variants in the context of human diseases. The tools I developed will contribute to improved annotation and understanding of genetic variants; they also can assist geneticists in the discovery of genetic factors contributing to human diseases, thereby ultimately facilitating the development of more efficacious diagnostic strategies and therapeutic interventions.

Table of Contents

List of Tables

List of Figures

Preface

I want to express my heartfelt gratitude to my PhD supervisor, Dr. Dennis Kostka. I wouldn't be having such a wonderful and fulfilling journey in my PhD life without him. He is always resourceful and brilliant in giving me very useful feedback on my research whenever I have a problem. Beyond the academic realm, Dennis is a very caring mentor who not only discusses the progress of the thesis with me but also gave me guidance in my personal and career development, time management and presentation skills. His passion for research has greatly influenced me and ignited my love for scientific research.

I would also like to extend my gratitude to my thesis committee members, Dr. Daniel Weeks, Dr. HJ Park, and Dr. Cecelia Lo. Dr. Weeks has extensive knowledge of statistical genetics and provided valuable assistance with the statistical and human genetics aspects of my research. Dr. Park is an expert in bioinformatics and machine learning and offered me valuable insights and feedback in that field. Dr. Lo has extensive expertise in the areas of developmental biology and congenital heart disease, and she has provided me with valuable feedback on the biomedical and developmental biology aspects of my work.

I would also like to thank all the professors, staff, and colleagues at the human genetics department. The wealth of knowledge I have acquired through the courses has been invaluable, and the department has truly become a second home to me. The annual retreats at Pymatuning have left indelible memories.

I would also like to thank all my lab members (both Dennis and Maria lab) for listening to my talks, giving me feedback on lab meetings, helping me solve technical questions, and mentally supporting me throughout all the time. I am so lucky to have such a wonderful group.

A special thank you to my friends who have been my companions throughout the PhD journey. Whether we were studying together, eating together, or sharing the happiness and challenges of my life, your camaraderie has been a source of joy and support.

Last but not least, I would like express my gratitude to my family, for believing me and supporting me all the time. I couldn't acheveve this without you.

1.0 Overall research goal and specific aims

Both protein-coding and non-protein-coding genetic variations play significant roles in the development of human diseases. Protein-coding variants can alter protein structures or functions, thereby contributing to disease progression [1, 2]. In contrast, non-coding variants primarily influence disease by regulating nearby gene expression, consequently affecting protein abundance [3, 4]. In this dissertation, I will explore coding and non-coding genetic variants contributing to human diseases.

In the first part, I will focus on non-coding variants. Although non-coding genetic variations are gaining importance for contributing to human diseases, studying which genetic variants contribute to which diseases remains challenging [5]. Therefore, there is a critical need for an accurate annotation and prioritization tool for non-coding variants. Existing tools are either designed for the entire organism (e.g., CADD [6]) or specific tissues (e.g., Genoskyline [7]). However, limited tools are designed for prioritizing genetic variants for specific diseases. Thus, the objective of our research is to develop a disease-specific variant prioritization method that enhances the accuracy of variant prioritization for specific diseases.

In the second part, I will delve into a specific type of coding variant that can trigger nonsense-mediated decay (NMD). Variants harboring premature termination codons (PTCs) can result in transcripts either undergoing NMD or evading NMD. The outcome, whether NMD is triggered or not, can have diverse implications for human diseases and is crucial in therapeutic interventions [8, 9, 10]. However, current NMD annotation tools have limitations, such as their inability to predict frameshift variants [11, 12]. Therefore, our research aims to create a computational tool for annotating NMD outcomes in genetic variants and utilizing this tool to analyze the outcome of NMD in complex genetic diseases.

To accomplish these goals, I have outlined the following specific aims:

Aim 1: Develop a disease-specific approach to improve the non-coding genetic variants prioritization.

Aim 2: Improve the disease-specific variant prioritization approach by incorporating single nucleotide variants (SNVs) associated with related diseases.

Aim 3: Construct a tool for nonsense-mediated decay (NMD) annotation for genetic variants and employ this tool to analyze the outcome of NMD in complex genetic diseases.

By successfully achieving these aims, we will have a more accurate tool for annotating and prioritizing non-coding genetic variants for specific diseases, as well as a robust NMD annotation tool that will deepen our understanding of NMD's role in genetic diseases. These advancements will empower researchers to utilize these tools in the investigation of coding and non-coding genetic variants, ultimately deepening our understanding of genetic variations contributing to human genetic diseases.

2.0 Introduction

Protein coding and non-protein coding genetic variants can both contribute to human genetic diseases. Protein-coding variants can lead to amino acid changes of the protein through missense mutations or lead to truncated/no protein through nonsense mutations [1]. These changes have the potential to alter the protein's structure or function, such as protein stability, protein-protein interactions or sub-cellular localization, which can contribute to the development of human diseases [2]. Non-protein-coding variants located in introns, upstream or downstream of coding regions, or intergenic regions, may also contribute to disease through regulatory mechanisms. For example, variants located in promoter or enhancer regions can impact nearby gene expression, while variants in 5' and 3' untranslated regions (UTRs) can affect mRNA stability, leading to Mendelian or complex diseases [13].

In my dissertation, I will explore coding and non-coding genetic variants related to human diseases. In the first part of the introduction, I will discuss the importance of prioritizing non-coding genetic variants in disease research and introduce various strategies that can be used to identify and prioritize these variants. This topic is directly related to the background in Chapters 3 and 4. Next, I will focus on a specific type of coding genetic variant that can lead to nonsense-mediated decay (NMD). I will discuss the role of NMD in human genetic diseases and the different existing tools for predicting NMD outcomes. This topic is directly related to the background in Chapter 5. By exploring these two distinct research areas, this dissertation aims to contribute to our understanding of the genetic factors that underlie human genetic diseases and provide insights into potential diagnostic and therapeutic strategies for these disorders.

2.1 Non-protein coding variant prioritization

2.1.1 Importance of non-coding variant prioritization

Non-protein coding genetic variants, which were once thought of as 'junk DNA', have gained increasing importance in the study of human genetic diseases [13, 4, 5, 14]. Recent genome-wide association studies have shown that more than 90% of single nucleotide variants (SNVs) associated with human diseases are located in non-coding regions (**Figure** 2.1). Furthermore, more studies into the regulatory role of non-coding DNA also shed light on the functional role of non-coding DNA in the development of human diseases [13, 15, 16].

Whole-genome sequencing (WGS) has become increasingly popular for studying non-coding genetic variants implicated in human diseases due to the decreasing cost of sequencing and its ability to analyze both common and rare, and de novo non-coding genetic variants [17]. However, whole genome sequencing (WGS) poses significant challenges, as it can detect millions of genetic variants per genome and most of the current variant prioritization procedures such as traditional genome-wide association tests or variant annotation tools like VEP or Ensembl are often insufficient in filtering these variants, making it difficult to identify causal variants for a disease of interest. Therefore, a better way to characterize and prioritize functional genetic variants potentially associated or causal to a disease is increasingly needed in the field of human genetics research.

Recent advancements in computational methods that employ machine learning strategies to integrate various functional and genetic datasets have shown promise in prioritizing non-coding genetic variants. Those methods combine various datasets and provide a unified score (we will later refer to it as **variant score**) that quantify the functional or pathogenic effect of a variant that could potentially lead to a phenotype or disease. Although research in this area is ongoing, studies have demonstrated the usefulness of these methods in identifying causal variants in human diseases. For example, CADD has been instrumental in prioritizing variants for WGS in autism spectrum disorder [6, 18]. Furthermore, these methods are scalable and can handle the growing functional and annotation datasets.

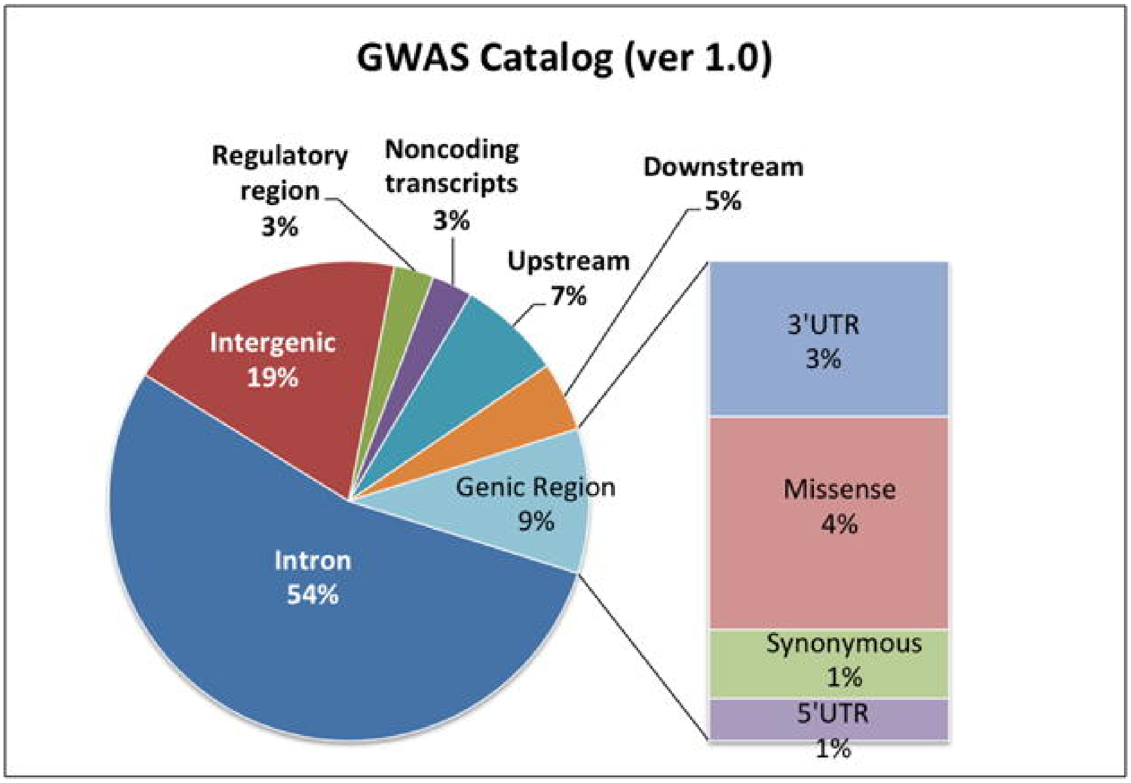

Figure 2.1: *Different categories of genetic variants in complex diseases.* Reproduced from Lee et.al. [5] (Figure licensed for re-use by Springers Nature and Copyright Clearance Center).

2.1.2 Datasets used in non-coding variant prioritization

2.1.2.1 Comparative genomics data

It has been commonly believed genetic mutations contributing to human diseases that can affect fitness consequences are typically subject to negative selection [19, 20]. As a result, these pathogenic alleles are gradually eliminated over time, and the genetic regions in which they occur tend to be conserved. It is predicted that 3% to 5% of the human genome is conserved between vertebrates and other species [21]. More than half of the highly conserved elements are located outside of protein-coding gene in the human genome [21].

Researchers have developed several methods to find conservation among different species, such as GERP [20], Phastcons [19], Siphy [22], and PhyloP [23]. These methods detect potential functional variants by identifying nucleotide substitution rates that deviate from neutral drift. For example, GERP and Phastcons can detect regions with a slower substitution rate compared with neutrally evolving regions [20, 19]. Siphy captures the sequence with the change of mutation rate and also uncovers characteristics of substitution patterns underlying natural selection [22]. PhyloP, implemented with four statistical and phylogenetic tests, can detect both faster and slower substitution rates compared with neutral drift and also in a clade-specific manner [23].

The conservation score can serve as a valuable tool for prioritizing functional genetic variants due to its ability to indicate pathogenicity. However, relying solely on this score can present certain drawbacks. One limitation is that the conservation score does not consider the specific functional context of genetic variants. Moreover, certain genetic variants associated with complex diseases, like cardiovascular diseases, may experience weak evolutionary selection, making it challenging for evolutionary scores to accurately predict their impact [24].

2.1.2.2 Functional genomics data

In addition to conservation scores, functional genomic data can also be used to generate variant scores. Non-coding genetic variants exert their function to cause human genetic diseases usually by their regulatory roles; therefore, variants that are located in functional elements can have the potential to cause human diseases. Functional genomics data catalog the functional elements in the genome, which include non-coding RNA and regions that show reproducible biochemical signatures such as protein binding or chromatin structure [25]. It is estimated that around 80% of the human genome has at least one biochemical activity in at least one cell type [25]. This is a much larger region than the conserved regions in the genome and a larger and more complicated dataset to analyze.

The Encyclopedia of DNA Elements (ENCODE) and the Roadmap Epigenomics Program are two large-scale databases that systematically provide us with functional genomics data; the ENCODE aims to catalog all the functional elements in the human genome and the Roadmap aims to investigate epigenetic modifications of the human genome [25, 15]. Other projects such as Fantom5 catalog promoters and enhancers using Cap Analysis of Gene Expression (CAGE) technology and GTEx provide us tissue-specific eQTLs that regulate expression [26, 27].

Some commonly used functional genomics data are histone modification markers (e.g. H3K4me3, H3K27ac, H3K27me3), DNA methylation, protein binding regions, open chromatin regions, chromatin 3D interactions, etc. Those markers are indicative of regulatory regions in the genome. For instance, H3K4me3 is thought to be related to promoters, H3K27ac is related to active enhancers and promoters, and H3K27me3 is associated with repressed regions [28]. Open chromatin regions are also commonly used as indicative of cis-regulatory elements [29].

Compared with conservation data, functional genomics data is context-specific, where the context can be a specific tissue or cell type, a specific time point, or one person or a group of people. This gives us the flexibility to use the data to accomplish a specific role. But it also comes with great challenges, such as how to integrate the large amount of data and how to find out the disease-relevant context.

2.1.2.3 Other annotation data

There are some other annotation data that are commonly included to prioritize non-coding variants. For example, ***minor allele frequency (MAF)*** can be used as an annotation, as the alleles that are pathogenic to human are tend to eliminated through history and are kept in low allele frequency. Therefore, allele frequency (AF) can be an annotation for variants indicating pathogenicity. ***GC content*** (the GC percentage over a window (e.g., 75kb) of a specific variant) is an informative annotation, as GC content has been found to be associated with mutation rate [30], chromatin accessibility [31], and DNA methylation [32]. ***Distance to nearest TSS*** is also a commonly used annotation because pathogenic variants tend to reside within a certain distance (e.g. 1kb) of a transcription start site (TSS) [33].

2.1.3 Methods used for non-coding variant prioritization

Typically, researchers use *supervised* or *unsupervised* machine learning methods to combine the datasets above to generate unified variant scores; those score summarizes the datasets and predicts the variant's function or deleteriousness. For supervised machine learning methods, such as CADD [6] and GWAVA [33], researchers use a set of benign variants and a set of pathogenic variants. Then, they build a model to utilize the annotations above to best distinguish the benign and pathogenic variants. Therefore, supervised methods are sensitive to the labeling data. On the contrary, unsupervised machine learning methods, such as Eigen [24] or GenoCanyon [34], do not use any labeled data to build the model; therefore, it is generally acknowledged that they are less biased and not sensitive to the different sets of the labeling data. The advancements achieved by unsupervised methods are usually modest, leaving room for improvements [5].

Current variant prioritization methods mainly fall into two categories: organism-level variant scores and tissue-specific variant scores. Organism-level scores integrate genomic features across multiple tissues/cell types into one score and predict the functional effect of variants for an organism overall; tissue-specific scores, on the contrary, retain some tissue-specificity from functional genomic data and predict the functional impact of variants in a specific tissue/cell type. Both two types of methods are disease-agnostic. Some recent studies generate variant scores for specific diseases (disease-specific scores), either for only one or a small group of diseases or for many diseases spanning various disease categories.

2.1.3.1 Organism-level variant scores

Organism-level variant scores, using either supervised or unsupervised machine learning approach, summarize genomic features across all tissues/cell types and generate one score that summarizes the functional or pathogenicity of the variant for the whole organism.

Some of them use a supervised machine learning approach. For instance, CADD (Combined Annotation Dependent Depletion) is among one of the earliest variant scores that combine various functional genomics data and conservation scores that derive a unified variant score [6]. CADD uses human-derived variants as benign variants and simulated de novo mutations as possibly deleterious variants to train the model, and it applies a linear support vector machine to train the model. The key highlight for CADD is that it does not use human-curated labeled pathogenic variants from databases such as Clinvar or Human Gene Mutation Database (HGMD) but uses labeled variants based on evolutionary selection; therefore, it has a much larger number of training variants. GWAVA (genome-wide annotation of variants) was trained using three random forest algorithms and predicts non-coding genetic variants [33]. It uses variants from the HGMD as pathogenic and matched variants from 1000 genome projects (1KG) as benign control variants. GWAVA uses various annotation resources, including histone modifications, open chromatin, conservation scores, and other annotations such as CG context, allele frequency, and distance to the nearest TSS.

There are some unsupervised machine learning methods. GenoCanyon used an EM-based algorithm that combines conservation scores and biochemical signals into a single score [34]. Eigen assumes variants have two unknown groups, function and non-functional, and blockwise conditional independence among different functional annotations [24]. Using the correlated structure among different annotations, Eigen generates a score using a linear weighted combination of the annotations. LINSIGHT employs a probabilistic model to analyze evolutionary data across various species and within human trajectory, thereby enabling the prediction of the fitness consequences of genetic variants [35]. Additionally, LINSIGHT also combines the probabilistic model with a generalized linear model, so it integrates functional genomic datasets into the analysis.

Variant scores	Model	Model type	Coding or Non-coding	Annotations type	Training data
CADD [6]	Linear support vector machine	Supervised	Both	CONS, FUNC	proxy-neutral and proxy-deleterious SNVs
eigen [24]	Unsupervised spectral approach	Unsupervised	Both	CONS, FUNC, OTHR	Variants from 1000 genomes
GenoCanyon [34]	A statistical framework	Unsupervised	Non-coding	CONS, FUNC	the GWAS Catalog and surrounding SNVs
GWAVA [33]	Random Forest algorithm	Supervised	Non-coding	CONS, FUNC, OTHR	HGMD variants and matched control variants
LINSIGHT [35]	Probabilistic model and a generalized linear model	Unsupervised	Non-coding	CONS, FUNC, OTHR	Polymorphism data from 54 unrelated individuals and divergence data from UCSC

CONS: comparative genomics data; FUNC: functional genomics data; OTHR: other annotation data

Table 2.1: *Examples of organism-level variant scores.*

2.1.4 Tissue-specific variant scores

Organism-level scores predict the variants' function for the whole organism; however, these methods do not provide information about which tissues those mutations have the most impact. Therefore, some researchers generate tissue-specific scores that can indicate the function of variants for specific tissues/cell types.

GenoSkyline is an example of a tissue-specific score, an extension of the GenoCanyon

[7, 36]. Similar to GenoCanyon, it assumes SNVs are a mixture of two groups (functional vs. nonfunctional) and fits a statistical framework to calculate the posterior probability of a SNV being functional given genome annotations. However, Genoskyline uses a broader set of epigenomic datasets (including eight different Chip-seq calls, such as H3K4me3 and H3K9ac and DNA methylation data), and it generates 127 scores for each variant that can indicate the tissue/cell type specificity. The Genoskyline score predicts the DNA functionality for 127 tissues/cell types.

Another example of a tissue-specific score is Fitcons2 [37]. Fitcons2 uses two types of genomic features that contain four context-specific epigenomic annotations (e.g., DNase-seq) and five context-agnostic annotations (e.g., Transcription Factor binding site across cell types). Fitcons2 uses a probabilistic evolutionary model to fit human and non-human sequence data and gradually builds a decision tree using the genomic features that best fit the evolutionary model. Finally, variants will be mapped to clusters using the decision tree and the Fitcons2 score, given by ρ, which indicates the probability that the mutation can have fitness consequences based on the cluster in which the variant resides. Fitcons2 score uses 115 different cell type-specific data, and it can predict variant deleteriousness in 115 different contexts.

Variant scores	Model	Model type	Coding or Non-coding	Annotations type	Training data
Genoskyline [7]	A statistical framework derived from GenoCanyon [34]	Unsupervised	Non-coding	CONS, FUNC	the GWAS Catalog and surrounding SNVs
Fitcons2 [37]	A probabilistic evolutionary model and a decisin tree	Unsupervised	Both	CONS, FUNC, OTHR	Population genomics data from 69 genomes and comparative genomics data from UCSC

CONS: comparative genomics data; FUNC: functional genomics data; OTHR: other annotation data

Table 2.2: *Examples of tissue-specific variant scores.*

2.1.5 Disease-specific variant scores

Tissue-specific variant scores can predict a variant's function in specific tissues. However, it is typically unclear which cell line/tissue combinations are best to distinguish disease risk variants from benign ones for a given disease. Therefore, a few studies have developed variant scores that can work on specific diseases.

Some disease-specific scores can predict variants for one or a small group of diseases. For example, heartENN uses a convolutional neural network (CNN) and a broad range of heart-related epigenetic datasets to predict the changes in the molecular effect of genetic mutations for congenital heart disease [38]. Yousefian-Jazi et al. generated variant scores for amyotrophic lateral sclerosis (ALS) using a CNN model on 2525 functional genomic features and ALS GWAS dataset [39]. Yousefian-Jazi et al. developed a model that can prioritize variants for 21 autoimmune diseases [40]. They trained a random forest algorithm on 2026 functional features using SNVs associated with 21 autoimmune diseases from HGMD and clinvar variants. Those methods can predict genetic variants for one or a few diseases, and they depend on hand-curated genomic features or GWAS datasets for a specific disease, which is laborious and hard to apply to a broader range of diseases.

Some disease-specific scores cover a broader range of diseases. One computational framework, ARVIN (Annotation of Regulatory Variants using Integrated Networks), built disease-specific gene regulatory networks (GRN) and used features derived from the disease-specific GRN and other genomic features to train a random foretest classifier [41]. Another disease-specific score PINES, used epigenomic data from Roadmap and ENCODE and scored a variant based on comparing the input variant with the background variants (variants with no disease-relevance) [42]. To get disease-specific PINES score, users can either input disease-associated SNVs from GWAS to calculate tissue enrichments or specify the disease-relevant tissues; in either way, PINES will upweight disease-relevant annotations and output disease-specific variant scores.

DIVAN (DIsease-specific Variant ANnotation) can also score disease-specific variants [43]. DIVAN used an ensemble framework and trained 45 disease-specific models using epigenomic and genomic annotations and SNVs associated with 45 diseases. DIVAN improved performance in distinguishing disease-associated SNVs compared with other disease-agnostic methods such as CADD or GWAVA.

2.1.6 Information sharing across different diseases

Disease-specific variant scores can prioritize genetic variants for specific diseases. However, it is important to note that diseases do not exist independently but are correlated. Researchers have also revealed that different diseases can share key information, such as genetic background, molecular mechanisms, or disease-associated gene expression patterns.

For example, immune diseases can share information. Cotsapas et al. analyzed 107 immune disease-associated SNVs associated with seven immune system diseases. Surprisingly, nearly half of these SNVs were associated with more than one disease [44]. They calculated the distances between SNVs and grouped them into four distinct clusters to better understand these relationships. Within each cluster, they examined the interactive proteins. They found that the proteins within each cluster interacted with each other and were uniquely expressed in immune-related cell types compared to other cell types. This suggests that diseases can share similar mechanisms and involve similar tissues. Similarly, Li et al. analyzed 28 shared genetic variants in pediatric autoimmune diseases (pAID)[45]. They found that pAID-associated genes (genes significantly associated with those shared genetic variants) are highly expressed in immune-related cell types compared with non-immune-related cell types.

Psychiatric disorders can also share information. For example, Wingo et al. found shared genetic background and shared causal proteins between psychiatric and neurodegenerative diseases [46]. Those shared proteins are more expressed in brain-related cell types and tissues. In another study [47], they also found shared SNVs among pairs of psychiatric disorders and found that brain tissues have enriched expression levels in those diseases. However, no SNVs are shared in more than two psychiatric disorders.

There is also evidence suggesting shared information between type 2 diabetes and Alzheimer's disease. For example, this study suggests shared genetic variants and shared causal pathways between type 2 diabetes and Alzheimer's disease [48]. But they didn't mention shared tissues between them. Another study suggests shared mechanisms and signaling pathways in disease etiology between type 2 diabetes and Alzheimer's disease [49].

2.1.7 Disease-category specific variant score

Some researchers developed disease-specific variant scores for a disease category. For example, eyeVarP [50] is a computational tool that can prioritize genetic variants for eye diseases (HPO ontology, HP:0000478), which can include many child disease terms such as glaucoma, coloboma, corneal disease, etc. In another study, researchers developed a non-coding variant framework to prioritize genetic variants for 19 autoimmune diseases by using pathogenic non-coding variants from autoimmune diseases and immune cell-related epigenetic features [51]. Another group developed a disease category-specific variant prioritization method called CASAVA [52]. In this paper, they developed a supervised method to score variants in 24 broad disease categories, such as cardiovascular disease, while each disease category contains 137 to 8065 disease-associated risk variants.

Overall, in this section, I introduced the background of different types of variant scores. These are related to my work in my Aim 1 and 2. I will discuss rationales for Aim 1 and 2 in **Section** 2.3.1 and 2.3.2.

2.2 Nonsense-mediated mRNA decay (NMD) in genetic diseases

2.2.1 Nonsense-mediated mRNA decay

Nonsense-mediated mRNA decay (NMD) is a cellular quality control process that degrades mRNAs that contain premature termination codons (PTCs) or other types of abnormal translation termination signals [53, 54]. PTCs can be introduced by single nucleotide variants, insertions, deletions, or splice site mutations that cause one of the stop codons (UAG, UGA, or UAA) to occur earlier than the canonical stop codon [8]. NMD is a process that is thought to remove truncated proteins that can potentially have some deleterious effect [53].

The mechanism of NMD in mammalian cells is usually referred to as the 'EJC model' [53].

An EJC (exon junction complex) is a multi-protein complex that resides on the junctions of two exons during the pre-mRNA splicing process, and it includes proteins such as the core protein eIF4A-III (eukaryotic initiation factor 4A-III) and additional proteins Magoh and Y14 [55]. In the event of the protein translation, the ribosome moves through the mRNA to the stop codon and removes the EJC along the way. However, in the case of an aberrant mRNA containing a premature termination codon (PTC), the ribosome stops at the PTC before the normal stop codon, and the EJC after the PTC cannot be removed. The retained EJC will act as a signal to initiate the NMD degradation process [54].

2.2.2 Rules for nonsense-mediated mRNA decay escape

NMD can cause mRNAs containing PTCs to degrade; however, many transcripts bearing PTCs can escape the NMD process [8]. For instance, the NMDective tool predicts that roughly 51% of PTC variants may escape nonsense-mediated decay to some degree [12]. Based on the EJC model and research on somatic and cancer cells, researchers have developed a set of rules to determine the NMD escapes based on the location of the PTC within the transcript [8, 12]. Those rules are usually categorized as ”canonical” rules and ”non-canonical” rules (illustrated in **Figure** 2.2):

- Canonical rules
  - Last exon rule: the PTC that is located in the last exon
  - 50-nt rule: the PTC that is located in the last 50 nt of the penultimate exon
- Non-canonical rules
  - Start proximal rule: the PTC located within 150 bp of the start of the first exon.
  - Long exon rule: the PTC located within an exon longer than 407 bp.
  - Single exon rule: the transcript has only one exon.

Canonical rules can be mostly explained by the molecular mechanism of the ’EJC model.’ PTCs in the last exon can escape from NMD as the EJC has been removed through the translation process. Similarly, PTCs in the last 50 nt of the penultimate exon can also lead to the EJC being removed because of the elongation/footprint of the ribosome [56]. The canonical NMD rules have been largely tested and validated by many large-scale human genomic studies [8]. For example, using the paired human genome and transcriptome data from the GTex, researchers revealed that PTC variants located in regions following the ”last-exon” and ”50 nt” rules had the lowest rate of allelic imbalance, indicating they escape NMD [57]. While these canonical rules can predict some NMD escapes, researchers found exceptions that deviate from the rules [58, 59, 57]. This suggests additional rules may allow some mRNAs to evade NMD.

Non-canonical rules were later proposed by using large-scale studies of human cancer genomes [60]. The start proximal rule describes that NMD efficiency is greatly reduced for transcripts bearing PTC in the first 150 nt. This could probably be explained by translation re-initiation, where the ribosome does not cycle back after translation termination but instead keeps scanning and starts at the downstream start codon [61, 62]. Cancer data also suggest long-exon rule [60], which may be explained by the hypothesis that the EJC needs to be in physical contact with the PTC to initiate the NMD process [8, 63]. Some researchers also found some evidence supporting the single exon rule by studying intronless genes such as human histone H4, mouse heat shock protein 70, and human melanocortin 4-receptor gene [64, 65].

Overall, according to one large cancer genome study, it is suggested that canonical rules can explain roughly 50% of the variance in NMD efficiency, while non-canonical rules can explain 25% [60].

2.2.3 NMD and human genetic diseases

In general, PTC-containing transcripts that undergo NMD lack protein production, leading to loss-of-function outcomes. Conversely, PTC-containing transcripts that evade NMD may lead to the production of truncated proteins, which can exhibit a range of functional outcomes, including loss of function, partial function (hypomorphic), dominant negative (antimorphic), or gain of function (neomorphic) [8].

When a mRNA escapes NMD, the resulting production of a truncated protein can have different biochemical effects. These effects can either exacerbate or ameliorate the disease. For example, an upstream PTC in the beta-globin gene can be seen by NMD, therefore

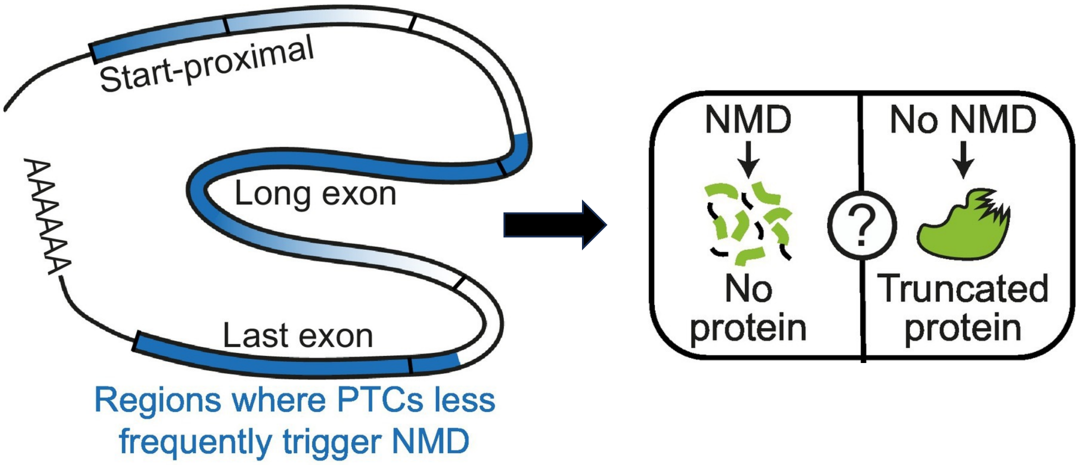

Figure 2.2: *NMD escape rules.* Reproduced from Supek et al. [8]. (Figure licensed for re-use by Elsevier and Copyright Clearance Center).

leading to a recessive form of β-thalassemia; on the contrary, if the PTC locates at the 3' end of the gene, it can escape from NMD, leading to a truncated beta-globin protein, which precipitates in toxic inclusion bodies, resulting in a dominant form of the disease. A contradicting example is Duchenne muscular dystrophy (DMD). The PTC mutations in the 3' of the dystrophin gene can lead to the milder DMD phenotype because the truncated protein can retain partial function. In contrast, the PTC mutations in the upstream of the gene result in a severe phenotype due to the loss of expression of the protein with partial function (reviewed in [8, 9, 10]).

NMD can have varying effects on human diseases; however, the overall impact of NMD on these diseases remains unclear. To delve deeper into this topic, Lindeboom et al. conducted a study analyzing the impact of NMD in evolutionary selection and pathogenic variants in different disease genes. They found that 52% of rare PTC variants can trigger NMD compared with 25% of the common PTC variants [12]. This suggests that rare variants, which usually lead to a more severe phenotype due to their elimination through purifying selection, exhibit a higher abundance of NMD trigger variants. Additionally, they found

49 disease genes that exhibited over a two-fold enrichment of predicted NMD-evading PTCs and 155 disease genes that displayed over a two-fold enrichment of predicted NMD-triggering PTCs [12]. This suggests that in more disease genes, NMD can aggravate the phenotype. Overall, these findings challenge the traditional belief that NMD protects individuals from truncated proteins, as the overall impact of NMD can exacerbate human genetic diseases, although this trend can vary across different disease genes.

2.2.4 Existing tools that can predict NMD for genetic diseases

There are several existing tools that can annotate NMD escaping for PTC variants.

An Ensembl VEP plugin can predict whether a PTC variant can escape from NMD. It uses four rules, including the last exon rule, 50nt-rule, start proximal rule, and single exon rule. VEP can annotate PTC variants that introduce stop gain, but it cannot annotate frameshift variants where the stop codon is located downstream of the variant. In addition, VEP can annotate NMD escape outcome, but it does not output which rule it used to predict the escape.

Another tool is NMDetective [12]. It trains a random forest model on 2840 PTC introducing mutations where the mRNA level of the PTC-bearing transcripts and the corresponding wild-type transcripts are measured. This model can predict the efficacy of the NMD for all PTC introducing single nucleotide variants, giving a score between 0 and 1, where 0 means complete NMD escape and 1 means NMD triggering. NMDetective can predict single nucleotide variants; however, it cannot predict insertions or deletions that cause PTCs.

There are some other tools that predict NMD escape. NMDEscPredictor can predict frameshift indel variants, but it considers only the canonical rules to make the prediction. [66]. ALoFT can predict the outcome of the PTC variant into benign, dominant, and recessive variants, and it can also predict NMD escape. At the same time, it does not specify the rules of the NMD escape prediction [67]. SNPEff is a variant annotator that can predict NMD escape but only considers two canonical rules (last exon and 50-nt rules) [68].

2.3 Rationales for specific aims

2.3.1 Aim 1: Develop an interpretable general framework for disease-specific variant prioritization

In-silico non-coding variant prioritization methods can predict variants' functional outcomes and are playing an increasingly important role in studying the genetic causes of human diseases [13, 4, 5, 14]. Current variant prioritization methods mainly fall into two categories: organism-level variant scores, such as CADD [6], which assess the functional impact of variants on the entire organism, and tissue-specific variant scores, like Genoskyline [7], which evaluate the functional consequences of variants within specific tissues or cell types. However, these scores are not disease-specific and do not provide insights into the functional impact of variants within the context of a particular disease. Therefore, it is crucial to develop variant prioritization methods tailored to specific diseases.

Disease-specific variant prioritization methods do exist. However, some of them are limited to a single disease (e.g. congenital heart disease [38]) or a small group of diseases (e.g. autoimmune diseases [40]). While other methods can prioritize variants for a broader range of diseases, they have their weaknesses such as requiring prior knowledge of tissues relevant to a disease [41] or using complex machine learning models (e.g., ensemble decision trees) that are difficult to interpret [43].

In this aim, we propose a simple logistic regression approach for converting tissue/cell-type specific variant scores into disease-specific scores. This not only allows for improved prioritization of disease-specific variants but also enables the calculation of disease similarities based on disease-relevant tissues and cell types. Through this method, we aim to demonstrate that a disease-specific approach outperforms current organism-level and tissue-specific approaches. Furthermore, by analyzing the tissue weights derived from the model, we expect to gain insights into the relationships between diseases and identify relevant tissues for specific diseases. Overall, our aim is to provide a straightforward and interpretable approach to disease-specific variant prioritization, which can enhance our understanding of the genetic basis of human diseases and improve the identification of disease-contributing non-coding variants.

2.3.2 Aim2: Combine SNVs in related diseases to improve disease-specific variant prioritization

In Aim 1, we developed a disease-specific approach to prioritize genetic variants in 111 diseases. While this method outperforms existing organism and tissue-specific variant scores, it still exhibits a moderate level of performance, leaving room for improvement. This could be partially due to the limited training samples available for each disease. Therefore, a method to increase the training dataset to improve disease-specific variant prioritization is needed.

The current method CASAVA increases the training dataset by aggregating SNVs within the same disease categories [52]. However, CASAVA did not assess the effectiveness of this grouping approach in comparison to considering individual diseases separately. Furthermore, CASAVA did not evaluate a more suitable metric for grouping related diseases, whether it be a genetic correlation, semantic similarity, or other disease-relatedness metric.

Thus, in Aim 2, we propose an information-sharing method to improve disease-specific variant prioritization by combining non-coding SNVs among disease terms. We will combine SNVs from two diseases in this aim, with the potential to extend this methodology to encompass more than two diseases. We will utilize it for all possible disease pairs in the GWAS Catalog. This approach allows us to systematically evaluate whether including SNVs in related diseases can lead to improved variant prioritization. In addition, it also allows us to find a better metric to group diseases by comparing three different disease similarity metrics. Overall, we anticipate that combining SNVs across disease terms will improve disease-specific variant prioritization.

2.3.3 Aim 3: Explore nonsense-mediated mRNA decay (NMD) in genetic diseases

Nonsense-mediated mRNA decay (NMD) is a cellular process responsible for degrading mRNAs that contain premature termination codons (PTCs) [53, 54]. PTC-bearing variants can either undergo NMD or escape from it, and they represent a significant proportion of pathogenic genomic variations with clinical relevance [12, 60]. However, the escape or retention from NMD can yield diverse biochemical consequences in the context of human genetic diseases. For example, escaping from NMD can worsen beta-thalassemia but alleviate the phenotype of Duchenne muscular dystrophy (DMD) [8, 9, 10]. Therefore, the accurate annotation of NMD escape outcomes for PTC-bearing variants is crucial for investigating the impact of these variants on human genetic diseases.

Existing tools for predicting NMD have certain limitations. For instance, VEP can only predict stop gain variants but not frameshift variants that lead to downstream stop codons [11]. Similarly, NMDetective can predict single nucleotide variants but not insertions or deletions [12]. Other tools, such as NMDEscPredictor and SNPEff, consider only a subset of NMD rules when making predictions [66, 68].

Aim 3 focuses on developing an NMD prediction tool that can accurately annotate transcript-variant pairs containing PTCs for predicting escape from NMD. This tool will incorporate functions that are not currently available in other existing methods. It will be based on established and experimentally validated rules for NMD escape. Furthermore, we will utilize this tool to annotate variants from the GWAS catalog and investigate the impact of NMD on human complex diseases by analyzing enrichment patterns. Through this analysis, we aim to enhance our understanding of the consequences of PTC-containing variants in human diseases.

2.4 Public health relevance

Genetic factors play an important role in human diseases, such as cancer, immune system diseases, mental or behavioral disorders, cardiovascular diseases, etc. [69]. Studying the genetic factors underlying human diseases can help us understand the molecular basis of diseases and, therefore, can help us provide early diagnosis and screening, disease treatment, personalized medicine, and gene therapy. For example, an accurate diagnosis of primary immunodeficiency disorders by genetic testing can find out which genes are disrupted for individuals and, therefore, inform decisions on targeted therapeutic options [70]. For another example, understanding the genetic causes of autism spectrum disorder can help us develop therapeutic targets that rescue the haploinsufficiency of gene expression [71]. The research work conducted in Chapter 3 and 4 can be used to prioritize and annotate non-coding genetic variants in genetic studies such as whole-genome sequencing studies. This can help us elucidate the genetic underpinning of human diseases, and ultimately improve healthcare by early diagnosis or precise treatment.

The research work conducted in Chapter 5 where we studied NMD in human genetic diseases can shed light on medical therapeutic strategies for diseases caused by PTC. For some patients, stimulating NMD could be beneficial, while for other patients inhibiting NMD is a potential therapeutic strategy [8]. Therefore, understanding the NMD effect on diseases is crucial in designing therapeutic strategies. For example, the stop codon read-through strategy can enable mRNA to evade NMD, and this approach has been used in the treatment of cystic fibrosis and Duchenne muscular dystrophy [72, 73, 74, 75]. Conversely, in a contrasting example, as NMD can suppress the production of toxic truncated globin protein, stimulating NMD may be potentially beneficial for beta-thalassemia therapy [76].

3.0 Disease-specific analysis improves prioritization of non-coding genetic variants

The following chapter was adapted from the manuscript Qianqian Liang, Abin Abraham, John A Capra, and Dennis Kostka, ”Disease-specific prioritization of non-coding GWAS variants based on chromatin accessibility” where I am the first author [77]. Minor revisions were made to address feedback from the thesis committee. Please see **Section** 3.6 for author contributions.

3.1 Abstract

Non-protein-coding genetic variants are a major driver of the genetic risk for human disease; however, identifying which non-coding variants contribute to diseases and their mechanisms remains challenging. In-silico variant prioritization methods quantify a variant’s severity, but for most methods the specific phenotype and disease-context of the prediction remain poorly defined. For example, many commonly used methods provide a single, organism-wide score for each variant, while other methods summarize a variant’s impact in certain tissues and/or cell-types. Here we propose a complementary disease-specific variant prioritization scheme, which is motivated by the observation that variants contributing to disease often operate through specific biological mechanisms.

We combine tissue/cell-type specific variant scores (e.g., GenoSkyline, Fit- Cons2, DNA accessibility) into disease-specific scores with a logistic regression approach and apply it to 25,000 non-coding variants spanning 111 diseases. We show that this disease-specific aggregation significantly improves association of common non-coding genetic variants with disease (average precision: 0.151, baseline=0.09), compared with organism-wide scores (GenoCanyon, LINSIGHT, GWAVA, eigen, CADD; average precision: 0.129, base- line=0.09). Further on, disease similarities based on data-driven aggregation weights highlight meaningful disease groups, and it provides information about tissues and cell-types that drive these

similarities. We also show that so-learned similarities are complementary to genetic similarities as quantified by genetic correlation. Overall, our approach demonstrates the strengths of disease-specific variant prioritization, leads to improvement in non-coding variant prioritization, and it enables interpretable models that link variants to disease via specific tissues and/or cell-types.

3.2 Introduction

Characterizing non-coding genetic variants in the human genome is essential for making progress toward better understanding the genetic components of the disease. Protein-coding sequence accounts for only about two percent of human DNA, and 90% of disease-associated variants discovered by genome-wide association studies (GWAS) are located in non-protein-coding regions [78]. Furthermore, whole-genome sequencing (WGS) discovers disease-associated variants genome-wide [79, 80] and is increasingly becoming an assay of choice. Therefore, approaches for characterizing and prioritizing non-coding variants can be expected to play an increasingly important role, especially when assessing discovered variants in the context of further functional follow-up experimental studies.

Efforts to (computationally) characterize and better understand non-coding variants take advantage of sequence, functional genomics, comparative genomics, and epigenomics data [81, 15, 16], amongst others. These data are combined and used to train and develop supervised and/or unsupervised models that attempt to quantify a variant's impact [5]. We find it conceptually useful to distinguish between variant scores that model overall impact (that is on the level of the whole organism, *organism-level* scores) and scores that quantify impact in a specific context, like a tissue or a cell-type (i.e., *tissue-level* scores). Examples for organism-level scores are CADD [6], Eigen [24], or LINSIGHT [35], while scores from methods like GenoSkyline [7], Fitcons2 [37], or FUN-LDA [82] are tissue-specific.

Often interest in a variant is from the perspective of studying a specific disease. In that case, organism-level scores are likely to be overly general. That is, a variant's impact might be considered high because it disrupts the functional role of a sequence element. However, that functional role may be unrelated to the disease of interest. In one study, for instance, organism-level scores like CADD and DANN were unable to discover an enrichment signal for brain-related traits, while context-specific variant scores focusing on relevant tissues were successful [83]. That is, tissue-specific scores can address the issue of disease specificity to some extent. However, aspects of disease-relevant tissues typically remain unknown, and often more than one tissue is implicated with a specific trait [84]. This suggests the use of *disease-specific* variant scores that characterize variants in the context of a specific disease phenotype of interest.

Computational methods for disease-specific variant prioritization do exist. Some approaches are geared towards one disease (e.g, congenital heart disease [38], amyotrophic lateral sclerosis [39]) or towards a specific class of diseases (e.g., autoimmune diseases [40]). This focus prevents them from being readily adapted to other disease types. Others, like DIVAN [43], PINES [42], and ARVIN [41], cover a broader range of disease types. Of these, ARVIN requires a priori knowledge of disease-relevant tissues, whereas DIVAN and PINES do not. PINES uses an enrichment-based method to predict and up-weight disease-relevant tissues/cell-types, whereas DIVAN uses a more complex machine learning algorithm. The PINES approach is evaluated on a relatively small set of traits ($\sim$10 different contexts), while DIVAN's more complex model renders understanding the relationship between different tissues and diseases difficult.

In this work, we derive disease-specific variant scores combining published tissue-specific scores. We use a carefully regularized logistic regression approach to derive data-driven disease-specific combination weights, which in-turn allow us to assess the similarity between different disease phenotypes. Using the NHGRI-EBI GWAS catalog [78] we compiled a benchmark dataset containing about 63k phenotype-associated non-protein-coding single nucleotide variants across 111 disease phenotypes (together with matched random controls). We then demonstrate that using disease-specific combination weights outperforms conventional organism-level approaches, that our interpretable model has competitive performance, and that it enables a disease similarity measure that captures information complementary to established measures like genetic correlation.

3.3 Methods

3.3.1 Data sources and processing

3.3.1.1 Disease-associated variants

Disease-associated non-coding single nucleotide variants were retrieved from the NHGRI-EBI Catalog of human genome-wide association studies database (GWAS catalog, version `2020-12-02`, downloaded from `https://www.ebi.ac.uk/gwas/docs/file-downloads`, no additional p-value threshold was imposed). These data contained 122,396 unique non-coding SNVs spanning 2,782 phenotypes, where non-coding was defined as variants not overlapping protein-coding sequence (GENCODEv36); we also excluded variants annotated as protein coding sequence variants (e.g. missense variants, frameshift variants) as a SNV's "functional class" in the GWAS Catalog. Variants in the GWAS Catalog are annotated with phenotypes using the Experimental Factor Ontology (EFO, `https://www.ebi.ac.uk/efo`) [85]. We then focused on variants with phenotype terms annotated in disease domain of the EFO (i.e., all terms/traits/phenotypes we consider are descendants of the term "disease" (EFO:0000408, EFO version `3.24.0`, accessed `2020-11-17`). Further on, we restricted ourselves to phenotypes with at least 100 annotated non-coding SNVs. This yields 121 diseases and 31,103 SNVs. Next, SNVs in the HLA region, and SNVs with minor allele frequency (MAF) less than 1% in the European population as reported by the International Genome Sample Resource were excluded (as they cannot be matched to control SNVs with the SNPsnap approach, see below). Out of 31103 SNVs, a total of 5225 SNVs were removed. Finally, we further removed diseases that had fewer than 100 SNVs after filtering in the previous step. As a result, we were left with 111 diseases and 25,561 SNVs, totaling 77,028 phenotype-SNV associations, since one SNV can be annotated to more than one phenotype. **Appendix Data** A.3.1 and A.3.2 contain 111 phenotypes and 77,028 phenotype-associated SNVs we used in this study. We also grouped SNVs in LD blocks (SNPsnap, $r^2 \geq 0.5$) and identify SNVs with the minimum p-value per block("representative SNV"); we provide this information, which we use in some of the analyses described below, in **Appendix Data** A.3.2.

3.3.1.2 Control variants

To adjust for the non-random distribution of disease-associated SNVs across the genome [33], we generated matched control variants. For each disease-associated SNV we generated matched control non-coding variants with MAF ≥1% using four different strategies, where the non-coding is again defined discussed above (**Section** 3.3.1.1). The four strategies are:

Random For each disease-associated SNV, we selected ten SNVs from common non-coding variants in 1000G EUR at random (i.e., equal probability for all SNVs) as controls.

TSS-matching We processed common non-coding SNVs and selected a subset of these variants as controls, where the distribution of distances to the nearest protein-coding gene's transcription start site (TSS) are matched between control set and disease-associated SNVs (similar to GWAVA, [33]). Specifically, we sorted all common non-coding SNVs by the distance to the nearest TSS and divided them into 50 bins, where each bin contains the same number of SNVs. Then, for each disease-associated SNV, we randomly selected ten control SNVs from the bin containing the disease-associated SNV's distance to the nearest gene.

SNPsnap-matching Using SNPsnap (an online tool that can select the matched SNVs for given SNVs using specific criteria) [86], we matched control SNVs to disease-associated variants in terms of minor allele frequency, gene density (r^2 cutoff `ld0.8`), distance to the nearest gene TSS, and number of SNVs in LD. Our parameters for maximum allowable deviation of the input SNVs were: 5%, 50%, 20% and 50%, respectively. We randomly selected ten control SNVs per disease-associated SNV from SNPsnap's results, and we ensured there are no duplicated control SNVs for different disease-associated SNVs. If there were less than 10 control SNVs returned by SNPsnap, we kept all of the control SNVs. If no control SNVs were matched, we removed the disease-associated SNVs (a total of 311 SNVs) from our analyses.

SNPsnap-TSS-matching Essentially the same as in SNPsnap-matching, but controlling **only** for the distance to the nearest genes (maximum allowable deviation: 20%); for three other attributes "maximum allowable deviation" is set to 10,000%. We note that in both SNPsnap-matching and SNPsnap-TSS-matching, distance is measured by distance to the nearest gene, whereas for TSS-matching only protein-coding genes are considered.

In all four matching strategies we excluded variants annotated in the GWAS catalog as control SNVs. One control variant can only be matched to one disease-associated SNV. In our research, we chose SNPsnap-matching for our main results, but we have compared the different performance of organism level scores using the four different matching strategies (See **Appendix Methods** A.1.1 and **Appendix Figure** A.1- A.2). And we provided the four sets of control variants in **Appendix Data** A.3.3.

3.3.1.3 Additional data sources, variant scores

We used pre-computed SNV annotations from the following sources:

- CADD v.1.3: `http://krishna.gs.washington.edu/download/CADD/v1.3/1000G_phase3.tsv.gz`
- EigenPC v.1.1:`https://xioniti01.u.hpc.mssm.edu/v1.1`
- Fitcons2: `http://compgen.cshl.edu/fitCons2/hg19`
- GenoCanyon: `http://genocanyon.med.yale.edu/GenoCanyon_Downloads.html`
- GenoSkylinePlus: `http://genocanyon.med.yale.edu/GenoSkylineFiles/GenoSkylinePlus/GenoSkylinePlus_bed.tar.gz`
- GWAVA v.1.0: `ftp://ftp.sanger.ac.uk/pub/resources/software/gwava/v1.0/VEP_plugin/gwava_scores.bed.gz`
- LINSIGHT: `http://compgen.cshl.edu/%7Eyihuang/tracks/LINSIGHT.bw`
- DIVAN: `https://sites.google.com/site/emorydivan`
- ncER: `https://github.com/TelentiLab/ncER_datasets`
- DHS accessibility: We downloaded Avocado-imputed [87] DNase1 hypersensitive sites (DHS) signal for 127 ENCODE biological contexts (tissues / cell types) from `https://noble.gs.washington.edu/proj/avocado/data/avocado_full/DNase/`.

3.3.2 Tissue-weighted variant prioritization based on tissue-specific score (e.g., DNase1 hypersensitivity)

3.3.2.1 A penalized logistic regression model for context-weighted score averaging

For each disease D_p, where $m \in \{1, 2, ..., 111\}$, we built a lasso logistic regression model. For predicting SNV's associations with a disease term, we consider SNVs as observations, and each SNV is described as a vector $\mathbf{x} \in \mathbb{R}^d$ of variant scores in d tissues/contexts; we arrange vectors $\{\mathbf{x}^i\}_{i=1}^n$ for n observations (n SNVs) in a matrix $X \in \mathbb{R}^{n \times d}$, together with a vector y of n binary entries, indicating for each SNV association with a specific disease term (no=0/yes=1). In addition, we denote the average score (across contexts) for a SNV i by $\bar{x}^i$, which is also a basline score because it aggregates across contexts.

We use a logistic regression model of the form

$$\log \frac{p_i}{1 - p_i} = \alpha \bar{x}^i + \boldsymbol{\beta}' \mathbf{x}^i \quad \text{s.t.} \quad \alpha \geq 0 \tag{1}$$

where $\alpha \in \mathbb{R}_+$ and $\boldsymbol{\beta} \in \mathbb{R}^d$ are regression coefficients, and p_i is the probability that SNV i is associated with a disease that is studied. We fit a regularized version of the negative log likelihood

$$\underset{\alpha_0, \alpha, \boldsymbol{\beta}}{\arg\min} \; -\frac{1}{n} \sum_{i=1}^{n} \left[\log(1 - p_i) + y_i \log \frac{p_i}{1 - p_i} \right] + \lambda ||\boldsymbol{\beta}||_2^2 / 2 \tag{2}$$

where the dependence on $\alpha, \boldsymbol{\beta}$ of the first term is through Equation (1). For large regularization parameters λ this will yield small $\boldsymbol{\beta} \to \mathbf{0}$ and recover the baseline ($\bar{\boldsymbol{x}}$) of unweighted averaging of context scores (scaled by a non-negative factor α). We implemented this approach using the R package glmnet (version 2.0-18, [88]) and determined the regularization parameter via a nested 5-fold cross validation (cv.glmnet function) through maximizing the area under the (cross-validated) ROC curve. In the nested 5-fold cross validation, we used the inner loop to select the regularization parameter λ, and used the selected λ to train and test the model in the outer loop. Class weights were employed to balance skewed class sizes. The method described in this section is referred to as ***tissue-weighted*** in the paper.

3.3.2.2 Disease similarities from context-weighted score averaging

Context-weighted score averaging, as described above, results in disease-specific coefficient vectors ($\{\boldsymbol{\beta}^{(p)}\}$, with p indexing disease terms), together with bootstrap estimates for the standard deviation of each coefficient (that can be arranged in corresponding vectors $\{\boldsymbol{\gamma}^{(p)}\}$). Specifically, we use 5-fold cross-validation repeated 10 times, yielding 50 coefficient vectors for each disease. We use their mean for our estimate of $\boldsymbol{\beta}^{(p)}$, and their standard deviation as an estimate of $\boldsymbol{\gamma}^{(p)}$.

For a pair of diseases (d_p, d_q) we then define a disease similarity through similarity of associated coefficient vectors $\boldsymbol{\beta}^{(p)}$ and $\boldsymbol{\beta}^{(q)}$, taking into account our estimates of coefficient variability. Specifically, we fit a weighted linear regression model (i.e., regressing $\boldsymbol{\beta}^{(p)}$ on $\boldsymbol{\beta}^{(q)}$), with regression weights taking into account coefficient variability as follows:

$$w_k^{(p,q)} = 1 \Big/ \sqrt{s_k^p s_k^q} \quad \text{and} \quad s_k^\circ = \alpha\gamma_k^{(\circ)} + (1-\alpha)m \quad \text{for} \quad \circ \in \{p, q\},$$

where we chose m to be the 25% quantile of all (estimated) standard deviations observed, and $\alpha = 3/4$. Therefore, s_k^p and s_k^q are shrunken versions of the standard deviations for the regression coefficients of disease p and disease q in tissue/context k, respectively. Finally, for disease pairs with a positive coefficient from the weighted linear regression we take the coefficient of determination (R^2) as a similarity measure; for disease pairs with a negative coefficient, we take $-R^2$. We note that for constant regression weights $\{w_k^{(p,q)}\}$ this is equal to the Pearson correlation between the coefficient vectors we obtain from context-weighted score averaging (i.e., $\mathrm{cor}(\boldsymbol{\beta}^{(p)}, \boldsymbol{\beta}^{(q)})$).

3.3.3 Variant prioritization performance

3.3.3.1 Tissue-weighted cross-validation performance

To measure the cross-validation performance of Tissue-weighted, we use repeated cross-validation [89] to reduce the variance (due to the random partitioning of data into 5 folds). Here, we repeated 5-fold cross-validation 30 times, and record the performance of each repeat. We later use the mean performance of the 30 repeats as the performance of that method and we also show the variance in figures such as Figure 3.4.

3.3.3.2 Comparing organism level scores

For each disease we have disease-associated and control SNVs, and corresponding precomputed organism-level scores. With this setup we calculate performance metrics of interest (area under the receiver operator characteristic curve (AUROC) and average precision (AUPR)), and obtain disease-specific performance metrics for each scoring approach. To compare performance between organism-level scores on the same disease we use performance measures computed on 30 bootstrap samples (each bootstrap sample randomly contains 90% of disease and control variants) and then employ the Wilcoxon signed-ranks test to test to assess differences in performance. This yields p-values as reported in Appendix Data A.3.4.

With respect to aggregating comparisons across diseases, we note that disease terms can (and do) share SNVs, so performance metrics in different terms are not necessarily independent. Also, disease terms can vary substantially in the number of annotated SNVs. We again use Wilcoxon singed-ranks test [90] on performance metrics (computed using all disease-associated- and control-SNVs for each disease term) to compare two organism-level variant scores aggregate across diseases. This approach yields p-values, as reported in Appendix Data A.3.5.

3.3.3.3 Comparing tissue-weighted scores

Tissue-weighted baseline scores (see above) are calculated in the same way as organism-level scores. For tissue-weighted scores with data-driven tissue-specific weighting (see above), we use cross-validated performance measure for each bootstrap sample and the same 30 bootstrap samples as when we compared between organism-level scores. And then we use the same Wilcoxon signed-ranks tests to measure the difference. For comparing scores aggregated across diseases we again proceed analogous to organism-level scores and use a Wilcoxon singed-ranks test on cross-validated disease-specific performance measures. Results are summarized in Appendix Data A.3.8 and A.3.9.

3.3.3.4 Comparing organism-level and tissue-weighted scores

For comparisons between organism-level and tissue-combined scores we again use a bootstrap approach: for a specific disease term we use the Wilcoxon signed-ranks tests as discussed above to compare performance measures from organism-level scores with tissue-weighted scores. We note that this approach does not take into account: ***(a)*** Variability in the organism-level scores originating form variability of the data they are derived from, and ***(b)*** The possibility that organism-level scores may have already used SNVs in their score derivation process, and we use them again for evaluation in their score derivation process. However, we don't expect these issue to substantially confound or results, and we note that incurred bias in our comparisons would expected to be in favor of organism-level scores. Results are summarized in Appendix Data A.3.6, A.3.7, A.3.10 and A.3.11.

3.3.3.5 DIVAN performance assessment and comparison.

To assess and compare our performance with DIVAN [43], we generated a test set of SNVs from the GWAS catalog that were *i)* added after DIVAN had been published (i.e., after 05/28/2016) and *ii)* not present in the database used to train DIVAN (Association Result Browser `https://www.ncbi.nlm.nih.gov/projects/gapplus/sgap_plus.htm`) and *iii)* not within 1kb distance around SNVs used to train DIVAN and *iv)* were annotated to a disease phenotype addressed by DIVAN.

Control SNVs were generated using SNPsnap matching, as described above. To be able to satisfy criterion *iv)*, we mapped our disease terms (EFO terms) to disease terms used by DIVAN (MeSH terms) using the EMBL-EBI Ontology Xref Service (OxO, `https://www.ebi.ac.uk/spot/oxo/`, retrieved on April 19, 2020) and were able to resolve 41 out of 45 terms (Appendix A.3.12). Of these, we keep terms with 20 or more disease associated SNVs in the test set and 50 or more SNVs in a training set that we also constructed (see below), yielding 29 overall disease phenotypes we used in our analysis. In order to fairly compare DIVAN with our logistic regression approach we constructed a training set using disease-associated SNVs from the GWAS catalog and the Phenotype-Genotype Integrator (PheGenI, `https://www.ncbi.nlm.nih.gov/gap/phegeni`) [91], excluding SNVs in the test dataset describe above, or SNVs within 1kb around test SNVs. Appendix Data A.3.13 summarizes test and training data used for this analysis. Results are summarized in Appendix Data A.3.14.

In addition, we also measured the difference in DHS-weighted variant score between disease-associated variants and matched controls in DIVAN test set. The delta value was used to quantify this difference. We found that overall, disease-associated variants had higher variant scores than controls in the DIVAN test set. In particular, for diseases such as Crohn's disease, the mean delta value can be as high as 0.1 in variant score. For more detailed information, please refer to Appendix Figure A.7.

3.3.3.6 Performance assessment using chromosome hold-out

To assess the performance of our DHS tissue-weighted score we also used a chromosome hold-out strategy, with test SNVs on different chromosomes from training data. Specifically, for each disease, we choose a set of chromosomes that contains approximately 20% SNVs with a 1/10 positive to negative ratio (the same as the cross-validation setting) as a test set. Selection of test chromosomes is performed for each disease term separately, as disease-associated SNVs differ. To automate the procedure, we deployed (binary) linear programming to pick out chromosomes in test set for each disease.

Specifically, for each disease term we solve the optimization problem

$$\begin{aligned}&\operatorname{argmax}_{\{x_i\}_{i=1}^{22}} \sum\nolimits_{i=1}^{22} c_i x_i\\ &\text{subject to } \sum\nolimits_{i=1}^{22} w_i^+ x_i \leq 0.2 \text{ and } x_i \in \{0,1\},\end{aligned}$$

where $\{x_i\}$ are binary indicator variables whether a chromosome is included in the test/hold-out set; w_i^- and w_i^+ are the fraction of disease-associated (w_i^+) and control SNVs (w_i^-) on chromosome i and weights in the objective function are defined as $c_i = w_i^+ - |w_i^+ - w_i^-|$. This approach selects, for each disease term, a set of chromosomes to hold out that contain about 20% of disease-associates SNVs and that approximately reflects the overall imbalance between disease-associated and control SNVs. **Appendix Figure** A.18 and A.19 contain performance evaluations on chromosome hold-out sets.

3.3.3.7 Performance assessment using one SNV per LD block

To assess the effect of SNV correlation on our results we also performed analyses using only a single representative SNV per LD block (defined by $r^2 \geq 0.5$, see **Section** 3.3.1.1). Results are shown in **Appendix Figure** A.20 and A.21.

3.3.4 Comparison with genetic correlation

We retrieved genetic correlation values of disease pairs from the GWAS atlas [92]. To be able to use these data we mapped EFO disease terms (used in the NIH-NCBI GWAS Catalog and in our study) to terms used in the GWAS atlas study. To do so, we extracted synonyms of each EFO term (as listed on EFO ontology) and compared each synonym to the "trait" and "uniqtrait" column in the GWAS atlas data. All matches (with one tolerated letter substitution) were used.

In this approache a single EFO term can map to multiple GWAS atlas traits and studies. To estimate the genetic correlation between two EFO terms (say d_i and d_j), we use a weighted combination of genetic correlation values:

$$r_g(d_p, d_q) = \sum_{l,m} w_{lm} r_g(s(d_p)_l, s(d_q)_m)$$

where $r_g(\cdot, \cdot)$ is the genetic correlation of two diseases, $\{s(d_p)\}_{p=1}^{r}$ and $\{s(d_q)\}_{q=1}^{s}$ are the GWAS atlas studies that are mapped to EFO term d_p and d_q, respectively; w_{lm} is a weight for each combination of the GWAS atlas studies accounting for the sample sizes of different studies used to estimate genetic correlation values. We choose

$$w_{lm} = \tilde{w}(s(d_p)_l) \cdot \tilde{w}(s(d_q)_m)$$

where

$$\tilde{w}(s(d_p)_l) = \text{size}(s(d_p)_l) / \sum_k \text{size}(s(d_p)_k)$$

where "size" denotes the sample size of a study. This schene puts higher weights on studies with large sample sizes and smaller weights to studies with smaller sample sizes.

3.3.5 Notes about epimap comparison, cluster annotation and display

3.3.5.1 Epimap trait-tissue association for table 3.5

We obtained the latest SNP-centric GWAS enrichments table from the EpiMap Repository at `http://compbio.mit.edu/epimap/`. We retrieve tissues with adjusted p-values for each disease. We map the tissue names used in our study (Standard Roadmap Epigenomes, as labed by EID) to tissue names used in epimap (biosamples, as labeld as BSS biosample id) by adapting the scripts from `https://github.com/cboix/EPIMAP_ANALYSIS/blob/master/metadata_scripts/get_roadmap_mapping.R`. If there are more than one biosamples tissues mapped to roadmap tissues, we reported the p value of the tissue with the most significant result.

3.3.5.2 Cluster names in table 3.6

To name each cluster/group of diseases/EFO terms we choose the EFO term that contains most of the cluster/group members. In Appendix Data A.3.21 we summarize the terms with high term frequency in each cluster, where term frequency is the fraction of *descendant* terms present. For example, the EFO term "immune system disease" (EFO:0000540) has a term frequency of 0.588 in the "immune-1 cluster"; this means that 58.8% of EFO terms in that cluster are descendants of EFO:0000540. We exclude the terms that are overly broad such as the term "disease" or "experimental factor ontology". For each cluster, we rank the cluster member EFO terms using term frequency and select as name a meaningful term with the high term frequency. For one cluster where no term had high frequency we chose the name "heterogeneous".

We also show a diagrams of EFO disease term relationships in each cluster in **Appendix Figure** A.11- A.17. Occasionally we include ancestor EFO terms not present in the cluster in a diagram, which are marked by asterisks.

3.3.5.3 Dimension reduction and coefficient heatmap

UMAP plot The two-dimensional UMAP plot of 111 EFO disease terms in **Figure** 3.7 is based on disease similarities based on context-weighted score averaging (see section 3.3.2.2). The `umap` function of the `uwot` `R` package was used with parameters `n_neighbors = 15`, `ret_model = TRUE`, `PCA_center = FALSE`.

Coefficient heatmap The heatmap in **Figure** 3.8 displays coefficient vectors of models for disease association (see section 3.3.2.1), normalized for each disease. Specifically, for each disease and tissue coefficient x_i

$$\tilde{x}_i = \begin{cases} (x_i - x_{\min})/x_{95} & x_i \leq x_{95} \\ 1 & x_i > x_{95} \end{cases}$$

where $x_{\min}$ is the minimum coefficient for a disease, and x_{95} is the 95% quantile.

Cluster-associated tissues For each cluster, we show the top-five tissues that are most associated with the cluster (**Figure** 3.8). To identify these tissues we conduct a two-sample Wilcoxon test (one-sided) on every tissue, where we compare normalized tissue coefficients for this cluster to the the other with the highest coefficients on average. The five tissues with the smallest p-value are then selected as top-five tissues.

Tissue-associated clusters For the heatmap with all tissues in **Appendix Figure** A.10, we assigned a cluster to each tissue. For each tissue, we calculated the median (across disease terms of a cluster) of the normalized coefficients for all clusters; the cluster with the highest median was assigned.

3.4 Results

3.4.1 Non-coding GWAS variants associated with disease phenotypes, and matched controls

In order to study variant prioritization methods, we created a dataset of “positive” (i.e., disease associated) non-coding variants, matched with a random set of “control” variants.

This setup allowed us to quantitatively assess prioritization methods based on their performance in discriminating positive from control variants.

3.4.1.1 Disease-associated non-coding SNVs

We used a subset of single nucleotide variants (SNVs) reported in the EBI/NIH GWAS catalog [78] to compile an inventory of disease-associated non-coding variants. Specifically, we focused in reported variants that *(a)* do not overlap protein-coding sequence (see **Methods**) and *(b)* that are associated with a disease phenotype as noted in the Experimental Factor Ontology (EFO) trait description, which is provided within the catalog. We define disease phenotypes as descendants of the EFO term "disease" (EFO:0000408). Focusing on disease terms with at least 100 annotated SNVs resulted in 26,080 associations involving 20,656 SNVs and 67 disease phenotypes. The EFO provides parent-child relations between disease terms (parent = more general, child = more specific), and propagating SNVs from child-terms to parent-terms increased the number of disease phenotypes with at least 100 SNVs, resulting in 77,028 association between 25,516 SNVs and 111 diseases. We find that most of the SNVs we recover are located in intronic (60.5%) and intergenic (25.8%) sequence (**Figure** 3.1**A**), and that a majority of SNVs are directly annotated to a single disease phenotype (**Figure** 3.1**B**). After propagating annotated SNVs from child to parent terms, SNV-to-disease annotations become predominantly many:many (**Figure** 3.1**B**). **Appendix Data** A.3.1 lists disease terms and corresponding numbers of disease-associated SNVs.

3.4.1.2 Control SNVs

For each disease-associated SNV we selected ∼10 matched control-SNVs using a re-implementation of the SNPsnap approach [86], while avoiding duplicate control-SNV across the overall dataset (see **Methods**). This yielded 255,137 control SNVs (for some disease associated SNVs we could not retrieve the full ten control SNVs); with this data we have access to data for 111 disease terms, containing disease-associated SNVs together with matched controls. **Appendix Data** A.3.2 and A.3.3 describe all SNVs used in this study.

3.4.2 Disease-specific non-coding variant prioritization with organism-level variant scores is moderately successful

We assessed how well current commonly-used organism-level variant scores are able to prioritize disease-associated vs. control-SNVs for the 111 disease terms we studied. **Figure** 3.2 summarizes results, where boxplots of two performance measures (area under the ROC curve and average precision (= area under the precision recall curve)) are shown for CADD [6], eigen [24], GenoCanyon [7], GWAVA [33], and LINSIGHT [35] scores. We find that organism-level scores, while improving upon random guessing, are only moderately successful in correctly prioritizing disease-associated non-coding variants. Comparing variant scores with each other we find that relative performance differences appear overall robust with respect to the metric employed (area under the ROC curve vs. average precision). It is qualitatively visible that CADD performs less favorably than other methods, but also that there are differences between the other methods. We therefore compared performance between different scores in more detail.

We studied the performance of different scores at two levels of resolution: In aggregate across all disease terms, and for each disease term separately. For both approaches we used Wilcoxon signed-ranks tests to decide whether one score significantly outperforms another score (significant p-value) or whether performance is tied (non-significant p-value); see **Methods** section, **Table** 3.1. We find that GenoCanyon has better performance than other variant scores, followed by LINSIGHT, GWAVA and eigen, while CADD is consistently outperformed by other methods. Performance differences between LINSIGHT, GWAVA and eigen are not significant when aggregating across disease terms (last three columns in **Table** 3.1); however, when counting individual disease LINSIGHT has most wins and fewest losses, while eigen has most losses and fewest wins. **Appendix Data** A.3.4 and A.3.5 contain results for all comparisons. Overall these quantitative results are in-line with the visual impression from **Figure** 3.2. Next, we investigated if the performance of organism-level variant scores could be improved by using tissue-specific scoring approaches.

3.4.3 Tissue-specific scores improve disease-specific variant prioritization

3.4.3.1 Disease-specific aggregation weights for tissue-specific variant scores

We studied three tissue-specific scores for variant prioritization to explore if their usage can improve on the performance of organism-level scores. Specifically, we used Genoskyline [7] and Fitcons2 [37] as scores specifically designed to prioritize variants, and DNase I hypersensitivity (DHS) profiles from the ENCODE project [16]. All of these scores are available for 127 contexts [15] spanning a diverse set of cell and tissue types, including heart, brain, immune cells, and more.

For each tissue-specific score we assess two approaches to prioritize variants. First, as a baseline approach we aggregate scores across tissues in a *disease-agnostic* way. That is, for a specific variant we average scores at the variant position across all tissues (termed ***tissue-mean***), essentially producing a organism-level type score, independent of the disease term under consideration. Second, we aggregate scores across tissues in a *disease-specific* way. Briefly, we train a regularized logistic regression model for each disease term that learns disease-specific tissue aggregation weights. In a nested cross-validation setup, learned weights are applied to held-out variants, allowing for a fair performance assessment of this approach (termed ***tissue-weighted***), see **Methods**. **Figure** 3.3 summarizes our findings.

In **Figure** 3.3**A** we show tissue-mean performance (as measured by average precision) for the three scores we study on the left, and tissue-weighted performance on the right. For all three scores tissue-weighted significantly outperforms tissue-mean (Wilcoxon signed-ranks test, p-values < 0.0001). **Figure** 3.3**B** shows tissue-mean vs. tissue-weighted comparisons for each score, and we see that in almost all disease terms tissue-weighted outperforms tissue-mean. See **Appendix Data** A.3.6 and A.3.7 for tissue-mean vs. tissue-weighted performances for each disease term, and for aggregated performances across all disease terms. The improvement remains evident if we use SNVs that are not in the same LD block or ensuring that the SNVs in the training and test datasets are not on the same chromosome (See **Appendix Figure** A.18 - A.21 and Supplemental material for more detail).

While performance-gain for tissue-weighted is broadly consistent across diseases, for some it is more pronounced than for others. To illustrate this, we selected four disease terms with a high performance gain, four terms with a medium gain, and four terms where we observed the least gain (Best improvement, ranking 1-4; middle improvement, ranking 20-23; least improvement, ranking 108-111). **Figure** 3.4 shows our findings, where variability in tissue-weighted performance induced by varying train-test-fold splits during cross-validation is also displayed. We see that for Celiac Disease (EFO:0001060), Systemic Scleroderma (EFO:0000717), Chronic Lymphocytic Leukemia (EFO:0000095) and Sclerosing Cholangitis (EFO:0004268) performance is consistently improved for all three tissue-weighted scores, while for Retinopathy (EFO:0003839), Endometriosis (EFO:0001065), Diabetic Nephopathy (EFO:0000401) and HIV-1 Infection(EFO:0000180) we find no improvement. We also note that disease terms with pronounced improvement appear to have better baseline (i.e., tissue-mean) performance than disease terms where we find little or no benefit of the tissue-weighted approach. Improvement for diseases shown in **Figure** 3.4 is largest for DHS, but, consistent with **Figure** 3.3, we see improvement for Fitcons2 and Genoskyline as well (but not as much).

3.4.3.2 DNase I hypersensitivity (DHS) scoring outperforms other tissue specific scores

To quantify relative performance of the three different scores, we proceed similarly to organism-level scores. Focusing on pairwise comparisons we find that DHS scores outperform Genoskyline and Fitcons2 for most disease terms, and on average (see **Table** 3.2). This observation is consistent with **Figure** 3.3 and 3.4, which often show higher average precision values for DHS than for the other two scores. Notably, baseline (i.e., tissue-mean) performance of DHS does not appear significantly better than that of Genoskyline (**Figure** 3.3. **Appendix Data** A.3.8 and A.3.9 contain details for comparisons between DHS, Fitcons2 and Genoskyline for all disease terms). Next, we explored whether disease-specific tissue weights outperform organism-level scores.

3.4.3.3 DNase I hypersensitivity (DHS) tissue-weighted scoring outperforms organism-level variant scores

To compare the DHS tissue-weighted score with organism-level scores, we directly contrasted their performance. Similar to before, **Table** 3.3 summarizes DHS "wins" (= significantly better performance of DHS tissue-weighted, p-value ≤ 0.05), losses, and ties, compared with other five organism-level variant scores, individually (per disease term) and aggregated (across disease terms). In addition, **Table** A.4 summarizes pair-wise comparisons between tissue-weighted DHS and each organism-level score. We find that DHS tissue-weighted outperforms all organism-level scores in the aggregated analyses, and that it outperforms all other scores on the majority of disease terms (it only performs significantly worse than any other score in 44 out of 550 comparisons).

GenoCanyon is the most competitive organism-level score, where DHS is significantly better for 92 terms out of 111 (∼83%). Interestingly, LINSIGHT performs better against DHS than GenoCanyon, which is the best overall performing organism-level score (see **Table** A.4). **Appendix Data** A.3.10 contains detailed results for each comparison. We also find that DHS outperforms organism-level scores when aggregating over disease terms (**Appendix Data** A.3.11).

To illustrate the gain in performance, we selected four example disease terms where disease-specific variant prioritization yielded high improvements, medium improvements, comparable performance, and worse performance, respectively. Selection was based on ranking differences between DHS and GenoCanyon: best improvement, ranks 1-4; medium improvements, ranks 25-28; comparable performance, ranks 64-67; GenoCanyon better, ranks 108-111. Results are summarized in **Figure** 3.5, where we find substantial improvements using tissue-weightes scoring for Systemic Sceleroderma (EFO:0000717), Celiac Disease (EFO:0001060), Sclerosing Chalangitis (EFO:0004268) and Multiple Sclerosis (EFO:0003885), for which we have already noticed substantial improvement of DHS tissue-weighted over DHS tissue-mean. Disease terms where GenoCenyon is performing better include Venous Thromboembolism (EFO:0004286), Diverticular Disease (EFO:0009959), Non-small Cell Lung Carcinoma (EFO:0003060), and Lung Adenocarcinoma (EFO:0000571).

To make DHS tissue-weighted scores available, we generated pre-computed scores for 111 diseases at every base across the genome (for chromosomes 1-22, available at `https://doi.org/10.7910/DVN/AUAJ7K`). Scores were calculated at 25 bp resolution, the same as DHS scores.

3.4.4 DNase I hypersensitivity (DHS) scoring performs well compared with DIVAN

Here we compare the performance of DHS tissue-weighted scores with DIVAN [43], a disease-specific variant score for 45 diseases. DIVAN is based on a more complicated feature-selection and ensemble-learning framework, and it uses a variety of other functional genomics features, in addition to DNase I hypersensitivity. To compare our method with DIVAN, we mapped our EFO disease terms to MeSH terms (as used by DIVAN) and use MeSH terms later for this section (See **Appendix Data** A.3.12). Because DIVAN uses as supervised learning approach, and because the published model was trained using GWAS SNVs, it was necessary to create specific train and test datasets to ensure a meaningful comparison between tissue-weighted DHS and DIVAN.

Therefore, to assess performance of both DIVAN and DHS, we created a test set of disease-associated variants (and their matched controls) that were published later than 2016 (DIVAN's publication date). That is, these variants are unlikely to have been a part of DIVAN's training data. We also created a training set for DHS tissue-weighted containing only SNVs published prior to 2016. This resulted in training data that (a) is distinct from the test set and (b) draws on similar information that was available for DIVAN's training. Further on, we only selected disease terms for this training/test data combination where at least 20 term-associated SNVs were present in the training data, and where at least 50 SNVs were present in the test data. This approach yielded 29 disease terms for this analysis. We then re-trained tissue-weighted DHS on this training data and compared with DIVAN on the test data. In addition, we added the organism-level GenoCanyon score as a reference.

To assess performance, we performed all pairwise comparisons for each disease term, and evaluated performance based on average precision. **Table** 3.4 summarizes observations, where we find that DHS performs significantly better than GenoCanyon and DIVAN in a majority of comparisons; however, there is a substantial number of comparisons (22 out of 58) where either GenoCanyon or DIVAN outperform DHS. **Figure** 3.6 further illustrates these comparisons. In panel **A** we show performance across disease terms, grouped by the best-performing method. We see that tissue-weighted DHS outperforms DIVAN and GenoCanyon substantially on Multiple Sclerosis (MeSH:D009103), Psoriasis (MeSH:D011565) and Inflammatory Bowel Disease (MeSH:D015212); DIVAN outperforms GenoCanyon and DHS on Arthritis, rheumatoid (MeSH:D001172) and Heart failure (MeSH:D006333); GenoCanyon outperforms DHS and DIVAN on Stroke (MeSH:D020521) and Alzheimer disease (MeSH:D000544). In panels **B-D** we directly summarize comparison results; we observe that the DHS tissue-weighted score often has an advantage in terms where prioritization efforts are overall more successful (upper right quadrants). Finding overall good performance for our approach, we next more closely examined the disease-specific tissue aggregation weights we use for our scores.

3.4.5 Disease-specific tissue weights reflect biomedical relevance

In addition to prioritizing SNVs, we can interpret the disease-specific tissue weights that our model learns in the context of disease mechanisms. Specifically, large tissue weights implicate tissues with a prominent role in associating SNVs with a disease in our model, and one may hypothesize that they have a function in the etiology of that disease. To investigate this hypothesis, we analyzed tissue weights of the top-performing models we derived, where each model represents a different disease.

Results are summarized in **Table** 3.5; they include the two top-performing models, Systemic scleroderma (rank 1) and Sclerosing cholangitis (rank 2). In order to report a diverse range of diseases, we next excluded any diseases that are descendants of immune system disease (EFO:0000540) or lymphoma (EFO:0000574). From the remaining diseases, we identify the next three highest-ranked models: Colorectal adenoma (rank 15), Atrial fibrillation (rank 20), and Cutaneous melanoma (rank 21). For each diseases, we list the five tissues with the largest tissue-weights, and their tissue group.

The tissues we associate with disease, overall, appear reasonable and generally are in-line with existing knowledge about disease mechanisms. Systemic scleroderma is an autoimmune disorder that can affect the skin and internal organs [93]. We find that GM12878 lymphoblastoid cells (a type of B cell) are among highest-weighted tissues, as were other types of B cells (primary B cell and B cell lymphoma, respectively). This is in-line with previous studies that have shown that B cells play a role in system scleroderma [94, 95]. Sclerosing cholangitis is an inflammatory condition that leads to scarring and narrowing of the bile ducts [96]. We highlight various inflammation-related types of blood cells, such as T cells and monocytes, which were previously suggested to play a role in the disease [97]. Colorectal adenoma is a benign tumor that develops in the lining of the colon or rectum. Our model identified rectal mucosa and stomach mucosa as the most-highly weighted tissues, and the function of rectal mucosa in colorectal cancer has been previously studied [98]. While the direct relationship between other gastrointestinal tissues and the development of colorectal adenoma has not been established, the association between gastrointestinal microbiome and colorectal adenomas has been discovered [99]. Regarding atrial fibrillation, our approach highlights fetal heart and lung tissues. In addition, we identified skeletal muscle cells. In the case of cutaneous melanoma, a type of skin cancer, our approach emphasizes foreskin melanocyte cells and a specific type of T cells. Apart from these, we highlight cervical carcinoma cell lines and endothelial primary cells.

Overall, we conclude that the tissue weights we derive carry biomedically meaningful information and are able to highlight tissue contexts that may play a role in disease etiology. To further explore this finding, we used a resource of the epimap consortium [84], where disease-tissue associations are reported that derived differently from the one we obtained in two key ways: First, epimap uses their enhancer definitions based on a much larger set of genome annotations. Second, epimap's enrichment test contrasts disease-associated SNV enrichment in a specific tissue's enhancer set compared to all enhancers, whereas our method effectively compares open chromatin harboring disease-associated SNVs vs control SNVs tissue-by-tissue. Nevertheless, results are summarized in **Appendix Table** A.7, and we find that out of the 25 tissues we associate with disease terms 14 have an estimated false discovery rate of less than 4% in the epimap analysis as well. Notably, a ground truth for these associations is generally unknown; but we interpret the overlap in associations as encouraging, while complementary associations are expected, given the differences in methodology. Based on this overall finding of meaningful disease-tissue associations, we next further explored the use of tissue-weights in disease characterization.

3.4.6 Disease-term similarity based on DHS tissue-weighted modeling reveals meaningful groups

Disease-specific tissue weights for aggregating DHS scores, which are learned by our approach, can highlight tissues and cell-types with a role in the disease (see previous section). Therefore, we derived and explored a measure for disease similarity based on these weights, which we detail in the following.

3.4.6.1 Disease similarities based on disease-specific tissue weights for non-coding variant prioritization

In our DHS tissue-weighted approach, for each disease term DNA accessibility across the same set of tissues is used to predict whether a certain SNV is disease-associated, or not. This results in disease-specific tissue aggregation weights (or coefficients) $\left\{\beta^{(i)} \in \mathbb{R}^d\right\}_{i=1}^n$, where i is indexing disease terms, n is the number of disease terms studied, and d denotes the number of tissues/cell-types with DHS scores. For our similarity measure between two diseases, say i and j, we then use a version of the Pearson correlation between $\beta^{(i)}$ and $\beta^{(j)}$that takes uncertainty in the estimated aggregation weights into account (see **Methods**). That is, if an overlapping set of tissues/cell-types drive the prioritization of SNVs for two diseases, similarity is high; if different tissues are used, similarity is low.

Using this approach we calculated disease similarities for the 111 disease terms we study. Resulting similarities are visualized in **Figure** 3.7, where we show a similarity-based 2D UMAP projection of disease terms. We observe that disease terms segregate into separate groups, with a coarse grouping between immune related diseases (lower left inlay, black) and others (lower left inlay, gray). A higher-resolution group structure was obtained by sub-clustering, where we grouped disease terms into seven groups (main panel, **Figure** 3.7).

Clusters names are based on EFO disease terms that include a large amount of cluster members as child-terms (see **Methods** and **Appendix Figure** A.11- A.17); **Table** 3.6 lists disease terms per cluster. In addition to the clear separation of immune-related diseases from others, we also find a very homogeneous group consisting of mental and behavioural disorders, containing terms like schizophrenia (EFO:0000692) and anxiety disorder (EFO:0006788), and a group of skin cancers. The remaining three groups are more heterogeneous, but with two of them containing several terms related to cardiovascular disease (EFO:0000319) and digestive system disorders (EFO:1000218), respectively. By design similar tissues in each group drive SNV-disease associations, and we next examined which tissues play a role in each of the clusters.

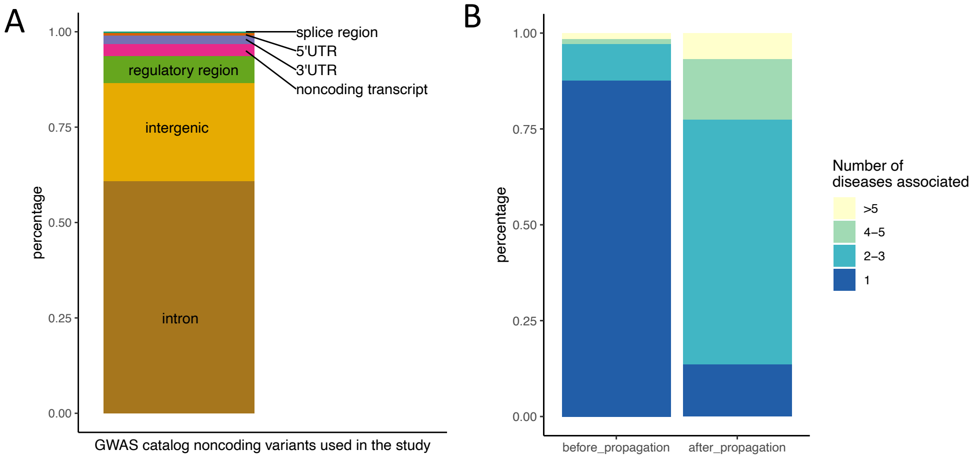

Figure 3.1: *20,656 disease-associated non-coding SNVs.* (A) Genomic context of non-coding SNVs used in this study. (B) Percentage of the SNVs used that are annotated to 1, 2-3, 4-5 or more than 5 disease phenotypes, before and after propagating SNV-phenotype associations according to EFO parent-child annotations. Genomic context annotation is adapted from the CONTEXT column from the GWAS catalog, where we combine splice donor, splice region and splice acceptor variants into splice variants and we combine TF binding variants and regulatory regions variants into regulatory region variants.

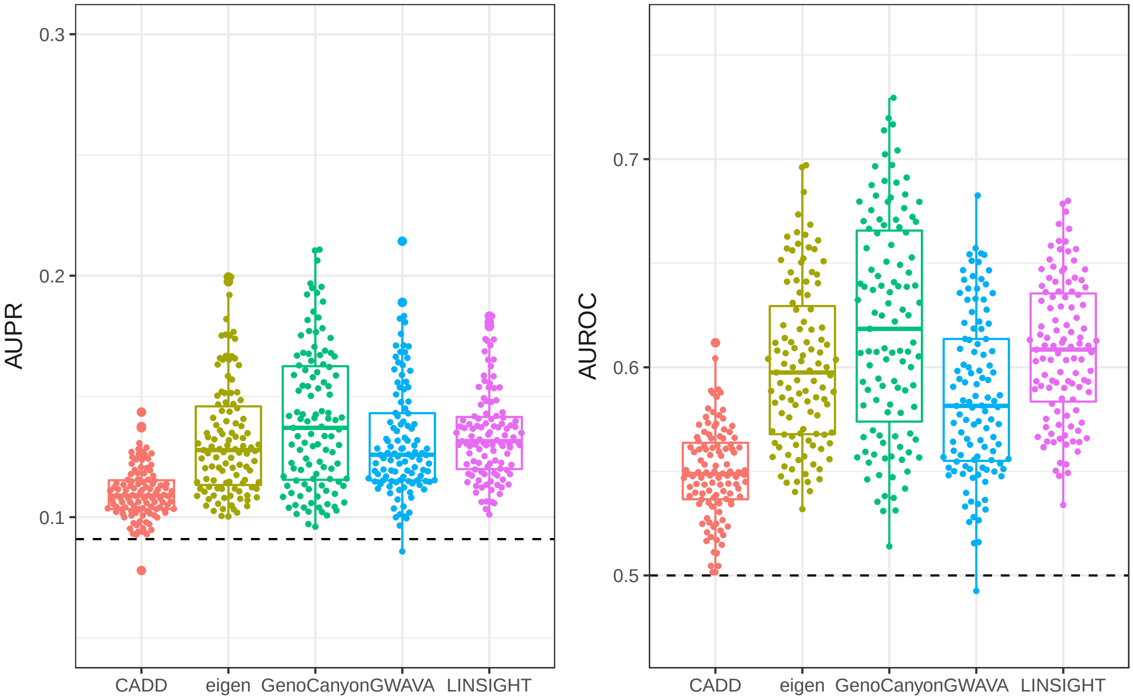

Figure 3.2: *Organism-level variant scores are moderately successful in prioritizing non-coding disease-associated variants.* Different organism-level variant prioritization scores are shown on the x-axis, the y-axis displays performance in terms of average precision (area under the precision recall curve, AUPR, left panel) and area under the receiver-operator curve (AUROC, right panel). Each point represents a specific disease term from the experimental factor ontology. Horizontal lines spanning data sets show expectations under random guessing.

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
GenoCanyon	307	106	31	4	0	0
LINSIGHT	281	146	17	1	1	2
GWAVA	221	196	27	1	1	2
eigen	219	201	24	1	1	2
CADD	24	403	17	0	4	0

Table 3.1: *Relative performance of organism-level variant scores.* Wins, Losses, Ties refers to significantly better (or worse, or tied) performance across all possible pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are four other methods and 111 terms, i.e. 444 comparisons), while the last three columns represent results of aggregate comparisons across terms. Average precision was used as the performance metric, and Wilcoxon singed-ranks tests to determine wins and losses (p-values larger than 0.05 are reported as ties).

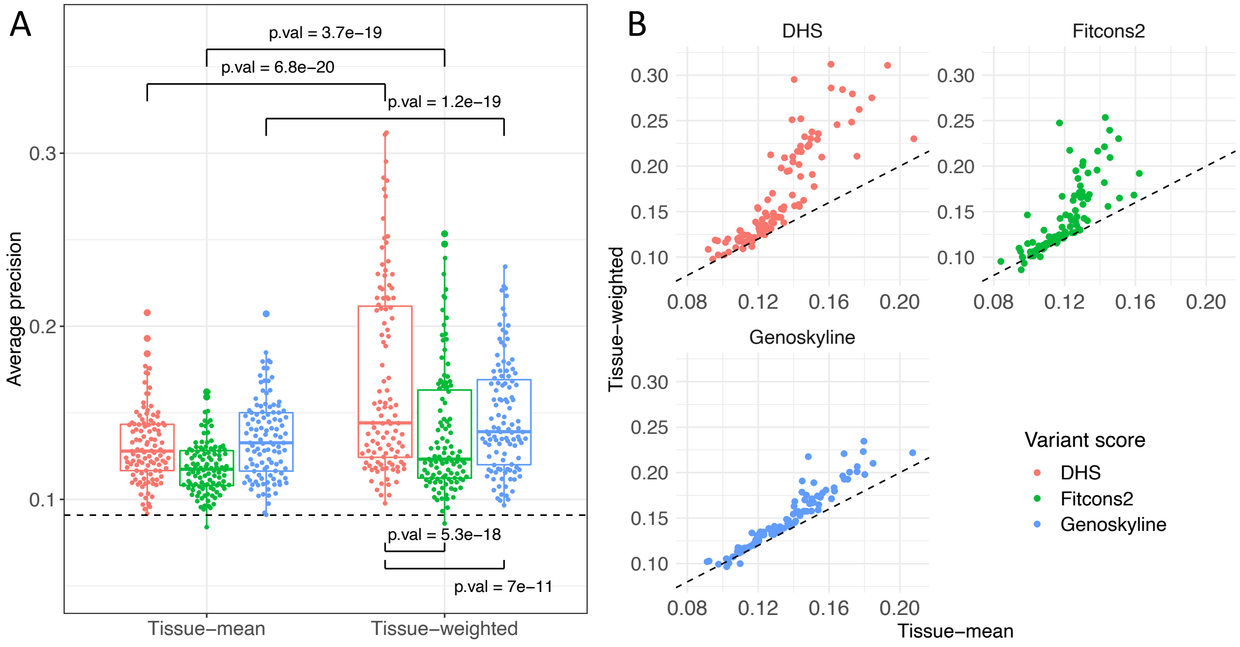

Figure 3.3: *Disease-specific tissue weights improve variants prioritization.* Performance of three tissue-specific variants scores (DHS, Fitcons2, Genoskyline) used to prioritize non-coding disease-associated variants for disease terms using two approaches: *tissue-mean* (i.e., disease-agnostic, baseline) on the left side and and *tissue-weighted* (i.e., disease specific) on the right side. P-values were calculated using a Wilcoxon signed-ranks test (A). Scatter plot of tissue-mean vs. tissue-weighted performance (average precision) for each tissue-specific score; dashed line denotes the diagnonal (B).

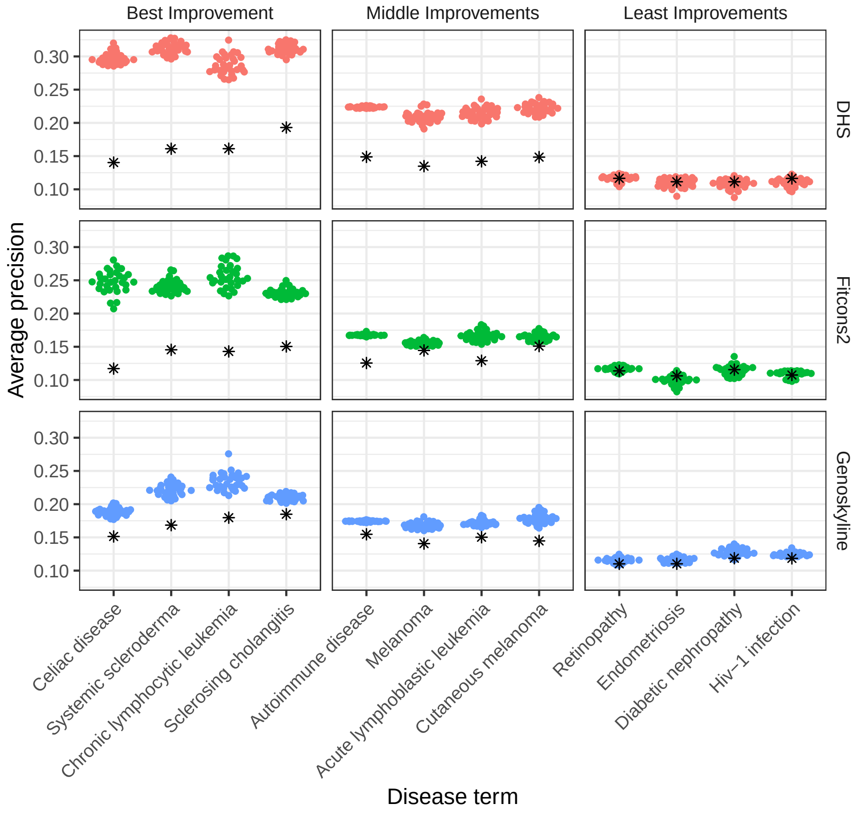

Figure 3.4: *Improvements of tissue-weighted variant scores for representative disease terms.* Shown is the performance of tissue-weighted variant scores (colored points) vs. tissue-mean (black asterisks) as a baseline, for three tissue scores (rows) and stratified by improvement observed: best improvement for the fist column middle for the middle column and least improvement for the right column. The x-axes denote disease terms, the y-axis average precision. Different points for tissue-weighted scores represent different data-splits in the nested cross validation procedure.

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
DHS	180	22	20	2	0	0
Genoskyline	96	94	32	1	1	0
Fitcons2	19	179	24	0	2	0

Table 3.2: *DHS outperforms other tissue weights.* Wins, Losses, Ties refer to significantly better (or worse, or tied) performance across all possible score pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are two other methods and 111 terms, i.e., 222 comparisons), while the last three columns represent results of comparisons aggregated over terms. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were reported as ties).

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
DHS	474	44	37	5	0	0
GenoCanyon	314	198	43	4	1	0
LINSIGHT	298	230	27	1	2	2
GWAVA	233	289	33	1	2	2
eigen	223	299	33	1	2	2
CADD	28	510	17	0	5	0

Table 3.3: *DHS outperforms organism-level variant scores.* Wins, Losses, Ties refer to significantly better (or worse, or tied) performance across all possible score pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are two other methods and 111 terms, i.e., 555 comparisons), while the last three columns represent results of comparisons aggregated over terms. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were reported as ties).

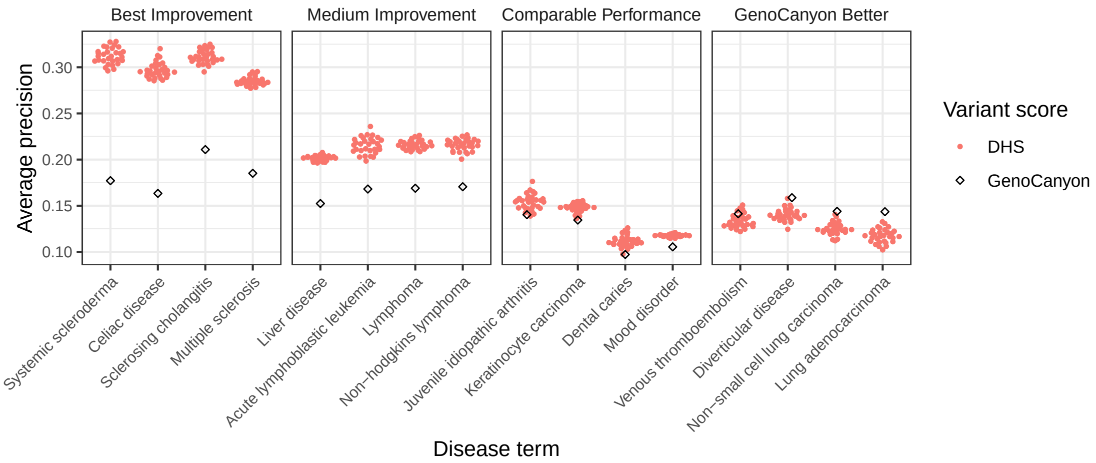

Figure 3.5: *DHS disease-specific scores improve variant prioritization compared with organism-level scores.* For four strata (best and middle improvement, comparable performance and worse performance) we selected for disease terms and compare performance results. GenoCanyon performance is denoted in black, DHS tissue-weighted in red. Different performances of DHS represent variation different data splits during nested cross validation (see **Methods**).

Score	Wins	Losses	Ties	Winning percent
DHS	34	22	2	61
GenoCanyon	26	31	1	46
DIVAN	25	32	1	44

Table 3.4: *DHS tissue-weighted disease-specific scoring outperforms DIVAN.* Across 29 disease terms, this table summarizes all pairwise comparison for DHS tissue-weighted, GenoCanyon and DIVAN using a specifically created test dataset. Wins, losses, ties refer to significantly better (or worse, or tied) performance. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were ties). Winning percent = #Wins/(#Wins+#Losses).

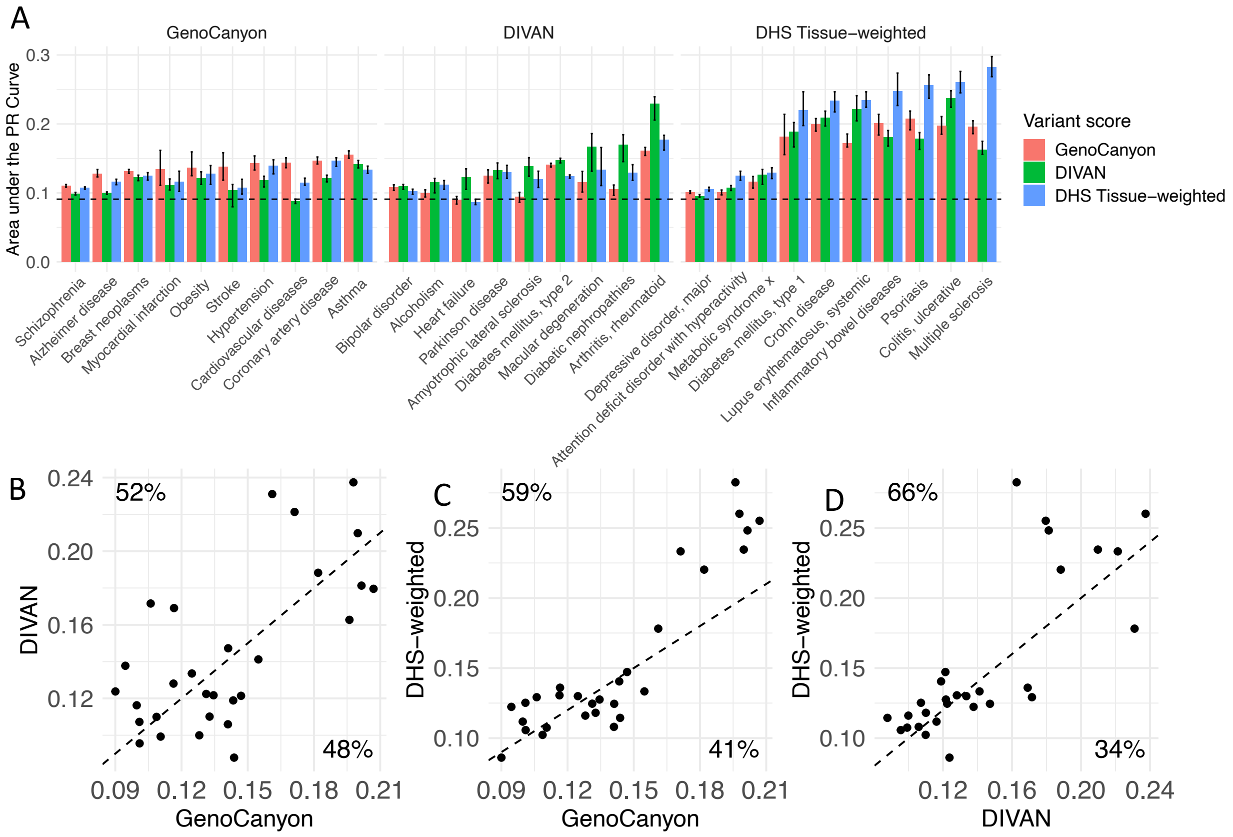

Figure 3.6: *DHS tissue-weighted scoring outperforms DIVAN.* Performance of DIVAN, GenoCanyon, and DHS tissue-weighted across the test set, with disease terms grouped by the best-performing method. Vertical striped indicates the minimum and maximum performance of 30 bootstrap samples (A). Performance scatter plots of GenoCanyon vs. DIVAN performance (B); GenoCanyon vs. DHS-weighted (C); DIVAN vs. DHS-weighted performance (D). Average precision was used for these plots; dashed lines denote equal performance. Percentages denote the fraction of points above and below the diagonal, respectively.

Rank	ID	Tissue name	Group
Systemic scleroderma			
1	E116	GM12878 Lymphoblastoid Cells	blood
2	E032	Primary B cells from peripheral blood	blood
3	E041	Primary T helper cells PMA-I stimulated	blood
4	E123	K562 Leukemia Cells	blood
5	E030	Primary neutrophils from peripheral blood	blood
Sclerosing cholangitis			
1	E116	GM12878 Lymphoblastoid Cells	blood
2	E061	Foreskin Melanocyte Primary Cells skin03	skin
3	E102	Rectal Mucosa Donor 31	gi_rectum
4	E041	Primary T helper cells PMA-I stimulated	blood
5	E029	Primary monocytes from peripheral blood	blood
Colorectal adenoma			
1	E102	Rectal Mucosa Donor 31	gi_rectum
2	E110	Stomach Mucosa	gi_stomach
3	E057	Foreskin Keratinocyte Primary Cells skin02	skin
4	E101	Rectal Mucosa Donor 29	gi_rectum
5	E028	Breast variant Human Mammary Epithelial Cells (vHMEC)	breast
Atrial fibrillation			
1	E083	Fetal Heart	heart
2	E108	Skeletal Muscle Female	muscle
3	E107	Skeletal Muscle Male	muscle
4	E088	Fetal Lung	lung
5	E120	HSMM Skeletal Muscle Myoblasts Cells	muscle
Cutaneous melanoma			
1	E061	Foreskin Melanocyte Primary Cells skin03	skin
2	E059	Foreskin Melanocyte Primary Cells skin01	skin
3	E117	HeLa-S3 Cervical Carcinoma Cell Line	cervix
4	E041	Primary T helper cells PMA-I stimulated	blood
5	E122	HUVEC Umbilical Vein Endothelial Primary Cells	vascular

Table 3.5: *Top-ranked tissues for five diseases.* For five diseases when show the top-five tissues with the largest tissue weights in the corresponding model we derive. The first column is the tissue rank, the second the tissue's roadmap ID, the third the tissue name, and the fourth column is the tissue group.

🟦 heterogeneous

- adolescent idiopathic scoliosis
- age-related macular degeneration
- 🟨 alcohol dependence
- amyotrophic lateral sclerosis
- chronic obstructive pulmonary disease
- 🟪 dental caries
- diabetic nephropathy
- 🟨 drug dependence
- □ endometriosis
- epilepsy
- gout
- hiv infection
- hiv-1 infection
- 🟫 lung adenocarcinoma
- 🟫 lung carcinoma
- neuropathy
- 🟪 non-alcoholic fatty liver disease
- 🟫 non-small cell lung carcinoma
- obesity
- 🟪 periodontitis
- peripheral neuropathy
- scoliosis
- 🟥 🟫 squamous cell lung carcinoma
- 🟧 venous thromboembolism

🟪 digest/cancer

- 🟩 ⬛ autoimmune thyroid disease
- 🟫 breast carcinoma
- 🟫 cancer
- 🟧 cardiovascular disease
- 🟪 colorectal adenoma
- 🟪 🟫 colorectal cancer
- 🟧 coronary artery disease
- 🟪 diabetes mellitus
- 🟪 🟫 digestive system carcinoma
- 🟪 digestive system disease
- female reproductive system disease
- 🟧 hypertension
- 🟩 🟫 multiple myeloma
- 🟨 neurotic disorder
- 🟪 🟫 pancreatic carcinoma
- 🟫 prostate carcinoma
- respiratory system disease
- 🟥 🟫 squamous cell carcinoma.
- 🟪 🟩 ⬛ type i diabetes mellitus
- 🟪 type ii diabetes mellitus

🟩 immune

- 🟩 🟫 acute lymphoblastic leukemia
- adult onset asthma
- 🟩 allergic rhinitis
- 🟩 allergy
- atopic asthma
- 🟩 🟪 celiac disease
- childhood onset asthma
- 🟩 🟫 chronic lymphocytic leukemia
- 🟪 cirrhosis of liver
- hypothyroidism
- 🟩 ⬛ juvenile idiopathic arthritis
- 🟩 🟫 lymphoid leukemia
- 🟫 lymphoma
- 🟩 🟫 neoplasm of mature b-cells
- 🟫 non-hodgkins lymphoma
- 🟩 ⬛ systemic lupus erythematosus
- 🟩 🟧 systemic scleroderma

🟧 cardiovascular/others

- 🟨 alzheimer's disease
- 🟧 atherosclerosis
- 🟧 atrial fibrillation
- 🟧 cardiac arrhythmia
- chronic kidney disease
- 🟪 diverticular disease
- glaucoma
- 🟧 heart failure
- metabolic syndrome
- 🟧 migraine disorder
- osteoarthritis
- 🟫 ovarian carcinoma
- parkinson's disease
- 🟧 peripheral arterial disease
- retinopathy
- 🟧 stroke
- uterine fibroid

⬛ immune/autoimmune

- ⬛ 🟩 ankylosing spondylitis
- asthma
- ⬛ 🟩 autoimmune disease
- ⬛ 🟩 🟪 crohn's disease
- 🟩 hypersensitivity reaction disease
- 🟩 immune system disease
- ⬛ 🟩 🟪 inflammatory bowel disease
- kidney disease
- 🟪 liver disease
- ⬛ 🟩 multiple sclerosis
- 🟩 psoriasis
- ⬛ 🟩 rheumatoid arthritis
- 🟪 sclerosing cholangitis
- skin disease
- ⬛ 🟩 🟪 ulcerative colitis

🟨 mental

- 🟨 anorexia nervosa
- 🟨 anxiety disorder
- 🟨 attention deficit hyperactivity disorder
- 🟨 autism spectrum disorder
- 🟨 bipolar disorder
- 🟨 eating disorder
- 🟨 mental or behavioural disorder
- 🟨 mood disorder
- movement disorder
- 🟨 obsessive-compulsive disorder
- 🟨 psychosis
- 🟨 schizophrenia
- 🟨 tourette syndrome
- 🟨 unipolar depression

🟥 skin cancer

- 🟥 🟫 cutaneous melanoma
- 🟥 🟫 keratinocyte carcinoma
- 🟥 🟫 melanoma
- 🟥 🟫 non-melanoma skin carcinoma

legend

- 🟪 digestive system disease
- 🟩 immune system disease
- ⬛ autoimmune disease
- 🟧 cardiovascular
- 🟨 mental or behavioural disorder
- 🟥 skin cancer
- 🟫 cancer

Table 3.6: *Disease groups based on model similarity.* For each disease group disease terms are shown. The colored squares denote the disease groups in the EFO ontology.

In order to find group-specific tissues, we examined for each cluster the top five tissues that (a) contribute most to disease association and (b) are cluster specific (see **Methods**). Results are summarized in **Figure** 3.8; we note that both disease groups related to the immune system highlight blood tissues (such as E043: Primary T helper cells from peripheral blood and E116: GM12878 Lymphoblastoid Cells, see **Appendix Data** A.3.23 for all names of standard epigenomes), with the group containing inflammatory bowel disease, Crohn's disease, and ulcerative colitis also containing rectum tissues (such as E101: Rectal Mucosa Donor 29). Brain tissues contribute to disease associations for mental and behavioral disorders, skin tissues to skin cancer, and gastro-intestinal / stomach tissue to the cluster with digestive system diseases. We also note that a clear association of specific tissues with disease groups correlates with better classification performance of our model for SNV-disease association (**Figure** 3.8, for example, the immune and immune/autoimmune clusters). We note, though, that not for all clusters the corresponding tissue associations are equally compelling, as illustrated in the same figure. While the clusters we derive resemble broader disease groups, for each disease a specific combination of tissues is used to derive whether a variant might be associated, and some tissues contribute to several clusters. For instance, one blood cell type (E116, GM12878 Lymphoblastoid Cells) contributes to both immune clusters, but also to diseases in the digestive/cancer, heterogeneous and skin cancer clusters. Another blood cell type (E043, Primary T helper cells from peripheral blood) displays a similar pattern. **Appendix Figure** A.10 shows the same heatmap as **Figure** 3.8, but for all tissues.

Overall, these results suggest that our modeling approach successfully identifies tissues with a role in disease etiology. Before exploring disease-tissue relations in more detail , we explore how our disease similarities relate to genetic similarities as measured by genetic correlation between two diseases.

3.4.6.2 Model-based similarities are complementary to genetic correlation.

Here we compare the disease-disease similarities we derived (s_m) with genetic correlations from the GWAS Atlas (s_g), where genetic correlation measures shared genetic causes

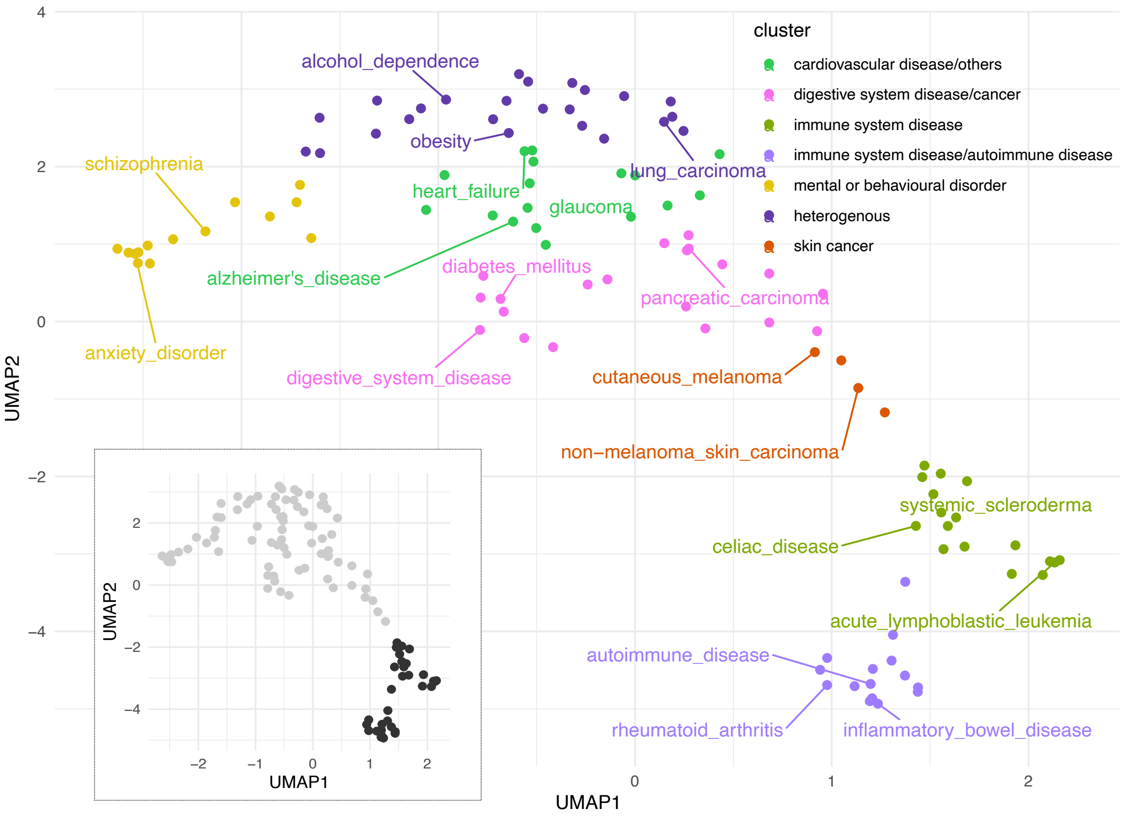

Figure 3.7: *UMAP plot shows disease-disease relationships among 111 diseases.* Two dominant clusters (inlay: immune system related disease terms (black) and others (gray)). Hierarchical clustering was used to group diseases into 8 clusters.

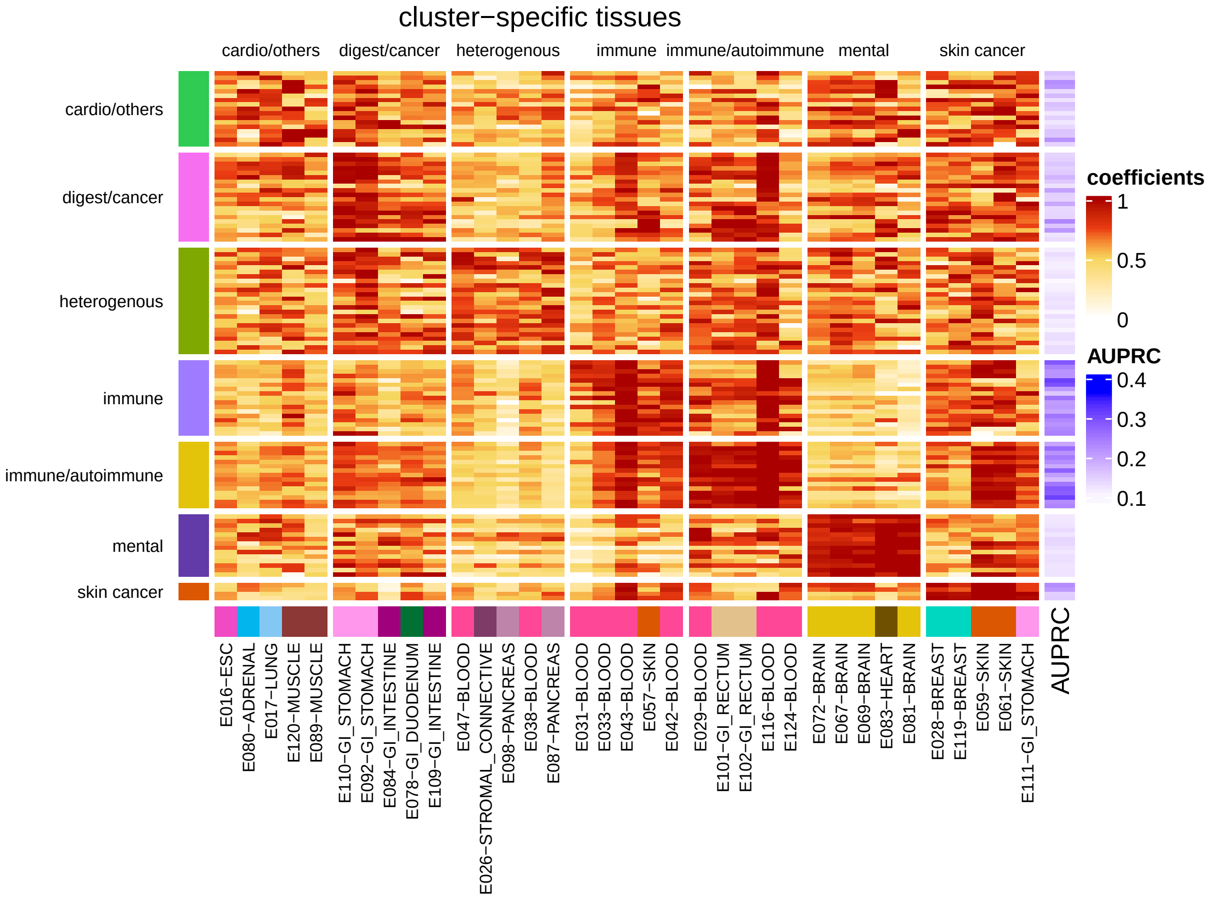

Figure 3.8: *Heatmap of top-five tissue-weights for 111 diseases.* Regularized model coefficients (i.e., tissue weights) of five disease-cluster-specific tissues (columns) are shown for 111 diseases (rows). Coefficients are scaled by disease, and rows are grouped into sets of cluster-specific tissues (see **Methods** section). Bottom annotation shows tissue names of cluster-specific tissues (names are shown in the format of 'Tissue name' - 'Tissue group'; annotation on the left side shows disease cluster, and annotating on the right side shows model performance in terms of AUPRC).

between two traits [92]. For 6,105 possible disease pairs of the 111 diseases terms we study, estimates for 595 pairs were available from the GWAS Atlas (see **Methods**). Overall, for these 595 disease pairs we observe only weak correlation between model similarities and genetic correlations ($r = 0.32$, $p\ value = 2.4E - 15$), where the scatter plot is shown in **Figure** 3.9**A**. We also see that most disease pairs are not annotated with substantial genetic correlations, or with model-based similarities (individually, 90% disease pairs has $s_m < 0.25$, and $s_g < 0.20$). Therefore, we explored three different regimes: Disease pairs where both similarity measures are high ($s_m \geq 0.25$ and $s_g \geq 0.20$), pairs with high genetic correlations and low model similarity ($s_m < 0.25$ and $s_g \geq 0.20$) vice versa (quadrants indicated in **Figure** 3.9**A**, named quadrants B,C and D). From each regime, we highlighted the top 8 extreme examples and we show them in **Table** 3.7. In the following we discuss one example from each regime. Here, we pick two immune system diseases for quadrant B; two mental or behavioral disorders for quadrant C; and one immune system disease and one mental or behavioral disorder for quadrant D. In addition, we pick example disease pairs without any parent-child relationships.

Ulcerative colitis (UC, EFO:0000729) and Crohn's disease (CD, EFO:0000384), for instance, have both high genetic correlation ($s_g = 0.53$) and model similarity ($s_m = 0.84$), see **Figure** 3.9**B**. This suggests that they share genetic causes, and that the same tissues are informative for SNV-disease association. While shared genetic causes for UC and CD have been pointed out (e.g., [100]), our model for SNV-disease association allows us to explore relevant tissue contexts. In **Figure** 3.9**B** we show a scatter plot of tissue weights for both diseases, where color indicates the importance of each tissue to model similarity (see **Methods**). We observe that open chromatin in blood (E116, GM12878 Lymphoblastoid Cells; E124, Monocytes-CD14+ RO01746 Primary Cells; E041, Primary T helper cells PMA-I stimulated) and rectum (E102, Rectal Mucosa Donor 31) is positively associated with SNV-disease association in both diseases, which is consistent with a previous study where blood cell types are found to be relevant in many autoimmune diseases, including UC and CD [101]; in addition, symptoms or complications in rectum is also observed in UC and CD [102]. Interestingly, open chromatin in GI-intestine (E085, fetal intestine small) is negatively associated with SNV-disease association, along with other intestine tissues (E084, fetal intestine large and E109, small intestine, with the 61th and 86th smallest tissue weight, respectively, amongst 127 contexts). This indicates fetal intestine or small intestine might not be less involved in UC and CD etiology, compared to their juvenile and adult counterparts.

Autism spectrum disorder (ASD, EFO:0003756) and anorexia nervosa(AN, EFO:0004215) is an example where we observe a low genetic correlation ($s_g = -0.05$) and a moderate high model similarity ($s_m = 0.34$), and as scatter plot of their tissue weights is shown in **Figure** 3.9**C**. We didn't choose one of the highlighted pairs in this quadrant since we want to look at examples from different groups rather than just immune system diseases where those highlight pairs are. We observe that both disease models give heart and brain tissue (E083, fetal heart and E081, fetal brain male) high tissue weights. This is consistent with the observation of brain abnormalities in ASD and AN [103, 104]. While the presence of fetal heart is less intuitive, we note that children with abnormal heart development are more likely to develop ASD, suggesting a connection between the disease and the fetal heart [105]. We also note that while the genetic correlation between ASD and AN is low, a link between the two diseases on the phenotypic level is being suggested [106, 107]; the tissue context we identified could provide information about a shared molecular disease etiology as well.

For obsessive compulsive disorder (EFO:0004242) and celiac disease (EFO:0001060) we observe low model similarities ($s_m = -0.26$) and moderately high genetic correlation ($s_g = 0.36$), and **Figure** 3.9 **D** shows a scatter plot of the tissue weights we derive. Several studies have shown that nervous system disease and immune related diseases have shared genetic background [108, 109]. However, in contrast to the other two examples, there is little relation between tissue weights in these two diseases. Blood cell types are highlighted in celiac disease, while brain and fetal heart tissues are highlighted in obsessive compulsive disorder. For celiac disease, the top six tissue contexts are blood cells, including different types of T cells (E041, Primary T helper cells PMA-I stimulated; E043, Primary T helper cells from peripheral blood and E034, Primary T cells from peripheral blood) and lymphoblasts (E116, GM12878 Lymphoblastoid Cells), which is consistent with findings that alterations in T cells and lymphoblasts can lead to celiac disease [110, 111].

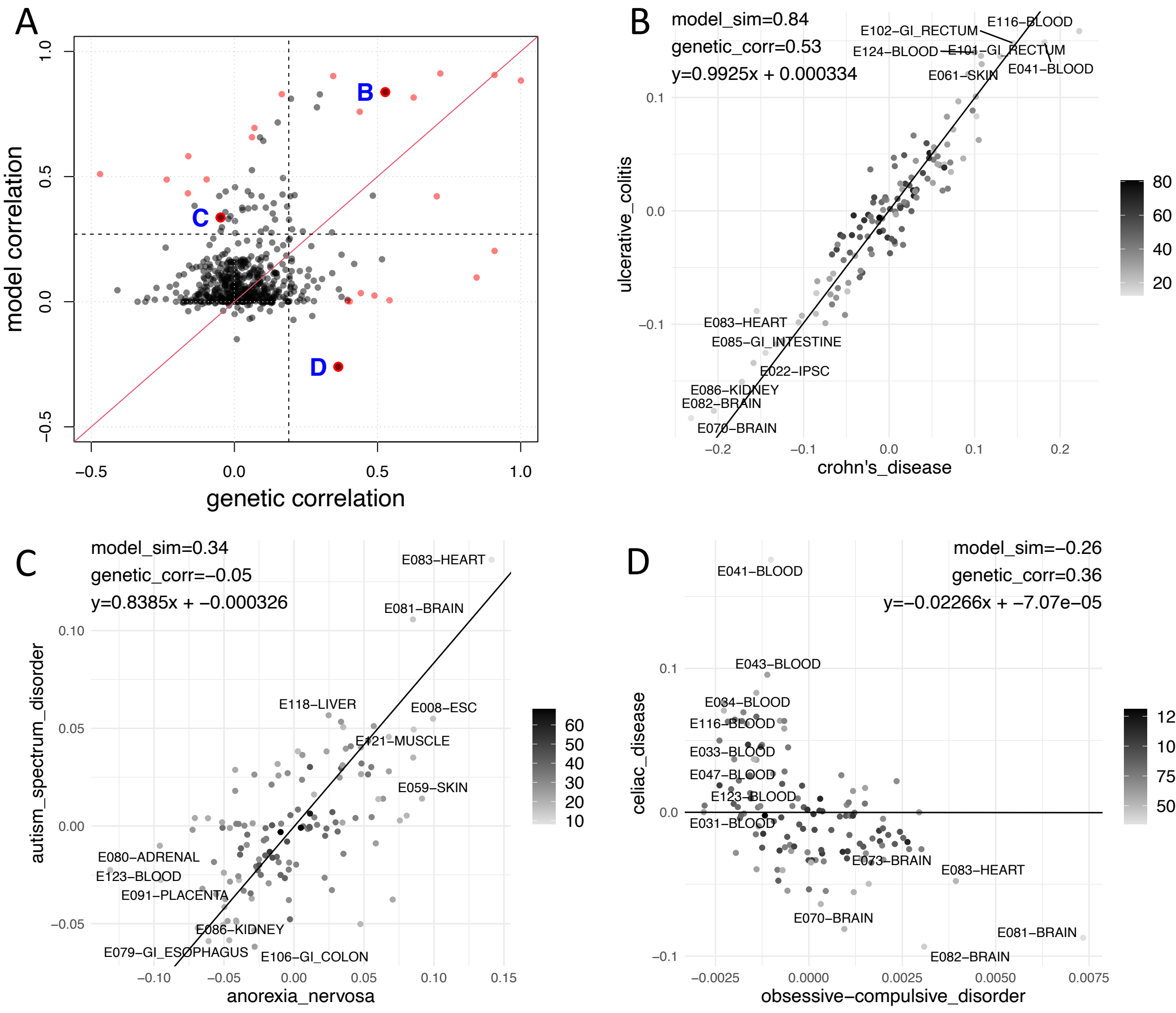

Figure 3.9: *Genetic correlation and model similarity.* (A) Genetic correlation vs model similarity for 595 disease pairs. Each dot is a disease pair, where the x axis denotes the genetic correlation and y axis is the disease model similarity. For B,C and D quadrants, we highlighted the top 8 extreme pairs, where highest 8 pairs with $s_g + s_m$, $s_m - s_g$, and $s_g - s_m$ are selected for quadrant B, C and D, respectively. (B-D) Scatter plot of tissue coefficients in three example disease pairs, where (B) shows Crohn's disease vs inflammatory bowel diseases; (C) shows anorexia nervosa vs autism spectrum disorder and (D) shows celiac disease vs obsessive compulsive disorder. Lines shows the weighted linear regression line. Color shows the weight for each disease pair when conducting weighted regression analysis.

Disease 1	Disease 2	s_g	s_m	Quadrant
Inflammatory bowel disease	Ulcerative colitis	1.00	0.88	B
Diabetes mellitus	Type ii diabetes mellitus	0.91	0.91	B
Crohn's disease	Inflammatory bowel disease	0.72	0.91	B
Sclerosing cholangitis	Ulcerative colitis	0.63	0.82	B
Crohn's disease	Ulcerative colitis	0.53	0.84	B
Ankylosing spondylitis	Sclerosing cholangitis	0.35	0.90	B
Inflammatory bowel disease	Sclerosing cholangitis	0.44	0.76	B
Bipolar disorder	Schizophrenia	0.71	0.42	B
Rheumatoid arthritis	Systemic lupus erythematosus	-0.47	0.51	C
Celiac disease	Systemic lupus erythematosus	-0.16	0.58	C
Sclerosing cholangitis	Systemic lupus erythematosus	-0.24	0.49	C
Crohn's disease	Sclerosing cholangitis	0.17	0.83	C
Rheumatoid arthritis	Sclerosing cholangitis	0.07	0.69	C
Crohn's disease	Rheumatoid arthritis	0.06	0.66	C
Systemic lupus erythematosus	Ulcerative colitis	-0.16	0.43	C
Crohn's disease	Systemic lupus erythematosus	-0.10	0.49	C
Type i diabetes mellitus	Type ii diabetes mellitus	0.85	0.10	D
Diabetes mellitus	Type i diabetes mellitus	0.91	0.20	D
Celiac disease	Obsessive-compulsive disorder	0.36	-0.26	D
Diabetes mellitus	Obesity	0.54	0.01	D
Obesity	Osteoarthritis	0.49	0.02	D
Attention deficit hyperactivity disorder	Obesity	0.44	0.03	D
Attention deficit hyperactivity disorder	Osteoarthritis	0.40	0.00	D
Obesity	Type i diabetes mellitus	0.40	0.00	D

Table 3.7: *Example disease pairs of genetic correlation and model similarities.* This table shows the genetic correlation and model similarity for some disease pairs as we selected. s_g: genetic correlation; s_m: model similarity. For quadrant B, C, D we pick 8 disease pairs, where $s_g + s_m$, $s_g - s_m$ and $s_m - s_g$ are the highest, respectively.

3.5 Discussion

Most variant scores prioritize non-coding variants either at the level of the whole organism (e.g, CADD [6], GenoCanyon [34]), or they provide tissue-specific scores (e.g, GenoSkyline [7], Fitcons2 [37]). Here we present a straightforward strategy to combine tissue-specific variant scores in a disease-specific manner. We show that for common genetic variants in the GWAS catalog [78] our approach leads to better performance than organism-level or tissue-specific scores (see **Figure** 3.5). Pre-computed disease-specific prioritization scores are available at `https://doi.org/10.7910/DVN/AUAJ7K`.

Comparing different variant prioritization methods we note that we use area under the precision-recall curve as an evaluation metric, and that the performance of all methods is modest. We believe that is because our analysis ***(a)*** focuses explicitly on non-coding variants, ***(b)*** stratifies SNVs by disease-phenotype, and ***(c)*** utilizes unbiased matching of control-SNVs (SNPsnap-matching, see Section 3.3.1.2). Each of these points affects the SNV sets we use for our analysis, and therefore the performance metrics we report. For transparency we provide all disease-associated variants we use (with matched negatives) in our supplemental data. More generally, associations reported in the GWAS catalog contain causal as well as non-causal SNVs, which will also contribute to sub-optimal performance measures of all the variant scores we assess.

We included a comparison with the DIVAN method in our evaluation, which also included comparing GenoCanyon with DIVAN. Part of this comparison is analogous to results reported in Chen et al. [43]; however, the performances we observed do not agree perfectly, as detailed in **Appendix Data** A.3.15. Broadly, looking at overlapping/matching disease terms, our results appear more favorable for GenoCanyon. These differences are likely due to different test sets used in the two evaluations (i.e., the GWAS catalog (this study) vs. GRASP (Genome-Wide Repository of Associations Between SNPs and Phenotypes)).

We note that there is other research associating variants with disease terms in a similar setting, notably PINES [42] and LSMM [112]. We did not compare directly with PINES, because no pre-computed scores are available; also, we note that while performance reported in this publication in terms of AUROC is higher than our results, a less stringent un-matched test set of random variants was used in these analyses. For LSMM we note that we leverage variants associated with EFO disease terms across studies, while LSMM uses summary statistics on a per-study basis. Using aggregate data from different studies allows our approach to consider parent-child relationships of the EFO ontology using variant aggregation (see Section 3.4.1).

We show that our approach can be used to calculate similarities between disease terms (”model similarities”), see **Section** 3.4.6.1. Since this similarity measure is derived from non-coding SNVs associated with disease, one could expect it is largely congruent with genetic correlation between disease traits. However, that is not the case (see **Figure** 3.9), most likely because we focus on a small subset of disease-associated SNVs reported in the GWAS catalog. For example, obsessive-compulsive disorder and celiac disease have a high genetic correlation ($s_g = 0.36$) but do not share non-coding SNVs in the GWAS catalog (and low model similarity $s_m = -0.26$), whereas autism spectrum disorder and anorexia nervosa have a low genetic correlation ($s_g = -0.05$) but share a number of significant SNVs in the GWAS catalog (and relative high model similarity $s_m = 0.34$). Further on, interpretation of model similarity between disease terms is different from genetic correlation; high model similarity implies that disease-associated SNVs reside in DNA-accessible regions in an overlapping set of tissues, but the identity of individual SNVs (and whether they overlap) is inconsequential. For example, asthma and rheumatoid arthritis have only 15 shared SNVs (out of 732 and 1283 SNVs in rheumatoid arthritis and asthma, respectively), but exhibit high model similarity ($s_m = 0.53$). This shows that model similarity between two diseases can involve similar tissues even if they do not share a genetic background. Further on, we noted that estimates of genetic correlation also may depend on the study used. For example, systemic lupus erythematosus (SLE) has a negative genetic correlation ($s_g = -0.47$) with rheumatoid arthritis (RA) (and other inflammatory diseases) when using the SLE summary statistics from Julià et al. [113] (as retrieved from the GWAS Atlas [92]), whereas another study (Lu et al., [114]) found SLE to have a positive genetic correlation ($s_g = 0.41$) with RA when using the SLE summary statistics from Bentham et al. [115].

We note that in our analyses we used the EFO ontology to aggregate variants annotated in the NIH/EBI GWAS catalog. That is, for each disease term directly-annotated variants were used, and, in addition, variants annotated to descendant terms in the ontology were also included. This approach allowed us to compile a more exhaustive set of variants per term. However, some amount of caution should be exercised when using disease models with more general terms, such as ”cardiovascular disease” for example, as they may encompass heterogeneous diseases.

Our approach is expected to improve as more variants are associated with disease, and as disease-associations get more refined. In addition, increasing amounts of epigenomics data, such as epimap [84] and ENCODE5 [16], could be incorporated and have the potential to improve the disease associations we learn.

One limitation of this method to be noted is that our prioritization scores are available only for diseases associated with a relatively large number of SNVs in the GWAS Catalog. Specifically, this applies to 111 diseases with more than 100 non-coding SNVs after filtering. However, it is expected that more diseases will be included as additional associated SNVs are discovered in the future.

It is important to note that nearly 80%-90% of the participants in the GWAS Catalog were of European descent [116, 117]; therefore, the disease-associated SNVs derived from the GWAS Catalog for training our model may not be generalized to other ethnic groups. However, a key aspect of our model is that it identified disease-related tissues, such as Rectal Mucosa in Colorectal adenoma; this is not driven by population structure, but rather driven by the overlap of open chromatin regions and disease-associated SNVs. While it is true that different population exhibit unique characteristics, for example allele frequency or LD patterns [116], the identification of disease-relevant tissues holds potential for applicability across various ethnic groups. Nonetheless, we need further research to support the effectiveness of this method in diverse ethnic populations.

In summary, we have provided a straightforward method to leverage tissue-specific variant scores for disease-specific variant prioritization. We show that this approach performs well compared with current methods, and we show that the resulting association models are interpretable and lead to useful characterization of disease terms. Overall, our contributions are useful for the following two reasons: Conceptually, because they highlight the value of disease-specific variant prioritization. In addition, we provide pre-computed prioritization scores for 111 disease terms that researchers can use in practice to interpret their variant data.

3.6 Author contribution

Qianqian Liang contributed to the chapter ”Disease-specific prioritization of non-coding GWAS variants based on chromatin accessibility” by:

- Conducting data preparation and analysis.
- Drafting the manuscript.
- Establishing the Github repository: `https://github.com/kostkalab/nc-gwassnps-score_manuscript`.

4.0 Information sharing between disease terms can improve the prioritization of non-coding genetic variants

4.1 Introduction

In the previous chapter, we introduced a disease-specific approach that can prioritize non-coding genetic variants in 111 different diseases. This approach outperforms current organism-level variant scores, tissue-specific scores, and another disease-specific approach, DIVAN; however, there is still room for improvement, considering its average precision of 0.151 across 111 diseases compared to a baseline average precision of 0.091. In addition, this disease-specific approach is only applicable to diseases with sufficient training samples, limiting its scope substantially.

The limitations mentioned above can be attributed, in part, to the scarcity of training risk variants for each disease. Among the 111 diseases we studied, only 17 diseases contain more than 1000 disease-associated variants and 70 diseases have less than 400 disease-associated variants. Additionally, over two thousand diseases in the GWAS Catalog contain fewer than 100 disease-associated SNVs, leading to their exclusion from our study due to the potential generation of inaccurate and unstable models.

On the other hand, studies have shown that diseases are related. For example, Cotsapas et al. discovered that seven immune system diseases not only share a similar genetic background but also exhibit shared uniquely expressed cell types [44]. Similarly, Wingo et al. found that some psychiatric and neurodegenerative diseases share genetic backgrounds and have highly expressed shared causal proteins in specific cell types and tissues [46]. In the previous chapter, we also observed that many diseases share similar models and highlight similar involved tissues. For instance, both Crohn's disease and ulcerative colitis models highlight the GI rectum and blood cell type while de-emphasizing brain tissue in the model (model sim = 0.84).

Given the relationships among diseases, one potential solution to the issue of inadequate positive training data could be to include risk SNVs in related diseases. Some researchers have developed computational models that can prioritize variants for a group of diseases. For example, eyeVarP is a computational tool that prioritizes genetic variants for various eye diseases, including glaucoma, corneal diseases, and more [50]. In another study, Chen et al. developed a non-coding variant prioritization method that effectively prioritizes variants for 19 autoimmune diseases [51]. Cao et al. developed CASAVA, a disease-category-specific variant prioritization method that effectively prioritizes variants for 24 different disease groups, such as cardiovascular diseases [52]. However, these studies typically group diseases based on disease categories without evaluating the effectiveness of grouping compared to considering individual diseases. In addition, they do not assess which metric is better for identifying related diseases.

Therefore, in this chapter, we propose an information-sharing approach that can share SNVs in two diseases and we systematically evaluate whether sharing information between different disease terms leads to improved variant prioritization performance. In this research, we leverage disease terms and SNVs from the GWAS Catalog and combine SNVs between terms with various disease-term-specific sample weights. We assess the performance improvement of variant prioritization, or lack thereof, using nested cross-validation. Our findings demonstrate that employing an information-sharing approach by combining SNVs from related diseases can enhance variant prioritization. Furthermore, we compare three different methods for identifying related diseases and find that utilizing model similarity derived from our previous chapter outperforms other approaches.

4.2 Methods

4.2.1 Diseases studied

Disease-associated and control variants were retrieved as described in the previous chapter (See **Section** 3.3.1), where we focused on 111 diseases that each contained 100 or more SNVs. For the purposes of the information sharing step, we obtained a subset of the diseases that are located at the relative bottom of the hierarchy plot within the 111 diseases based on The Experimental Factor Ontology (EFO, `https://www.ebi.ac.uk/efo/`). This ensured that no disease in the subset included another disease. This resulted in 68 diseases (**Appendix Table** B.2-**??**).

4.2.2 Pairwise information sharing model for two diseases

4.2.2.1 Model overview

In the information-sharing setup, we designate D_1 as the primary disease of interest and D_2 as an auxiliary disease introduced for potential information-sharing purposes. We aim to assess whether the inclusion of SNVs in D_2 can impact the performance of D_1. To achieve this, we gather all disease-associated single nucleotide variants (SNVs) and matched control SNVs for D_1; however, we exclude overlapping SNVs in D_2 that exist in D_1, so the combined SNVs collection does not contain any duplicate SNVs.

In this study, we utilized the regularized logistic regression algorithm, which was also employed in the previous chapter (refer to **Section** 3.3.2 for more details). The model was trained and evaluated using a five-fold cross-validation approach. Our training dataset comprised SNVs from both D_1 and D_2, while the test datasets exclusively contained SNVs from D_1.

During the training phase, we assigned weights to the SNVs from D_1 and D_2. This weight assignment serves two purposes: firstly, to prevent the model's performance from deteriorating significantly compared to using D_1 alone (see **Appendix Figure** B.1) and secondly, to determine the extent to which incorporating D_2 improves the model's performance. A higher weight is assigned to D_2 if it contributes positively, whereas a very small weight is assigned if D_2 significantly impairs the performance of D_1. The weight values are determined through a hyperparameter optimization process.

To establish a control scenario where no D_2 SNVs are added, we set the weight of D_2 to zero. Finally, we compare the performance of the information-sharing model containing SNVs in D_1,D_2 with that of the D_1 only model.

4.2.2.2 Weight setup in two diseases

In the disease pair consisting of D_1 and D_2, where D_1 is the disease of interest and D_2 is the auxillary disease, we define N_1 as the number of negative SNVs (control SNVs) in D_1, P_1 as the number of positive SNVs (disease-associated SNVs) in D_1, N_2 as the number of negative SNVs in D_2, and P_2 as the number of positive SNVs in D_2. We introduce w_{1n} as the weights for negative SNVs in D_1, w_{1p} as the weights for positive SNVs in D_1, w_{2n} as the weights for negative SNVs in D_2, and w_{2p} as the weights for positive SNVs in D_2. Furthermore, w_1 represents the overall weights for D_1 and w_2 represents the overall weights for D_2. Consequently, we obtain the following relationships:

$$w_{1p} \times P_1 + w_{1n} \times P_1 = w_1$$

$$w_{2p} \times P_1 + w_{2p} \times P_1 = w_2$$

To ensure equal weights for positive and negative SNVs within each disease, we set up the following equations:

$$w_{1p} \times P_1 = w_{1n} \times N_1$$

$$w_{2p} \times P_2 = w_{2n} \times N_2$$

Solving these equations, we find:

$$w_{1n} = \frac{w_1}{2N_2}$$
$$w_{1p} = \frac{w_1}{2P_2}$$
$$w_{2n} = \frac{w_2}{2N_2}$$
$$w_{2p} = \frac{w_2}{2P_2}$$

We define w as the ratio $w = \frac{w_2}{w_1}$, and we will optimize it using the weight hyperparameter optimization process. In our study, the emphasis is on the ratio w not the individual values of w_1 or w_2, because the model remains the same if both are proportionally increased or decreased.

4.2.2.3 Weight hyperparameter optimization

To determine the optimal hyperparameter (w) for our model, we employed a nested five-fold cross-validation approach. The entire dataset was randomly divided into five outer folds, each containing an equal ratio of positive and negative single nucleotide variants (SNVs) and an equal ratio of SNVs from D_1 and D_2. Within each outer fold, the training data was further divided into five inner folds, maintaining the same ratio of positive and negative SNVs and the same ratio of SNVs from D_1 and D_2. Notably, the SNVs from D_2 were excluded in both the inner and outer fold's test datasets.

Within the inner loop, we performed hyperparameter optimization for the weight parameter w using a grid search technique. We explored 11 different values of w including 10^{-5}, 10^{-2}, $10^{-1.5}$,10^{-1}, $10^{-0.5}$, 10^{0}, $10^{0.5}$, 10^{1}, $10^{1.5}$, 10^{2}, and 10^{5}. This process was repeated twice to reduce the impact of random variations in the data and obtain a more reliable estimate of the optimal w value. Consequently, we obtained 10 sets of validation performance, and for each set, we selected the weight with the highest area under the precision-recall curve. Among the 10 folds, we discarded the two highest and two lowest weights, retaining only the six weights in the middle. Finally, we calculated the mean of these six weights to obtain the optimized weight for that specific inner loop. See **Appendix Figure** B.2- B.10 for the performance of the validation set in the inner fold for example diseases.

In the plots above and Section 4.2.2.4, we experimented with 11 different lambda values spanning from 10^{-5} to 10^{5}. However, to prevent D_2 from being disproportionately weighted compared to D_1—thereby dominating it—we set the maximum weight (w) in the optimization process at 10^{1}.

By conducting this hyperparameter tuning procedure, we can identify the optimal w hyperparameter for our model. The outer loop of the nested cross-validation was then utilized to evaluate the model's performance on the test set.

4.2.2.4 Model on lambda: GAM

To identify the optimal hyperparameter w, we employed a grid search approach using 8 different values of w. However, the corresponding regularization parameter lambda can vary significantly with occasional outliers (see **Figure** 4.1 for example). This could result in the validation area under the curve (AUC) fluctuating (**Figure** 4.2a), not necessarily due to changes in w, but rather due to the outlier value of lambda (for example, we can observe outliers in fold 3, 4 and 5).

To address this issue, we constructed a Generalized Additive Model (GAM) regression approach to model the non-linear relationship between the hyperparameter w and lambda values. Here, we choose a smooth function with a cubic spline basis (bs = "cs"), 4 knots (k = 4) and a smooth parameter 50 (sp = 50), in order to result in a relatively smooth curve. We then used the fitted lambda values to generate predictions on the validation dataset. This approach allows us to account for the complexity of the lambda change due to w, while producing a relatively stable AUC result that reflects the impact of w alone. Here, we show an example of the validation performance before and after applying the GAM model on weight and lambda (**Figure** 4.2) for disease adult-onset asthma (D1) and ulcerative colitis (D2). We can see that the validation AUC is smoother after applying the model, representing the trend of the AUC on weight parameter w rather than the outlier of lambda.

4.2.3 Model performance

For each disease, we performed two repetitions of five-fold cross-validation, resulting in ten test sets (called repeated cross-validation, see [118] for more details). In each test set, we evaluated the performance using the Area Under the Precision-Recall Curve (AUPRC) in three settings. Firstly, we utilized the optimal weight for combining D_1 and D_2 (model performance referred to as $PR_{D_1D_2}$). Secondly, we set the weight to 0 ($w = 0$) to mimic a control scenario where no D_2 is added (referred to as PR_{D_1}). Lastly, we set the weight to 1 ($w = 1$) to assign equal weight to D_1 and D_2, thereby simulating a scenario without a weight selection process (referred to as $PR_{D_1D_2\text{-}w1}$).

We then computed the relative performance using the following equation:

$$Relative\ Performance = \frac{\overline{PR_{D1D2}} - \overline{PR_{D1}}}{\overline{PR_{D1}}}$$

where the $\overline{PR_{D_1D_2}}$ and the $\overline{PR_{D_1}}$ represent the average precision of $PR_{D_1D_2}$ and PR_{D_1} across

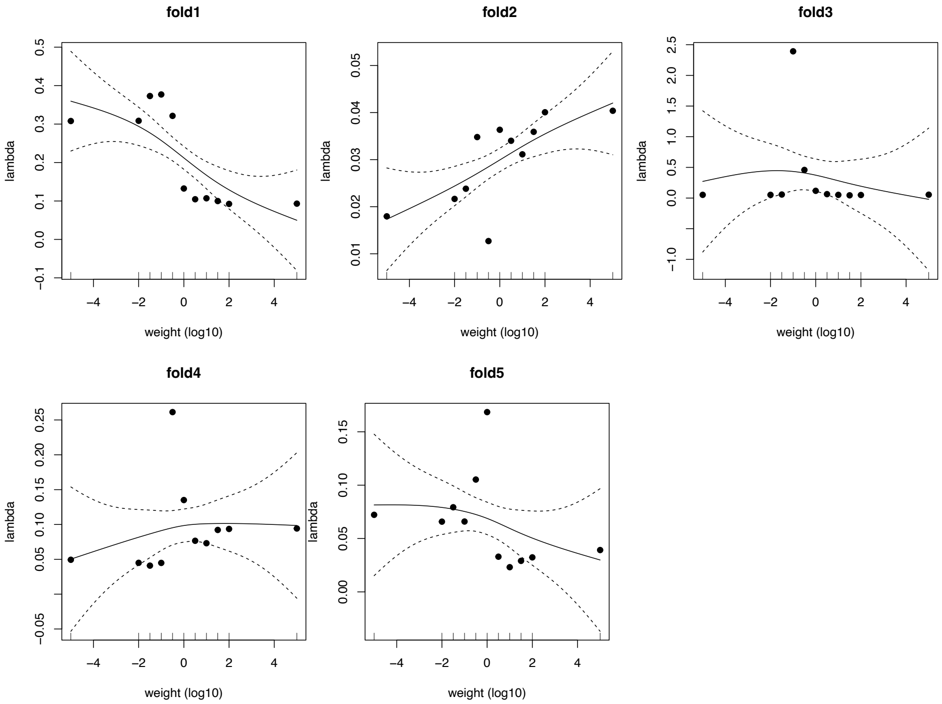

Figure 4.1: *An example of lambda value varies in different weight values in five different folds.* The figure illustrates the variability of lambda values across different weight values (w)in five different folds. The example focuses on two diseases: adult-onset asthma (D1) and ulcerative colitis (D2). Each plot corresponds to a different outer fold. On the x-axis, we have the logarithm base 10 of the weight values ($log10(w)$), while the y-axis represents the hyperparameter lambda in regularized logistic regression. The solid line represents the regression line, and the dashed line indicates the 95% credible intervals.

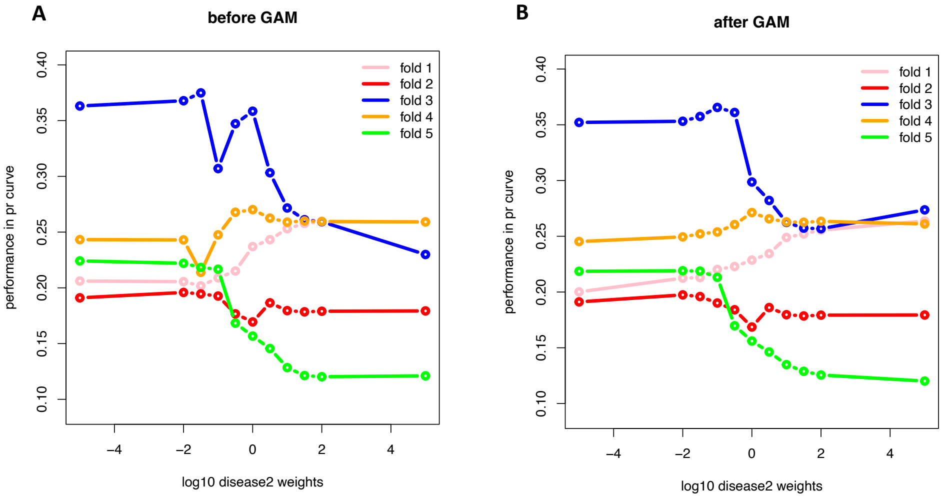

Figure 4.2: *The performance of validation set in five folds before and after applying GAM model.* The figure illustrates the performance of the validation set in five folds, both before and after applying the Generalized Additive Model (GAM). The example focuses on two diseases: adult-onset asthma (D1) and ulcerative colitis (D2). Each color represents a distinct fold in the inner loop. The x-axis represents the logarithm base 10 of the weight values ($log10(w)$), while the y-axis depicts the model's performance in terms of the Area Under the Precision-Recall Curve (AUPR) on the test set.

the 10 test sets. Relative performance greater than 1 indicates improved performance when incorporating D_2, whereas relative performance less than 1 suggests adding D_2 deteriorates the performance.

4.2.4 Disease similarity measurements

Model similarity

The model similarity is measured using the similarity of the beta coefficients from our disease-specific model from the previous chapter. See **Section** 3.4.6 for more details.

Genetic correlation

Genetic correlation is obtained from the GWAS Atlas [92], where it measured the shared genetic background of two diseases using GWAS summary statistics. See **Section** 3.3.4 for more details.

Semantic similarity

Semantic similarity, which measures the disease similarity based on the controlled biological vocabularies, such as Medical Subject Headings (MeSH), is calculated using Wang's method [119]. We retrieved Wang's method similarity using the meshes package developed by Yu et. al. [120]. In this process, we mapped the disease terms from the Experimental Factor Ontology (EFO) to corresponding MeSH terms. The mapping was performed using the EMBL-EBI Ontology Xref Service (OxO) available at `https://www.ebi.ac.uk/spot/oxo/`.

4.3 Results

4.3.1 Information sharing model can improve variant prioritization

4.3.1.1 Including related diseases can improve performance through information sharing model

The information-sharing model is built on the disease-specific regularized logistic regression model discussed in the previous chapter. We employed the same disease-associated and control variants, along with tissue-specific scores represented by DNase I hypersensitivity (DHS) profiles (see **Section** 3.3.1, 3.3.2 and 3.4.3 for more details). In this study, we focused on a subset of 111 diseases from the previous chapter that do not exhibit a parent-child relationship (for example, cancer is the parent and colorectal cancer is the child), resulting in a final set of 68 diseases. Within the information-sharing model, we defined D_1 as the disease of interest and D_2 as the auxiliary disease added to facilitate information sharing. To control the weight and importance assigned to D_1 and D_2, we developed a weighting scheme for positive and negative SNVs from both diseases (See **Section** 4.2.2.3). This weighting scheme involves assigning distinct overall weights to D_1 and D_2 (w_1 and w_2). By optimizing the relative weight $w = w_1/w_2$ through a hyperparameter optimization process, we were able to determine the appropriate weight for D_1 and D_2 for effective information sharing.

To assess the performance improvement achieved in the information-sharing model, we measured the relative performance using the area under the precision-recall curve across 10 test folds (See **Section** 4.2.3). The relative performance quantifies the performance gain (or loss) of the information sharing model for each disease pair (referred to as D_1D_2) compared to the model utilizing only D_1. **Appendix Data** B.1.1 summarize the relative performance of the information-sharing model in all pairs.

To visualize the relative performance between disease pairs, we plotted a heatmap using 20 sample diseases D_2 for each disease D_1. The selection of these samples was based on their model similarity, including the top 10 most similar diseases, followed by the middle 5 (ranking 18, 27, 36, 45, 54), and the least 5 similar diseases. The heatmap (**Figure** 4.3) reveals that the information-sharing model exhibits improvements in certain disease pairs across many diseases, as indicated by the redder color. Notably, diseases like multiple myeloma and squamous cell lung carcinoma demonstrate substantial performance improvements when incorporating specific D_2 diseases. Juvenile idiopathic arthritis also shows some improvement, albeit to a lesser extent. On the other hand, diseases like ankylosing spondylitis exhibit limited improvements through this information-sharing model. Additionally, we observed that disease pairs with higher model similarities tend to achieve greater relative performance increases compared to those with middle and least model similarities (**Figure** 4.3).

We performed the Wilcoxon signed-rank test for the disease pairs with top 10 model

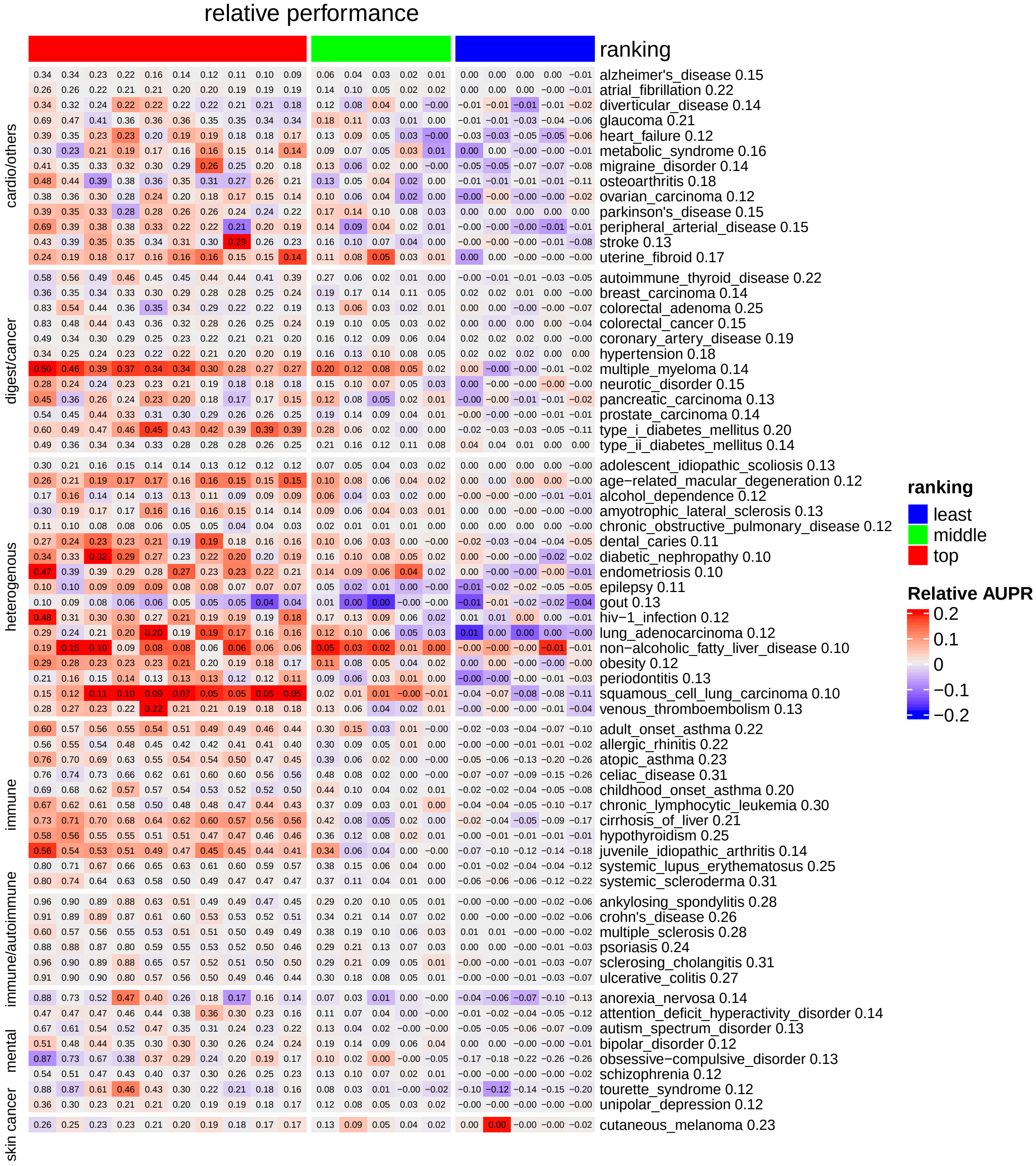

Figure 4.3: *Relative performance of disease pairs using an information sharing model.*

(description next page)

Figure 4.3: Continued figure legend for figure 4.3.

(Previous page) Color in the heatmap represents the relative Area Under the Precision-Recall Curve (AUPR) after applying the information-sharing model, while numbers indicate the model similarity of the two diseases. Rows correspond to Disease 1 (D_1), and columns represent Disease 2 (D_2), where each row represents a distinct disease. The heatmap displays a subset of D2 diseases based on their model similarity ranking: the top 10, middle 5, and least 5. The number next to the disease name in each row represents the baseline performance of D_1 without information sharing model. Disease grouping is based on the clusters discussed in **Section** 3.4.6.

similarity to test whether adding D_2 significantly increases (or decreases) the performance. We observed that out of 680 disease pairs, 102 of them have p value less than 0.05, and 36 of those exhibited a relative increase greater than 10%. After applying the false discovery rate (FDR) adjustment, 13 pairs remained significant with adjusted p-values less than 0.1 (See **Appendix Data** B.1.2).

To further visualize the results, we utilized a categorical scatter plot where the relative performance is shown on the y-axis. The x-axis represents D_1 diseases and is sorted by the number of SNVs in D_1. The plot demonstrates that the information-sharing model improves the performance of certain disease pairs (**Figure** 4.4). Moreover, disease pairs with higher model similarities (indicated by red) generally exhibit higher relative performance, whereas those with lower similarities (yellow or green) have lower or negative relative performance. Additionally, we noticed that diseases with a higher number of SNVs in D_1 (ranking of the x-axis) tend to show less pronounced improvements compared to diseases with fewer SNVs. However, the number of SNVs in D_2 (the point size) does not seem to have an obvious impact on the performance.

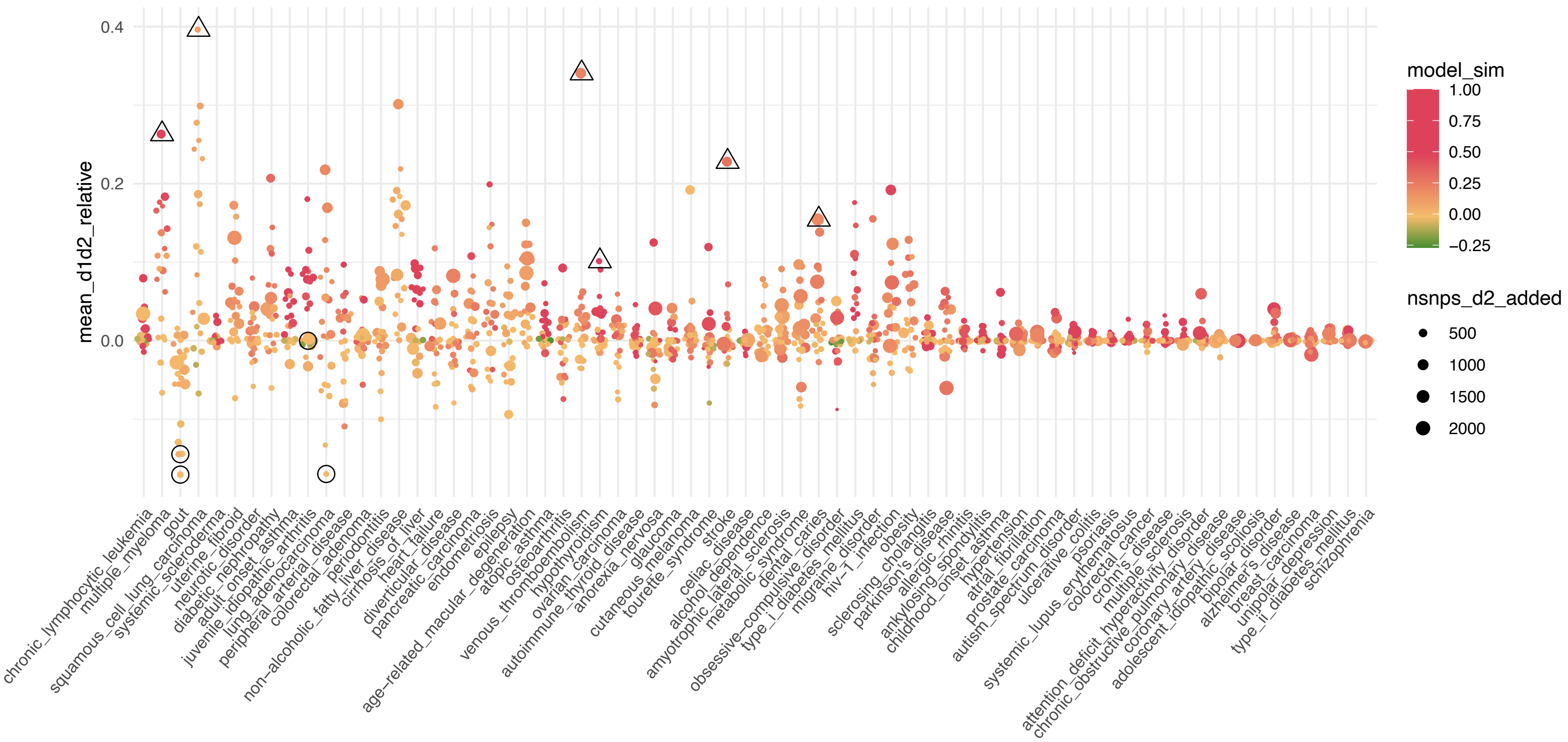

Figure 4.4: *Information sharing model increases the performance for some disease pairs.* The x-axis represents disease D_1, sorted by the number of SNVs in D_1, while the y-axis represents relative performance. Dots represent disease pairs with D_2 samples (top 10, middle 5, and least 5). The dot color indicates model similarity, and the dot size represents the number of SNVs in D_2 added. Disease pairs that discussed are highlighted with a black triangle (increased after applying the model) or black circle (no improvement or decreased after applying the model).

In addition, we calculated the average improvement for each disease D_1 when considering the top 10 most helpful diseases D_2. Our analysis revealed that lung carcinoma, diabetic nephropathy, multiple myeloma, and alcoholic liver disease experienced the greatest benefits from the information-sharing model, with average increases ranging from 36% to 20% **Figure** 4.5, red box). However, the improvement for diseases such as coronary artery disease was not as pronounced.

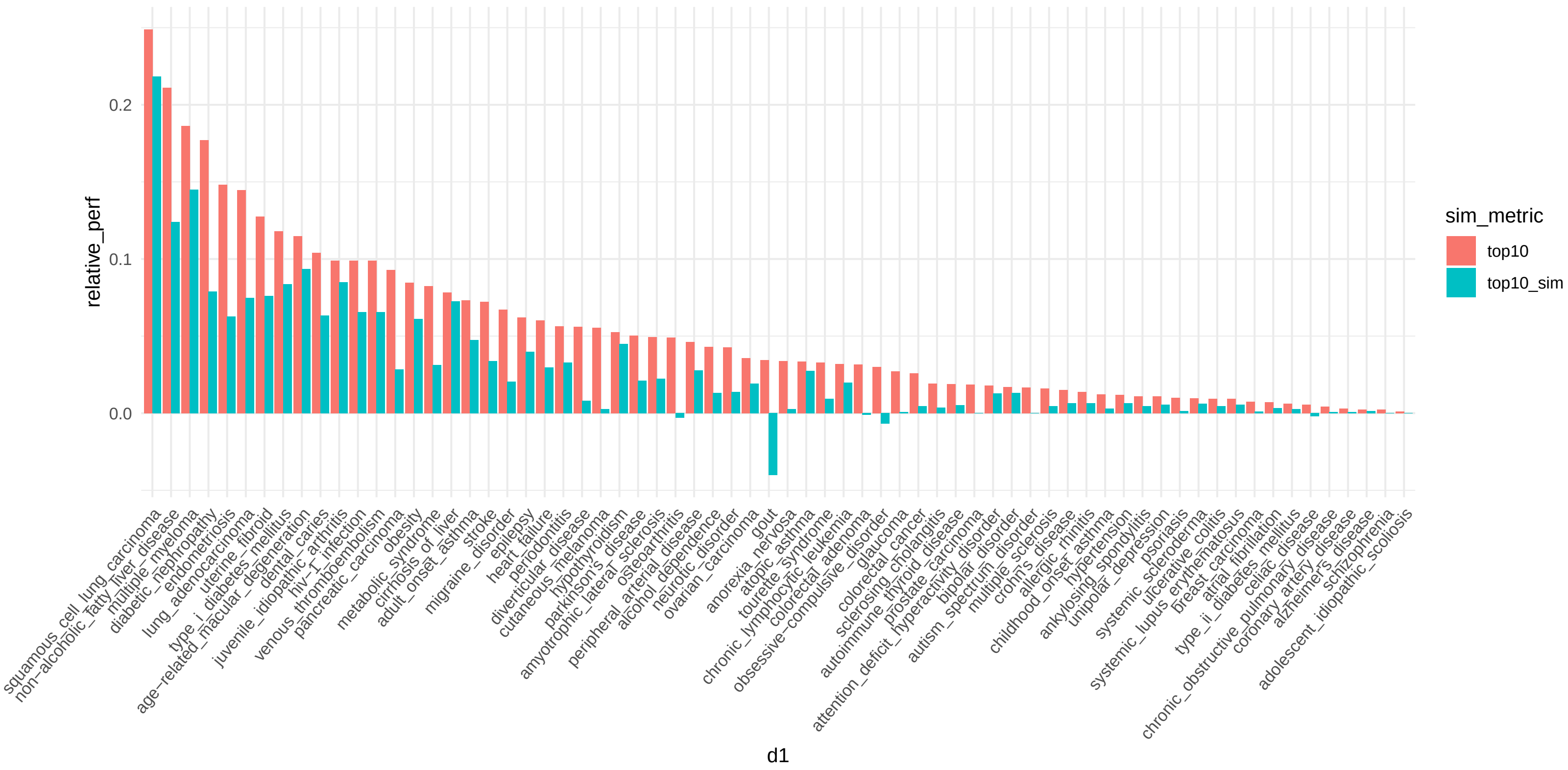

Figure 4.5: *Relative performance of the information sharing model considering the top 10 disease.* The red bar indicates top 10 diseases selected by the diseases with the highest performance. The blue bar indicates top 10 diseases selected by diseases with the highest similarity with D_1.

4.3.1.2 Information sharing model reveals biological relevance in disease pairs

Next, we examined several disease pairs that exhibited a significant increase or decrease (or no change) when applying the information sharing model (marked triangle or circle in **Figure** 4.4 and summarized in **Table** 4.1 and 4.2).

When incorporating SNVs in celiac disease to hypothyroidism, the performance of hypothyroidism improved by approximately 10%, despite hypothyroidism already having a relatively higher average precision of 0.25 on its own (**Table** 4.2). While celiac disease is a digestive immune disorder [121] and hypothyroidism is caused by a lack of the thyroid hormone [122], both of them are a type of immune disease, and they have high model similarity ($s_m = 0.56$). In addition, studies also found they coexist in patients [123]. It is worth noting that despite their shared model similarity and clinical relationship, these two diseases only share 4 overlapping SNVs. This implies that celiac disease presents a good information-sharing candidate for hypothyroidism by introducing an additional 184 unique SNVs.

Another example pair worth considering is multiple myeloma and multiple sclerosis. Despite being a type of white blood cell cancer [124] and a nervous system disease [125], respectively, these two conditions exhibit a 50% model similarity. Notably, multiple myeloma demonstrates relatively poor performance (AUPR = 0.14) when trained alone (**Table** 4.1). Multiple myeloma has 107 associated SNVs; in contrast, multiple sclerosis contains 503 disease-associated SNVs. Furthermore, previous studies such as [126] and [127] have explored the association between these disorders, providing additional support for the similar disease model and the potential benefit of utilizing multiple myeloma data to enhance multiple sclerosis predictions.

Another example pair to consider is squamous cell lung carcinoma and hypothyroidism. Despite their relatively low model similarity ($s_m = 0.10$), they achieve high relative performance (40%) (**Table** 4.1). Although the absolute value of disease similarity between them is not high ($s_m = 0.10$), hypothyroidism ranks fourth in terms of similarity to squamous cell lung carcinoma. While research has found biological relationships of hypothyroidism to certain cancers such as breast cancer [128, 129, 130], as far as we are concerned, we have not found any studies indicating a relationship between squamous cell lung carcinoma and hypothyroidism.

The information-sharing model improves the performance in many disease pairs; however, it has come to our attention that it leads to a decrease in performance in certain disease pairs (as shown by the circle in **Figure** 4.4, summarized in **Table** 4.2). Gout serves as an example in this case. Gout is a form of arthritis [131] and it exhibits low model similarity with almost all other diseases studied (with the highest model similarity being 0.1). Including SNVs associated with Parkinson's disease results in a decreased performance of 17% and metabolic syndrome results in a decrease of 14%. Both Parkinson's disease and metabolic syndrome have very low model similarity with gout (0.00 and 0.04, respectively), and the information sharing model assigns them relatively low weight ($10^{-1.13}$ and $10^{-1.11}$). In addition, we observe that adding almost all other diseases does not improve the model performance of gout (See **Figure** 4.4). This can be partially attributed to the low model similarity between gout and other diseases, although the exact reason remains unclear.

In addition, we would like to note another disease pair: juvenile idiopathic arthritis and anorexia nervosa. These two diseases belong to different disease groups, with juvenile idiopathic arthritis being an autoimmune disease that shows arthritis in children [132] and anorexia nervosa being an eating disorder [133]. They possess very dissimilar model similarity (s_m = -0.10) (**Table** 4.2). During the information-sharing process, we assigned a very low weight to anorexia nervosa ($10^{-3.5}$), resulting in a performance similar to using only juvenile idiopathic arthritis (0.0%). This example emphasizes the purpose of the weight-tuning process in the information-sharing model.

4.3.2 Factors influencing the performance of the information sharing model

As previously observed, certain factors, including the number of SNVs in D_1 and D_2 and model similarity, can affect the performance of the information-sharing model. Consequently, in this and the next section, we conducted a comprehensive investigation into the factors that could influence the relative performance of the model.

We first examined the number of SNVs in D_1. For each disease in D_1, we calculated

D_1	D_2	D1perf	impr	weight*	s_m[a]	nD_1[b]	nD_2[c]
squamous cell lung carcinoma	hypothyroidism	0.09	40%	0.14	0.10	109	174
venous thromboembolism	Alzheimer's disease	0.13	34%	0.62	0.22	173	940
multiple myeloma	multiple sclerosis	0.14	26%	0.69	0.50	107	583
stroke	coronary artery disease	0.13	23%	0.76	0.28	187	844
dental caries	unipolar depression	0.10	15%	0.26	0.19	207	1362
hypothyroidism	celiac disease	0.25	10%	0.58	0.56	174	188

[a] Model similarity, [b] The number of SNVs in D1, [c] The number of SNVs in D2

* All weight values are presented in logarithmic scale (log10).

Table 4.1: *Examples of disease pairs where information sharing model increases the performance.*

D_1	D_2	D1perf	impr	weight*	s_m[a]	nD_1[b]	nD_2[c]
gout	metabolic syndrome	0.14	-14%	-1.13	0.04	109	205
gout	Parkinson's disease	0.14	-17%	-1.11	0.00	109	244
juvenile idiopathic arthritis	anorexia nervosa	0.14	0%	-3.50	-0.10	124	182

[a] Model similarity, [b] The number of SNVs in D1, [c] The number of SNVs in D2

* All weight values are presented in logarithmic scale (log10).

Table 4.2: *Examples of disease pairs where information sharing model do not improve the performance.*

the average performance of the top 10 diseases. Our analysis revealed that the information-sharing model demonstrated noticeable performance improvements for diseases with fewer than 300 SNVs, while the improvement was less pronounced for diseases with more than 300 SNVs (refer to **Figure** 4.6 A). A similar trend was observed when we ranked the diseases based on model similarity (see **Figure** 4.6 B).

Furthermore, we explored the baseline performance of diseases in D_1. We discovered that the information-sharing model exhibited better results for diseases with lower baseline performance (below 0.15). However, diseases with moderate performance (e.g., around 0.2 AUPR) also demonstrated the potential for improvement. For instance, type I diabetes mellitus has an average improvement of 11.8% among the top 10 diseases and 8.3% among the top 10 similar diseases (refer to **Figure** 4.7).

We also investigated how the number of SNVs in D_2 could influence the performance. For each disease in D_1, we examined the correlation between the relative performance and the number of SNVs in D_2. We found that certain diseases, such as atrial fibrillation, exhibited a correlation between the number of SNVs in D_2 and the relative performance. However, for the majority of diseases (62 out of 68), we did not observe a significant correlation between the number of SNVs in D_2 and the relative performance (See **Appendix Figure** B.11 - B.12 and **Appendix Table** B.2 - B.2).

4.3.3 Model similarity can help find similar diseases that improve the performance

In the previous section, we explored some factors (e.g. nSNVs in D_1) that may influence the information-sharing model performance. In this section, we continue to investigate how model similarity can affect relative performance and whether it can serve as an effective criterion for efficiently selecting related diseases compared with other disease similarity metrics.

We first explore whether an increase in model similarity correlates with improved performance. Across all diseases, a slight increase in relative performance was observed with higher model similarity (see **Appendix Figure** B.13). To further investigate this trend,

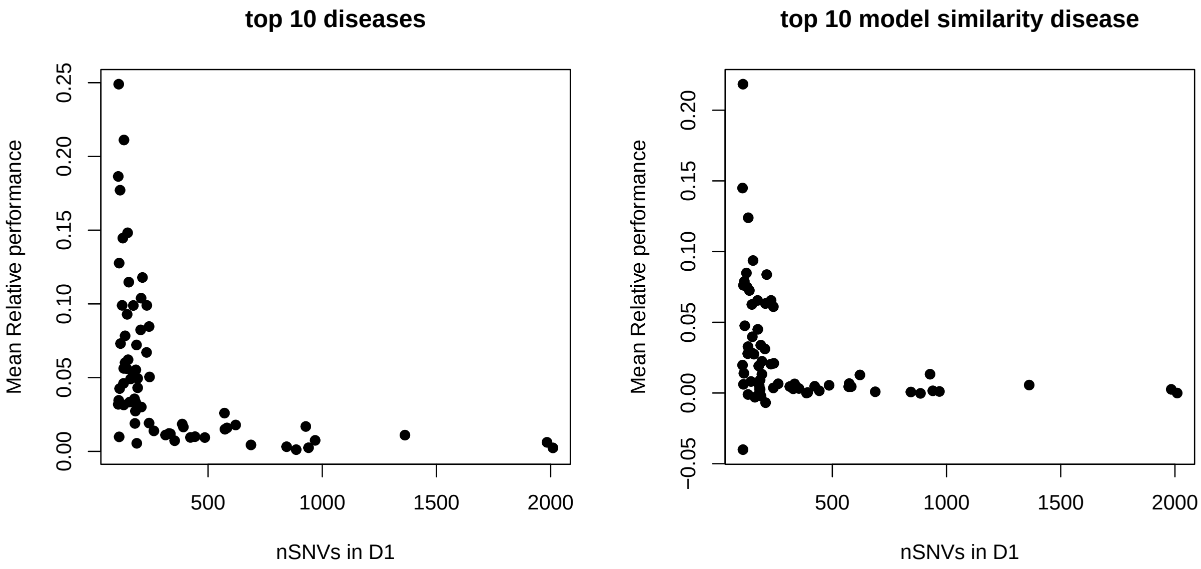

Figure 4.6: *Number of SNVs in D_1 can influence the information sharing model performance.* The x-axis is the number of SNVs in D_1 and the y-axis is the mean of the average precision of the top 10 related diseases with the highest performance (**left panel**), or the top 10 diseases with the highest model similarity (**right panel**).

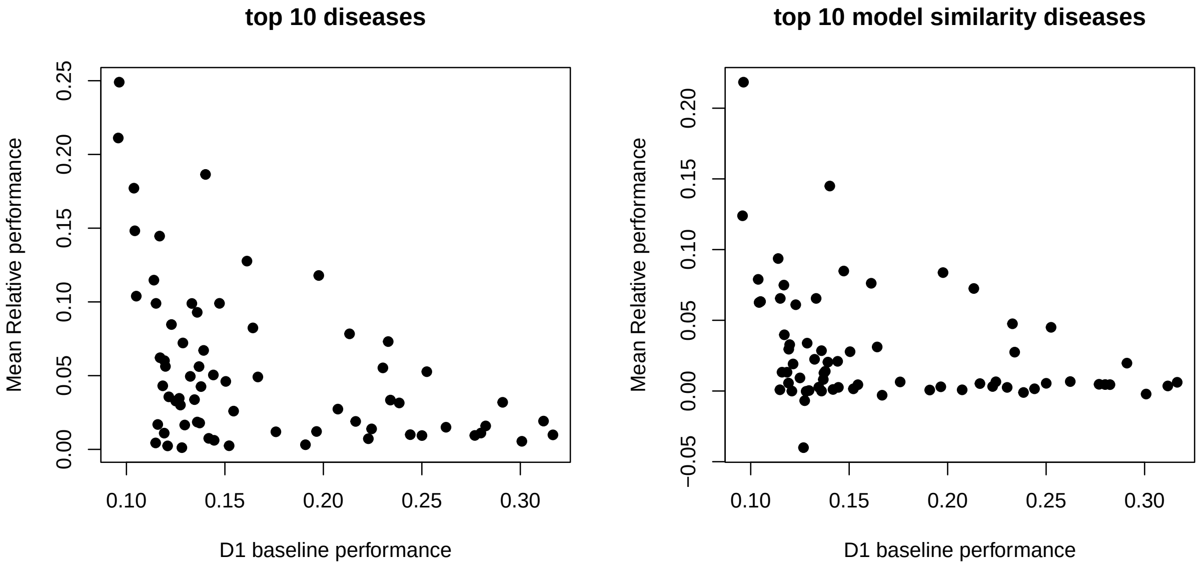

Figure 4.7: *Baseline performance of D_1 can influence the information sharing model performance.* X axis is D_1 baseline performance and y-axis is the mean of the average precision of the top 10 related diseases with the highest performance (**left panel**), or top 10 diseases with the highest model similarity (**right panel**).

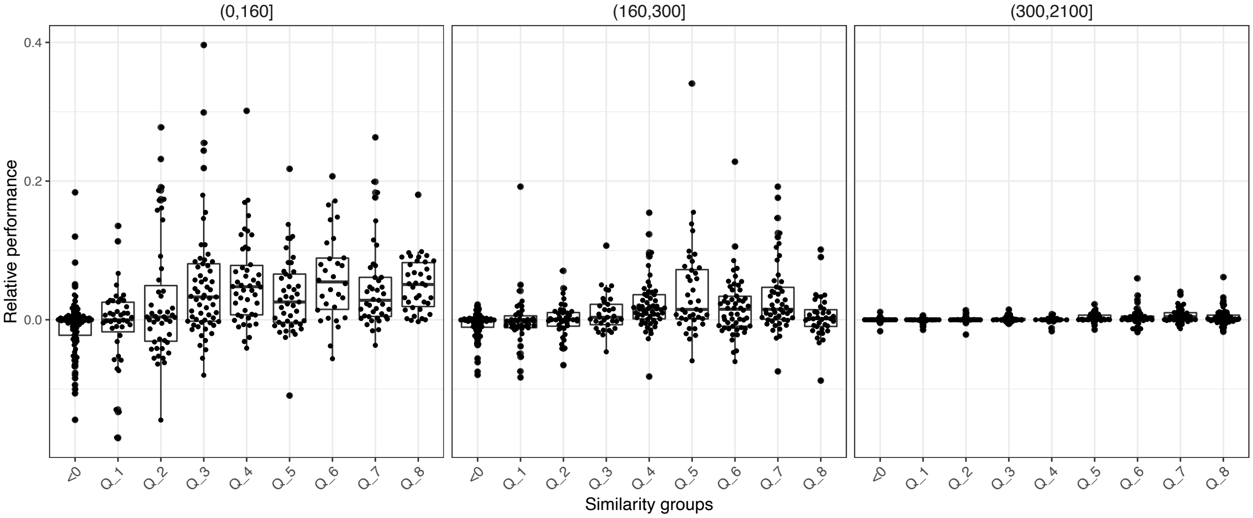

Figure 4.8: *Impact of model similarity on the relative performance of the information sharing model.* The x-axis represents the quantiles of model similarity (or model similarity is less than 0). Y axis is the relative performance of the information-sharing model. The boxplot summarizes the first quantile, medium and the third quantile of the data. The data are grouped based on the number of SNVs in D1.

disease groups with a similar number of SNVs in D_1 were analyzed. For disease pairs with fewer than 160 SNVs in D_1, an increase in model similarity corresponded to improved relative performance. In cases where D_1 contained SNVs ranging from 160 to 300, relative performance increased with higher model similarity, but it decreased in the highest quantile. However, when D_1 contained over 300 SNVs, there was a minimal change (See **Figure** 4.8).

Further analysis focused on the relationship between relative performance and model similarity within individual diseases. Some disease pairs demonstrated a strong correlation between relative performance and model similarity, such as hypothyroidism (correlation = 0.736, adjusted p-value = 4e-11), juvenile idiopathic arthritis (correlation = 0.736, p-value = 4e-11), and obesity (correlation = 0.635, p-value = 4.8e-8) (see **Table** 4.3.3 and **Figure** 4.9 for example diseases). Conversely, certain diseases, like coronary artery disease and breast carcinoma, exhibited less obvious correlations (see **Appendix Table** B.2- B.2).

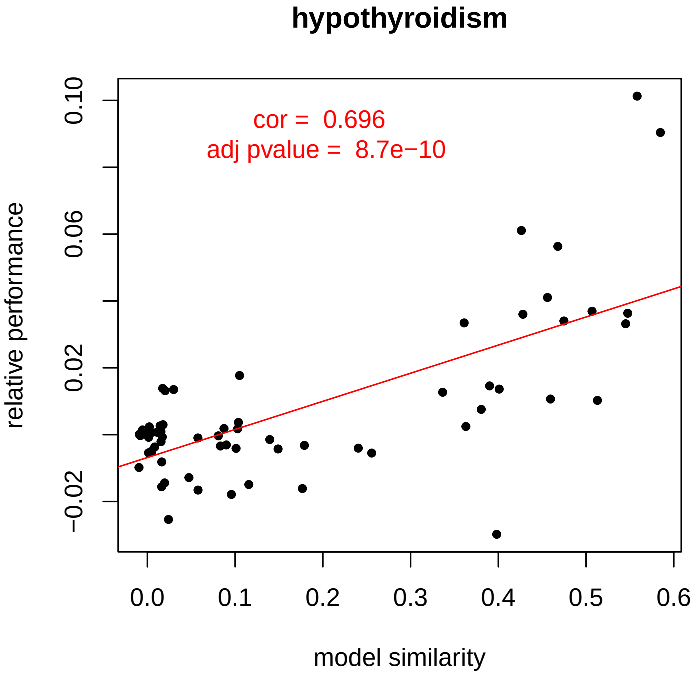

Figure 4.9: *Model similarity is highly correlated with relative performance for Hypothyroidism.* The x-axis represents the model similarity across all disease pairs, with D_1 being Hypothyroidism. The y-axis is the relative performance.

D1	Corre	pval	adj p val
juvenile_idiopathic_arthritis	0.829	4.6e-18	3.13e-16
cirrhosis_of_liver	0.764	5.77e-14	1.96e-12
multiple_myeloma	0.737	1.18e-12	2.67e-11
type_i_diabetes_mellitus	0.721	6.04e-12	1.03e-10
hypothyroidism	0.696	6.38e-11	8.68e-10
age-related_macular_degeneration	0.629	1.16e-08	1.31e-07
obesity	0.618	2.46e-08	2.39e-07
bipolar_disorder	0.614	3.32e-08	2.82e-07
atopic_asthma	0.606	5.53e-08	4.18e-07
hiv-1_infection	0.583	2.21e-07	1.5e-06

Table 4.3: *Correlation of relative performance and disease similarity for example diseases.* 10 diseases with the highest correlation were selected. A false discovery rate adjustment was applied to the p values to account for multiple comparisons.

Our next objective is to assess a more effective approach for the rapid identification of related diseases by comparing different disease similarity metrics. We compared our model similarity approach with two other disease similarity metrics: genetic correlation, which assesses shared genetic factors between diseases [113], and semantic similarity, which gauges shared clinical or biological characteristics within the ontology tree [120]. For each disease marked as D_1, we selected the five diseases with the highest similarity scores in each of these three metrics and subsequently analyzed their mean relative performance. Our findings reveal that model similarity outperforms both semantic similarity and genetic correlation, with a particularly notable improvement in comparison to genetic correlation (See **Figure** 4.10 and 4.11).

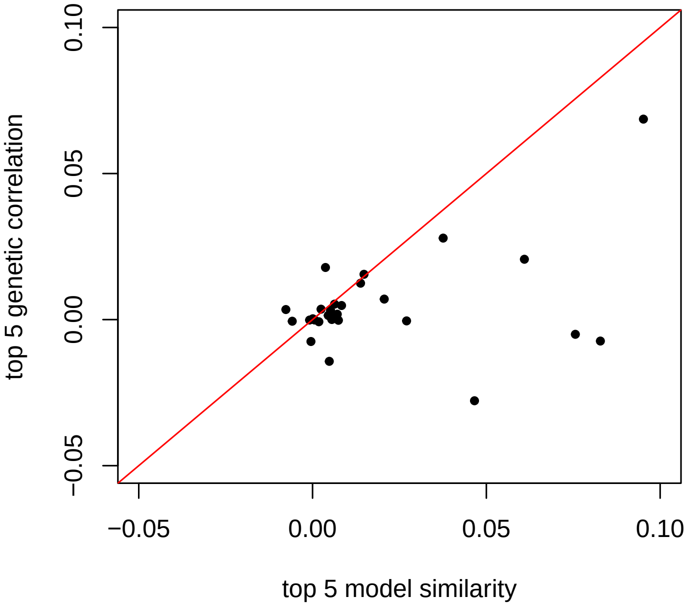

Figure 4.10: *Model similarity performs better than genetic correlation.* X-axis denotes the performance of each disease, measured as the mean performance of 5 diseases exhibiting the highest genetic correlation. Y-axis denotes the performance of each disease, measured as the mean performance of 5 diseases having the highest model similarity.

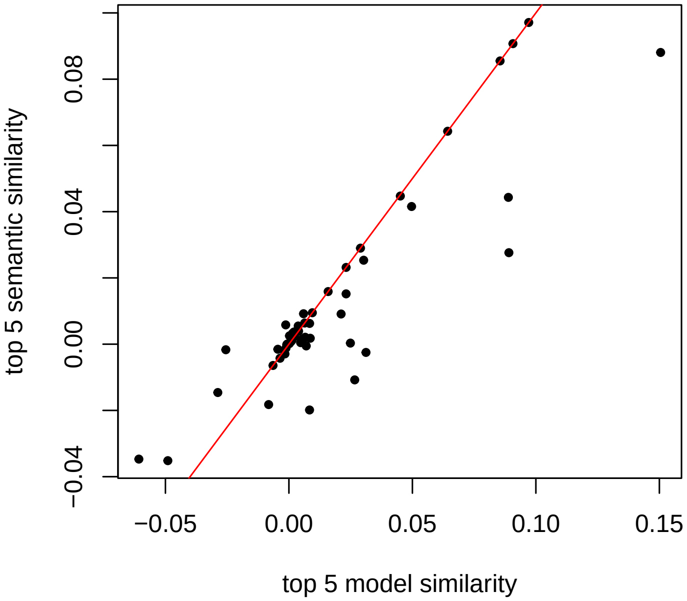

Figure 4.11: *Model similarity performs better than semantic similarity.* X-axis denotes the performance of each disease, measured as the mean performance of 5 diseases having the highest semantic similarity. Y axis denotes the performance of each disease, measured as the mean performance of 5 diseases having the highest model similarity.

4.3.4 Heterogeneous and immune disease group show higher improvements than other disease groups

Finally, we investigate which disease categories can benefit the most substantially from the information-sharing model. We used the disease group information introduced in the previous chapter (as detailed in **Section** 3.4.6). We calculated the mean relative performance of each disease, taking the 10 most similar diseases. We then identified 10 diseases with highest relative performance (defined later as top-level improvement). From **Table** 4.4, we observe that 5 diseases (30%) within the heterogeneous group experience substantial benefits from the information sharing model, with an average improvement of 12%. This follows diseases in the immune group, with 2 diseases (18%) demonstrating substantial improvements attributable to the model, resulting in an 8% average improvement. The Digest/Cancer group has 2 diseases displaying top-level improvements and an average enhancement of 11%. Lastly, one disease in the Cardio/Others category demonstrates an 8% improvement. The other three disease groups, immune/autoimmune, mental, and skin cancer do not have any diseases having top-level improvement.

Group	# Diseases	Top-level improvement	
		# Diseases	Ave imprv
Heterogeneous	17	5 (30%)	12%
Immune	11	2 (18%)	8%
Digest/Cancer	12	2 (17%)	11%
Cardio/Others	13	1 (8%)	8%

Table 4.4: *Heterogeneous and immune disease groups shows higher improvements.* Ten diseases with top-level improvement were selected among 68 diseases (improvement measured by the mean relative performance of the 10 most similar diseases). The first column is the number of diseases in each group. The 2-3 columns are the number of top diseases in the group and the percentage. The last column is the mean average relative performance for the top diseases in the group by using the most 10 similar diseases.

4.4 Discussion

In Chapter 3, we developed a disease-specific variant prioritization method and applied it to 111 diseases in NHGRI GWAS catalog. In this chapter, we improved the disease-specific method by developing an information-sharing approach that can include SNVs from related diseases. We also showed that model similarity is a better way to select related diseases compared to other disease similarity metrics including genetic correlation and model similarity.

Within the information-sharing model, we employed a weight-tuning technique to get the relative overall weight, denoted as w, assigned to Disease 1 (D_1) and Disease 2 (D_2) ($w = w2/w1$). Notably, we observed that in some disease pairs, we can select an overall relative weight greater than 1. This means that the disease we added (D_2) can outweigh the disease of interest (D_1). For instance, in the case of multiple myeloma and multiple sclerosis, the relative weight was 4.8 and this information-sharing model leads to a performance

D_1	D_2	model_sim	nD_1	nD_2	nD_2added
colorectal cancer	colorectal adenoma	0.83	572	131	23
ankylosing spondylitis	sclerosing cholangitis	0.96	314	242	30
crohn's disease	sclerosing cholangitis	0.89	574	242	31
psoriasis	sclerosing cholangitis	0.88	443	242	31
ulcerative colitis	sclerosing cholangitis	0.90	423	242	31
autism spectrum disorder	anorexia nervosa	0.52	392	182	39
autism spectrum disorder	tourette syndrome	0.61	392	184	41
autism spectrum disorder	obsessive-compulsive disorder	0.67	392	208	53
crohn's disease	ankylosing spondylitis	0.89	574	314	60
ulcerative colitis	ankylosing spondylitis	0.90	423	314	61

Table 4.5: *Examples of disease pairs in the Figure 4.8 middle panel, Q_8 group.* Ten disease pairs were selected based on the nD_2 added. The smallest 10 in the largest quantile group were selected.

improvement of 26%. This could come from the reason that D_2 possesses significantly more SNVs than D_1 (nearly fivefold, 583 vs. 107, see **Table** 4.1), assigning a weight greater than 1 allows us to leverage the information from D_2, ultimately enhancing model performance.

Another key finding in this chapter is the effectiveness of model similarity in identifying related diseases. However, it's important to note that a minimum threshold of SNVs is necessary to train the model and calculate the model similarity between two diseases. In our study, we limited the inclusion of diseases to those containing more than 100 non-coding SNVs to enable model similarity calculations. For diseases with very few SNVs, e.g., only 10 SNVs, utilizing the model similarity metric becomes challenging. Nevertheless, we recognize that in such cases, semantic similarity can serve as a viable alternative, representing a solid strategy for identifying related diseases.

As depicted in **Figure** 4.8 (middle panel), we observed a slight decline in relative performance within the highest quantile of model similarity groups. This trend can be attributed to certain disease pairs in this group sharing SNVs, resulting in fewer additional SNVs when incorporating D2. For example, **Table** 4.4 illustrates examples from this quantile that exhibit high model similarity but limited SNVs added to D2 due to SNV overlap between D1 and D2.

In this paper, we introduced an information-sharing approach between two diseases to enhance disease-specific prioritization. By applying this approach, we can use it to derive disease-specific variant scores that achieve better performance than the scores from Chapter 3. Particularly this approach holds the potential to increase the performance for diseases (D_1) with relatively few disease-associated SNVs, such as those with fewer than 300 SNVs (**Figure** 4.8). To facilitate further research, we have compiled a table detailing diseases D_1 with under 300 SNVs alongside their corresponding D_2 that contribute the most substantial performance improvement **Appendix Table** B.2 and B.2. This can serve as a guide to help researchers to identify diseases to include in future studies

In future research, this information-sharing model can be extended to incorporate more than two diseases. Furthermore, the insights gained from our study regarding the effectiveness of model similarity in identifying related diseases can be applied in future work to select diseases in a disease group. Additionally, we currently focus on diseases with more than 100 associated SNVs, but in the future, we can broaden our scope to include diseases with a moderate number of associated SNVs (e.g., between 50 to 100) or even diseases with fewer associated SNVs (e.g., less than 50) to expand our knowledge.

5.0 Annotation and analysis of predicted escape nonsense-mediated mRNA decay (NMD) for human genetic variants

Section 5.1 - 5.4 from the following chapter has been taken from the manuscript: Jonathan Klonowski*, **Qianqian Liang***, Zeynep Coban-Akdemir, Cecilia Lo and Dennis Kostka, ”aenmd: Annotating escape from nonsense-mediated decay for transcripts with protein-truncating variants” where I am the co-first author with Jonathan Klonowski (See [134] and [135]). Please see **Section** 5.5 for author contributions and **Section** 5.6 for additional methods and results.

5.1 Introduction

Nonsense Mediated mRNA Decay (NMD) is a well-characterized, evolutionarily conserved quality-control mechanism that is essential for embryogenesis and other developmental processes, and it is known to play a role in human disease [136]. NMD guards against compromised transcripts by affecting their degradation vs. translation, including transcripts with variants that introduce premature termination codons (PTCs). PTC-causing variants where a resulting transcript is subject to NMD can exert loss-of-function (LOF) effects in case of haploinsufficiency, where transcripts from both chromosomes are required for normal protein function. For PTC-harboring transcripts that escape NMD, there are additional possibilities of dominant-negative (DN) or gain-of-function (GOF) effects, where the altered protein may interfere with the wild-type version (DN) or where it can possess an altered molecular function or activity domain (GOF). While molecular mechanisms of DN/GOF effects are generally less well understood compared with LOF effects [137] and the pathogenicity of NMD escaping variants is gene/transcript-specific, they do play a significant role in human disease [136, 138, 139, 66, 140, 141, 142, 143, 10, 9, 144].

Given the potential contribution of PTC variants with NMD escape in causing disease, significant new insights into mechanisms of disease pathogenicity can emerge from annotating

PTC-containing transcripts with a prediction about their escape from NMD [138, 141, 142, 143, 10, 9, 144, 145, 146, 147, 148, 149, 150, 151, 152, 153, 8, 154, 155, 156]. Using an exon-exon junction complex-dependent model of NMD, a notable fraction of PTC-harboring transcripts is predicted to escape NMD [66]. However, this model only describes about half of NMD-escaping human variation accurately, prompting the development of additional approaches for predicting transcript escape from NMD [136, 148, 8, 157, 158, 159, 160, 161, 162, 163, 56, 60, 164, 12, 165, 59, 57, 166]. Nevertheless, and despite the relevance of annotating PTC-causing variants with respect to a modified transcript's susceptibility to NMD, there is a lack of scalable and accessible software addressing that task comprehensively (i.e., for all types of PTC-causing variants). Therefore, we developed aenmd, a software tool for comprehensive annotation of PTC-causing variant-transcript pairs with a (predicted) escape from NMD. aenmd makes use of well-established and experimentally validated rules based on PTC location within a transcript's intron-exon structure [60, 12], and it integrates well into existing variant analysis pipelines. In the following, we describe aenmd in more detail and report statistics of NMD escape for PTC-causing variants in the Clinvar [167], gnomAD [168], and NHGRI-EBI GWAS catalog [169] resources.

5.2 Methods

5.2.1 Annotating escape from NMD

aenmd predicts escape from NMD for combinations of transcripts and PTC-generating variants by applying a set of NMD-escape rules, which are based on where the PTC is located within the mutant transcript. First, the location of the 5'-most (novel) PTC is determined, and then escape from NMD is predicted by the following five rules [60, 12]: Whether

- the PTC located in the last coding exon (last exon rule),
- the PTC located within d_pen bp upstream of the penultimate exon boundary (penultimate exon rule; default: d_pen = 50)

- the PTC located within d_css bp downstream of the coding start site (css rule; default: d_css = 150)
- the PTC located within an exon spanning more than 407bp (407 bp rule)
- the transcript is intronless (single exon rule)

See **Figure** 5.1**A**. Distances (in bp) are calculated using the PTC nucleotide closest to the coding start site (CSS) or exon boundary for the css and penultimate exon rules, respectively; variants are assumed to be left-normalized [170] (aenmd provides this functionality). Variants that overlap exon-intron boundaries or splice regions are not currently analyzed by aenmd. Variant-transcript pairs with a PTC conforming to any of the above rules will be annotated to escape NMD, but results for all rules are reported individually by aenmd; this allows users to focus on subsets of rules, if desired. aenmd is implemented in the R programming language [171], making use of the VariantAnnotation [172] and vcfR [173] packages for importing/exporting variants from/into vcf files, and the Biostrings [174] and GenomicRanges [175] packages for calculating rules. An index containing all PTC-generating SNVs is pre-calculated for a given transcript set and stored in a trie data structure for lookup, using the triebeard package. For non-SNV variants, alternative alleles for overlapping transcripts are explicitly constructed and assessed. This strategy allows us to assess frameshift variants where a PTC is produced downstream of the variant location, and it accounts for both the size and content of sequence insertions, deletions, and insertion-deletions.

5.2.2 Data on genetic variants and transcript models

We obtained gnomAD version v2.1.11liftover GRCh38, Clinvar version 20221211, and the NHGRI-EBI GWAS version 20220730, catalog from their respective download sites and annotated variants using aenmd. For our analyses we used transcript models from ENCODE version 105, where we focused on protein-coding transcripts on standard chromosomes that: (a) have an annotated transcript support level of one (or NA for single exon transcripts), and (b) have a coding sequence length divisible by three.

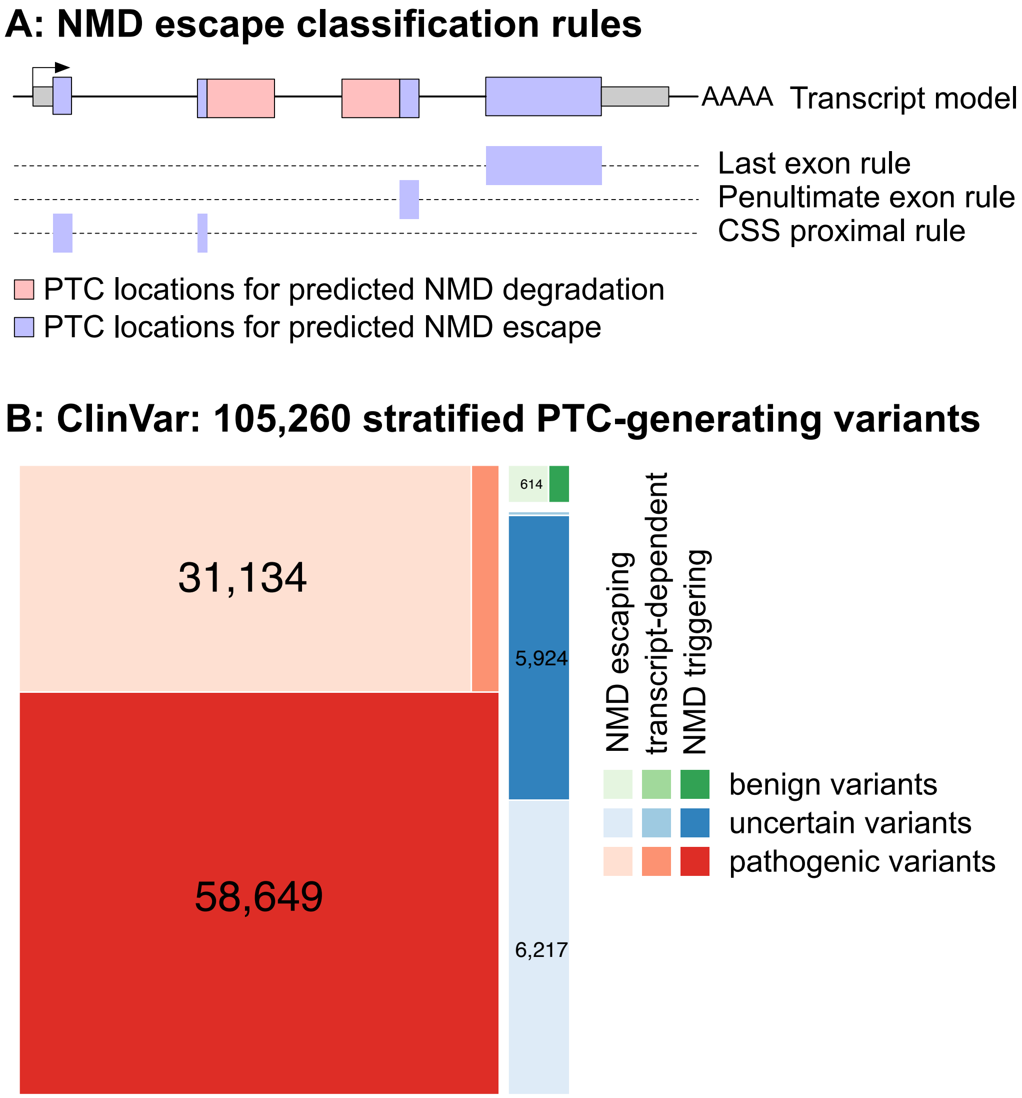

Figure 5.1: *NMD escape rules and clinvar variants.***Panel A** illustrates rules for predicting escape from NMD, with purple-shaded regions indicating areas that would harbor predicted NMD-escaping PTCs. The single exon rule and the 407 bp rule are not shown. **Panel B**: ClinVar variants, stratified by pathogenicity and annotated with predicted escape from NMD. transcript-dependent: the same variant overlaps multiple transcripts and has differing NMD escape predictions.

5.3 Results

5.3.1 aenmd R package

The aenmd R-package provides functionality to annotate variant-transcript pairs for predicted escape from NMD within the R ecosystem. Data dependencies (i.e., transcript models) are implemented via specific data packages (see below), and functionality for data import and export (vcf files) is also provided, as is functionality for variant left-normalization. Key differences that set aenmd apart from currently available tools for annotating escape from NMD are: all types of PTC-causing variants (including frameshift variants that do not cause stop codons at the variant site) are annotated, variants are annotated at scale, inserted sequence is considered for indels, and differentiated (i.e., rule-specific) output is provided for each transcript-variant pair where NMD-escape rules are applicable. This enables users to focus on the subset of rules most applicable to their situation; for example, some users may choose to focus on the exon-exon junction complex related “canonical” NMD rules only and ignore the “css proximal”, “single exon”, and “407 bp plus” rules (see **Section** 5.2). In addition to the R package we also provide a command line interface to aenmd‘s functionality.

5.3.2 aenmd_cli command-line interface

We constructed a containerized version of aenmd with all dependencies, which also provides a command-line interface. This allows end-to-end annotation of variants. An input vcf file is read, PTC-generating variants that overlap a specific transcript set (see the aenmd data packages section below) are annotated, and the annotation results are then included in the INFO column of an output VCF file. In this way, the aenmd_cli command line tool makes aenmd easily accessible and its results reproducible; there are no external dependencies, no knowledge of the R programming language is required, and it can be seamlessly integrated into existing variant processing workflows.

5.3.3 aenmd data packages

Annotation for (predicted) escape from NMD is based on the location of a PTC in the context of a transcript model. With aenmd, we provide precompiled annotation packages that provide comprehensive protein-coding transcript sets for the GRCh37 and GRCh38 assemblies of the human genome (data packages: aenmd.data.gencode.v43 and aenmd.data.gencode.v43.grch37, respectively), based on GENCODE version 43 annotations. We also provide a more stringently filtered transcript set based on ENSEMBL (version 105), containing transcripts with the highest level of transcript support (data package: aenmd.data.ensdb.v105). The aenmd package provides functionality to select between different transcript sets, allowing convenient prediction of NMD escape for GRCh37 and GRCh38 variants.

5.3.4 Annotation of gnomAD, Clinvar and the GWAS catalog

We used aenmd with a high-quality ENSEMBL transcript set (aenmd.data.ensdb.v105 annotation package, see above) to annotate the gnomAD, Clinvar, and GWAS catalog databases of human genetic variation for PTC-generating variants predicted to escape NMD. Our results are summarized in Supplementary Tables S1 - S3. We observe that the fraction of NMD-escape PTC-generating variants varies between 36% (ClinVar), 50% (gnomAD), and 57% (GWAS catalog). The fraction of coding variants in each database that introduce PTCs also varies (10% for ClinVar, 4.1% for gnomAD, and 4.5% for the GWAS catalog). While the absolute number of PTC-generating variants is low for the GWAS catalog (most of its variants are non-coding), we learn from gnomAD that half of the 300k PTC-generating variants recovered from 125k exome sequences are predicted to escape NMD. Analyzing the ClinVar database (**Figure** 5.1**B**, Supplementary Table C), we find that for the subset of variants that are considered pathogenic and generate PTCs, 34% (nearly 31k variants) are predicted to escape NMD. This suggests that escape from NMD may play a substantial role in the disease mechanisms underlying variants of clinical significance annotated in ClinVar.

5.3.5 Comparison with VEP NMD plugin

We note that Ensemble's Variant Effect Predictor (VEP) [11] provides an NMD annotation plugin that annotates escape from NMD for ”stop_gained” variants. This set of variants does not include frameshift variants with a downstream PTC, so the set of variants considered by aenmd and the VEP plugin are inherently different. For example, aenmd annotates 200k variant-transcript pairs for ClinVar, while VEP considers 77k due to its restrictions on variant type (Supplementary Table C).

Nevertheless, we systematically compared VEP and aenmd NMD escape predictions for the ClinVar database for variants that overlap in transcript set and variant type between the two methods. Overall, we find high consistency of NMD escape predictions (97.5% identical predictions), with 773 (out of 75,840) variant-transcript pairs annotated as NMD escaping by aenmd but not VEP, and with 1,096 pairs annotated as NMD escaping by VEP but not aenmd. We manually examined a limited set of 20 randomly selected variants with different predictions, and differences are often due to understandable technical differences in the implementation of NMD escape rules.

5.4 Discussion

Here we present aenmd, a self-contained, accessible, and scalable computational tool for annotating (predicted) escape from nonsense-mediated decay (NMD) for variants that generate premature termination codons (PTCs) in a transcript. While there exist other tools that annotate escape from NMD for PTC variants, aenmd is unique in its specific features. For instance, VEP annotates NMD escape but is lim-ited to “stop_gained” variants. Additionally, a user is unable to readily interpret which NMD escape rules underlie a certain prediction. This interpretability shortcoming is shared with NMDetective [12], a tool that annotates annotate escape from NMD and provides a NMD efficacy prediction, a feature that aenmd lacks; however, NMDetective's approach does not consider inserted sequence for indels that cause PTCs. The latter is also the case for the tool NMDescPredictor [66], which is also different from aenmd in that it does not provide batch annotation functionality, and it implements a smaller set of NMD escape rules. The tool ALoFT [67] more generally predicts pathogenicity of loss of function variants, but it has the capability to annotate NMD escape. However, its output is less fine-grained than aenmd's as to which specific rules drive NMD escape annotations. Similarly, the variant annotator SNPEff [52] provides NMD escape prediction, but it only considers two NMD escape rules (penultimate and last exon rules). In summary, aenmd stands out in terms of its functionality, flexibility, and interpretability of results.

We also performed a detailed comparison of aenmd with the VEP NMD plugin, which annotates fewer variant types, uses a smaller set of NMD escape rules, and does not report the outcome of individual rules. While aenmd annotates substantially more variants, we nevertheless found that overlapping predictions were highly consistent

We note that the rules aenmd (and other tools) utilize for predicting escape from NMD do not yield perfect annotations, and not all of the rules are believed to work equally well. For instance, Lindebloom et al. [12] in their NMDetective-B model observe the highest efficacy for the "last exon" rule, followed by the "CSS proximal" rule, followed by the "penultimate exon" rule, followed by the "407bp plus" rule when analyzing can-cer data. Further on, it is conceivable that the efficacy of different rules changes across different tissues/cell-types where affected transcripts are expressed. However, a recent study leveraging GTEx data [176] found that NMD effects were highly stable across tissues and individuals, and it concludes that NMD prediction tools' predictive power should be stable across tissues.

In addition, we note that NMD plays a role in designing CRISPR gene editing experiments [12], and therefore aenmd's functionality will potentially be useful in this context as well.

In summary, aenmd's comprehensive features, flexibility, and ease of use allow for improved annotation of PTC-generating variants at low computational cost.

5.5 Author contribution

Qianqian Liang contributed to the manuscript "aenmd: Annotating escape from nonsense-mediated decay for transcripts with protein-truncating variants" by:

- Conducted software testing and debugging to ensure the accuracy and reliability of the analysis.
- Performed the analysis for the GWAS catalog analysis, which involved data preprocessing, statistical analysis, and interpretation of results.
- Analyzed the ClinVar dataset and contributed to the generation of the ClinVar analysis plot (**Figure** 5.1**B**)
- Participated in setting up the GitHub repository for the aenmd_manuscript, which facilitated collaboration and version control among the co-authors.
- Debugged the code and contributed to the development of supplemental tables that provided additional information and context for the main results.

The next section (**Section** 5.6) is an extension of the manuscript by applying the aenmd tool to analyze variants in the GWAS dataset. This section is conducted by Qianqian Liang.

5.6 Additional methods and results

5.6.1 Additional introduction

Premature termination codon (PTC) genetic variants are known to contribute to human diseases, and their impact can be various based on whether it escape or not from nonsense-mediated mRNA decay (NMD) [12, 8]. For example, PTC variants that escape NMD can exacerbate beta-thalassemia by introducing toxic truncated proteins, while those that escape NMD may result in a milder form of Duchenne muscular dystrophy by producing truncated proteins with partial function [8, 9, 10]. Numerous studies have investigated how NMD escape can aggravate or alleviate disease phenotypes; however, most of these studies have focused on individual diseases, making it necessary to explore escaping from NMD in a broader context [144, 143, 142, 139].

In a recent study, Lindeboom et al. analyzed NMD escape and NMD triggering enrichment in PTC variants, albeit only in terms of the disease gene [12]. Therefore, this study aims to analyze the enrichment of NMD or NMD-escape disease-associated PTC variants for specific phenotypes by comprehensively analyzing all possible trait terms using the GWAS Catalog. In this section, we will utilize the PTC-introducing coding variants available in the GWAS Catalog database and leverage the aenmd tool developed in the previous section of this chapter to annotate these variants. Subsequently, we will analyze the enrichment of NMD escape versus NMD triggering across all potential terms documented in the GWAS Catalog. Given the limited number of coding variants in the GWAS Catalog, we will utilize the ontology to gather genetic variants annotated to the term itself and its child terms to increase the number of variants for the terms we analyzed.

By investigating the enrichment patterns of NMD escape and NMD triggering in disease-associated PTC variants across a wide range of phenotype terms, this study offers a comprehensive understanding of the relationship between NMD escape, disease phenotypes, and potential therapeutic implications.

5.6.2 Additional methods

5.6.2.1 Process and annotate the GWAS Catalog dataset

Genetic variants, including single nucleotide variants and insertions and deletions, were retrieved from the GWAS Catalog version 20220730 [169]. We conducted the following filtering/preprocessing steps to get the potential PTC causing variants: 1. get variants that are located in coding regions; 2. get variants that are potential PTC causing variants (annotated as frameshift/stop gain in GWAS catalog CONTEXT column); 3. normalize variants so that variants can have standard format to work with aenmd package (see **Section** 5.2), especially for insertions and deletions; 4. get reference and alternative allele using ncbi_snp_query function from rsnps package [177], as the GWAS catalog did not have alternative allele information for some variants; 5. annotate variants so that each variant is mapped on a single

EFO phenotype. If one variant is mapped to more than one EFO phenotype, we will generate multiple entries for that variant.

Using the preprocessed potential PTC variants, we use the aenmd package to filter for PTC variants and annotate NMD escape. If one variant is annotated as NMD escape and NMD triggering in different transcripts or alternative alleles, we predicted this variant as NMD escape.

5.6.2.2 Term frequency generation

In the previous section, we created a table that includes information on the variants, their associated phenotypes, and their NMD escape annotations. Using this table, we computed the term frequency for all possible EFO terms, where the EFO frequency represents the percentage of phenotypes in the table that are descendants of that particular EFO term. For instance, if a particular EFO term, such as "hematological measurement," has a term frequency of 21%, it means that 21% of the phenotypes listed in the table are subtypes of hematological measurement.

We sort all EFO terms based on the term frequency and use the top 50 terms for further analysis. We removed terms that are too broad, such as 'experimental factor' or 'information entity' from further analysis, as they provide no biological meaning. Terms that contain less than ten unique variants were also removed.

5.6.2.3 Statistical tests

For each EFO term we analyzed, we got all GWAS PTC variants that were associated with that term or to the descendants of that term. Variants were annotated as NMD triggering or NMD escape using aenmd package (**Section** 5.6.2.1). To investigate the association between NMD escape/triggering and the presence/absence of a specific term, we performed Fisher's exact test on a contingency table. The contingency table was constructed with the rows representing the presence or absence of a specific term (e.g., "term X" or "not term X") and the columns representing the NMD status (escape or triggering).

Variants may annotate to more than one trait terms. For each term X, we retrieve all the variants that are associated with either term X or descendants of term X. Rest of the variants are annotated as not term X. Each variant is only counted once in the contingency table. It is recognized that some variants assigned to term X may also be associated with a trait that is not in term X. However, due to the limited number of variants, we do not exclude these variants from our analysis.

Pleiotropy effect analysis is also performed using Fisher's exact test. The contingency table considered the number of traits associated with a variant (one vs. more than three) and its NMD status (escape or triggering).

5.6.3 Additional results

5.6.3.1 NMD escape annotation for the GWAS Catalog

We first use the aenmd to annotate all possible PTC variants in the GWAS Catalog. In total, the GWAS Catalog (as of version 20220730) contained 197,442 unique variants, of which 5,703 were coding variants when considering only those overlapping with the aenmd default transcript set. Among these coding variants, only 254 were found to cause PTC. This relatively low number can be attributed to the prevalence of non-coding variants in the GWAS Catalog (See **Table** 5.1).

Among the GWAS PTC variants, when considering both canonical and non-canonical rules, 57.1% were found to escape NMD. This percentage is higher than the NMD escape rate observed for variants in the ClinVar Database (38.3%) (See **Appendix Table** C). When examining subtypes within ClinVar variants, the GWAS PTC variants exhibited a higher NMD escape rate compared to pathogenic variants (36.2%), but a lower rate compared to benign variants (64.8%) in ClinVar (See **Figure** 5.1**B** and **Appendix Table** C). Similar trends were observed when analyzing PTC variant-transcript pairs. Among the GWAS PTC variants-transcript pairs, 49% were found to escape NMD, whereas the NMD escape rate for ClinVar variants was 34.8%.

Overall number of variants:	197,442	
Variants overlapping AENMD default tx set:	5,703	
	number	percentage
Unique PTC variants	254	100
NMD triggering	109	42.9
NMD escaping canonical	96	37.8
NMD escaping non-canonical	42	16.5
Transcript dependent predictions	7	2.8
NMD escaping overall	145	57.1
PTC variant-transcript pairs	437	100
NMD triggering	223	51
NMD escaping canonical	150	34
NMD escaping non-canonical	64	15
NMD escaping overall	214	49

Table 5.1: *Distribution of NMD escape annotations for GWAS catalog based on unique variants or variant-transcript pairs.*

5.6.3.2 Term enrichment analysis revealed phenotypes that are enriched with NMD variants

Using aenmd tool, we annotated the NMD outcome for variant-phenotype pairs for all PTC variants in the GWAS catalog using reference and alternative alleles. We then generate a list of EFO terms using those phenotypes and calculate the term frequency for all possible EFO terms and take the top 50 terms for further analysis (See **Section** 5.6.2.2). We then conduct Fisher's exact test to test all those EFO terms and calculate the enrichment of NMD escape or NMD triggering for all those terms.

Our analysis revealed significant enrichments of PTC variants that undergo or escape

NMD in various phenotypic terms. Specifically, we observed one notable enrichment of NMD triggering variants in glycoprotein measurement (EFO:0004555 odds ratio = 0.14, 95% confidence interval [0.01-0.67]). There are some other terms such as leukocyte count(EFO:0004308, odds ratio = 0.41 [0.09-1.68]), erythrocyte indices (EFO:0004306, odds ratio = 0.188 [0.14-1.68])and platelet count (EFO:0004309, odds ratio = 0.48, [0.14-1.57]) show some enrichment of NMD triggering, but not significant. On the contrary, there are some other terms such as triglyceride measurements(EFO:0004530, odds ratio = 3.18, [0.83-17.98]), anthropometric measurement(EFO:0004302, odds ratio = 1.9, [0.76-5.23]), and body weights and measures (EFO:0004324, odds ratio = 1.79, [0.7-4.95]) show some enrichment of NMD escape, but not significant. (See **Figure** 5.2). Some broader terms, such as disease (EFO:0000408, odds ratio = 0.88 [0.46-1.68]), do not show significant enrichment in NMD triggering or NMD escape.

We examine specific examples to elucidate these findings further. The term glycoprotein measurement (EFO:0004555) shows a significant enrichment of NMD triggering variants (odds ratio 0.14, p-value = 0.005). It exhibited a term frequency of 2.2% and encompassed 3 descendant traits (HbA1c measurement, EFO:0004541; sex hormone-binding globulin measurement, EFO:0004696; and erythropoetin measurement, EFO:0008391) within the GWAS Catalog, encompassing a total of 12 variant-phenotype pairs (with 12 unique variants). Notably, 17% of PTC variants associated with glycoprotein measurement evaded NMD, compared to 59% of PTC variants associated with other terms (See **Figure** 5.2 and **Table** 5.2).

Let us now consider another example. The term triglyceride measurement (EFO:0004530) shows some enrichment of NMD escaping variants but is not significant (p-value = 0.104, odds ratio 3.18). It exhibited a term frequency of 2.8% and encompassed one descendant trait (triacylglycerol 50:5 measurement, EFO:0010412). This term comprised a total of 15 variant-phenotype pairs (15 unique variants). Among the PTC variants associated with triglyceride measurements, 80% evaded NMD, compared to 56% of PTC variants associated with other terms (See **Figure** 5.2 and **Table** 5.3).

We also investigated the term ”disease” (EFO:0000408) as we are interested in whether NMD triggering or NMD escape are enriched in a disease vs a non-disease trait. The term

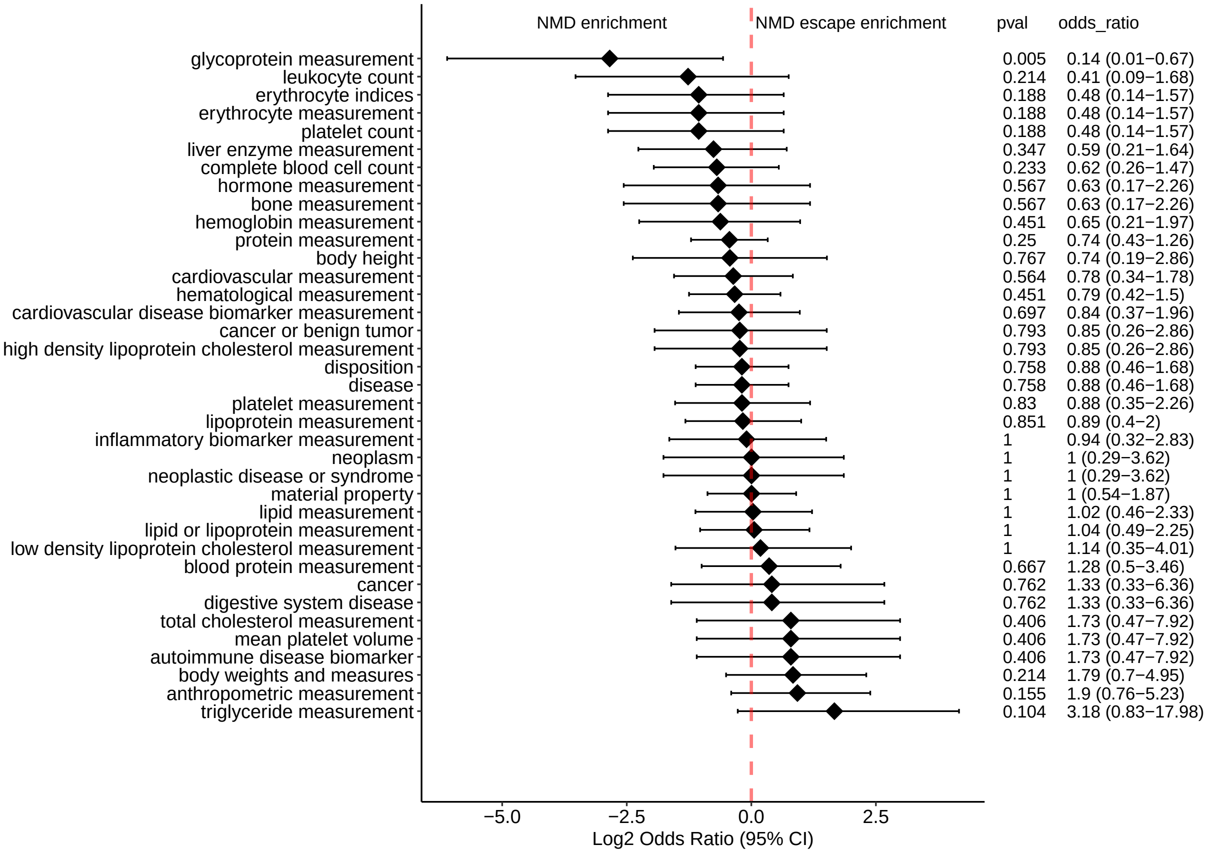

Figure 5.2: *NMD enrichment for EFO terms.* The X axis is the Log2 Odds ratio from the enrichment analysis. The y axis is the EFO terms sorted by Log2 odds ratio.

”disease” encompassed all traits categorized as diseases (e.g., Crohn’s disease, EFO:0000384; inflammatory bowel disease, EFO:0003767). This term included 60 distinct disease terms and 90 variant-disease pairs (55 unique variants). When comparing disease traits to non-disease traits, we do not observe any enrichment in NMD triggering or NMD escape (odds ratio 0.88, p-value = 0.758) (See **Figure** 5.2 and **Table** 5.4).

	NMD escape	NMD triggering	escape percentage
Glycoprotein measurement	2	10	17% escape
Not glycoprotein measurement	143	99	59% escape
Total	145	109	57% (254 total)

Table 5.2: *NMD enrichment analysis for glycoprotein measurement.*

	NMD escape	NMD triggering	escape percentage
Triglyceride measurement	12	3	80% escape
Not triglyceride measurement	133	106	56% escape
Total	145	109	57% (254 total)

Table 5.3: *NMD enrichment analysis for triglyceride measurement.*

	NMD escape	NMD triggering	escape percentage
Disease	30	25	55% escape
Not disease	115	84	58% escape
Total	145	109	57% (254 total)

Table 5.4: *NMD enrichment analysis for disease vs non-disease.*

5.6.3.3 Pleiotropy analysis revealed marginal significance of pleiotropy traits are enriched with NMD escaping variants

Lastly, we explored the effect of NMD on pleiotropy. Specifically, we tested whether NMD escape is enriched in SNVs associated with one trait or SNVs associated with multiple traits. Here, we found that we have a higher rate of NMD escape in PTC variants associated with three more traits (75% escape) than PTC variants associated with only one trait (56% escape) (See **Table** 5.5), but with only marginally significant (pval = 0.056, odds ratio = 0.43 [0.16-1.03]).

pleiotropy	NMD escape	NMD triggering	escape percentage
N trait >3	27	9	75% escape
N trait = 1	82	63	56% escape

Table 5.5: *NMD enrichment analysis for pleiotropy effect in GWAS Catalog.*

5.6.4 Additional discussion

In this section, we used our new tool aenmd to annotate PTC genetic variants in one of the genetic datasets: the GWAS Catalog. To investigate the outcome of NMD in particular trait terms, we analyze the enrichment of NMD escape and NMD triggering PTC variants for specific phenotype terms. We discovered one trait, glycoprotein measurement, that is significantly enriched with NMD escape variants compared with other traits. In addition, we found that traits such as triglyceride measurement show some enrichment with NMD escaping variants, however, it is not statistically significant. Moreover, we also revealed that NMD has some function in pleiotropy, while there is an enrichment of NMD escapes in the variants associated with three more traits, compared with variants associated with only one trait.

However, we have to admit that the number of the PTC variants we analyzed overall is relatively small (254 unique variants). This is due to that the GWAS Catalog mostly contains non-coding genetic variants (more than 95%), and the percentage of the PTC bearing variants among all coding is lower in the GWAS Catalog (4.4%) compared with Clinvar (10.3%) (See **Table** 5.1 and **Appendix Table** C). However, this study is meaningful, as most other people have studied Clinvar, which mainly contains variants in Mendelian diseases [12]; our study broadens our understanding by studying the impact of NMD in complex traits.

We acknowledge that there could be some ascertainment bias in the study as the PTC variants we studied are not randomly picked but the ones that have at least some function in a trait or disease. Further studies on a more general population of randomly picked PTC variants could be conducted to understand the function of NMD further.

In our analysis, we found that PTC variants that escape NMD are enriched for disease traits compared to non-disease traits. However, Lindeboom et al. observed that in more disease genes, pathogenic PTC variants triggering NMD are overrepresented [12]. These seemingly discrepant findings could be explained by differences in the types of diseases studied, while we studied complex disease traits Lindeboom et. al. are focusing on Mendelian diseases. This suggests that while NMD may have deleterious effects in the context of Mendelian diseases, for complex diseases, NMD escape may contribute more to disease associations compared to non-disease traits. This highlights the complex interplay between NMD and disease mechanisms, emphasizing the need for further investigation and a nuanced understanding of NMD's role in different disease contexts.

6.0 Conclusions

6.1 Summary

In this dissertation, we presented computational approaches that can characterize and prioritize protein-coding and non-protein coding genetic variants, and by analyzing or utilizing these methods, we revealed significant biological findings related to human disease. In Chapter 3, we presented a disease-specific non-coding variant prioritization method that showed improved performance than current disease-agnostic methods. Using a relatively simple logistic regression model in this method, we were able to highlight relevant tissues for specific diseases and find meaningful disease groups. The findings in this chapter highlight the value of disease-specific variant prioritization. In Chapter 4, we improved the performance of the disease-specific variant prioritization method by combining SNVs from related diseases. Using the information sharing model, we showed that by adding SNVs from related diseases, the performance of the disease-specific variant score can improve up to 40%. These findings suggest that data sharing is a promising avenue for improving the performance of disease-specific models. In Chapter 5, we developed a tool ”aenmd” which can annotate mRNA non-sense mediated decay (NMD) that addresses the limitations of current NMD annotation tools. Then we utilized this tool to annotate coding variants associated with traits from the GWAS Catalog and we found traits that are enriched with NMD triggering or escaping variants. This expands our knowledge of the function of NMD in the field of complex traits. Overall, we present better computational approaches to annotate coding and non-coding genetic variants, thereby enhancing our understanding of the genetic basis of complex human diseases and traits.

6.2 Significance

As most current variant prioritization methods are disease-agnostic, my work first expands our knowledge by demonstrating that disease-specific scores can improve our ability to prioritize disease-associated variants. In addition, by utilizing a relatively simple logistic regression approach, our method is able to highlight significant tissues relevant to diseases and find disease groups with biological meaning. Next, we improved the disease-specific scores by incorporating SNVs from related SNVs. This suggests that data sharing is a promising avenue for improving the performance of disease-specific models. Lastly, the NMD annotation tool *aenmd* contains many functions that are absent in current tools and provides a user-friendly way to annotate NMD. By applying the tool to the GWAS Catalog, our findings expand our knowledge of the role of NMD in complex diseases and traits.

There are a few things that we can consider as part of our future plan for the thesis. Firstly, functional datasets like ENCODE are expanding their datasets by using biosamples instead of the standard 127 roadmap tissues. We can include these datasets as features to train our model in the future, which has the potential to increase the model's performance and help us find more disease-relevant tissues. Secondly, we can conduct further research to better understand the impact of population stratification on our method. Lastly, in Chapter 5, we identified some traits that are enriched with NMD escape or NMD trigger variants. To better understand the role of NMD, we can conduct further investigations into the biological relevance of these traits.

Appendix A Supplemental materials for aim1

A.1 Supplemental material

A.1.1 Four different matching strategies for control SNVs

For each disease-associated SNV, we have matched control SNVs from four different matching strategies: 1) random; 2) SNPsnap_TSS; 3) SNPsnap; and 4) TSS (see **Method**). We measured the performance of five organism level scores on four different control sets in 111 diseases. Among them, random matching is considered as the least stringent way as we don't have any constraint on it. Therefore, we choose the random matching as the baseline, and we normalize the performance of the other three control sets on random matching for each disease. We plot the normalized performance in Appendix Figure A.1 and A.2, using each disease as a panel.

From here, we observe that three normalized performances in CADD are all distributed around 1. This indicates that the CADD is robust in different matching strategies. The normalized performances of Eigen, GenoCanyon, GWAVA and LINSIGHT are all less than 1. This indicates that those three matched control sets are more stringent than randomly selecting control variants. Among those three control sets, TSS is the most stringent, followed by SNPsnap and SNPsnap_TSS.

It is important to note that TSS and SNPsnap TSS are both matched using the distance to the nearest TSS; however, TSS uses the distance to the nearest protein-coding gene while SNPsnap_TSS uses the distance to the nearest gene. TSS matched SNVs have similar distribution to disease SNVs in both all genes and protein-coding genes; in contrast, SNPsnap_TSS SNVs have similar distribution to disease SNVs in all genes but not in protein-coding genes (Appendix Figure A.3 and A.4). Therefore, TSS is more stringent than SNPsnap_TSS and also more stringent than SNPsnap even though SNPsnap has matched with additional three criteria.

Here, SNPsnap matching strategy is neither too stringent nor too loose and it matches with four criteria (see **Method**). Thus, we choose the control set matched by SNPsnap in our study.

A.1.2 DHS-weighted performance using two additional strategies to prevent overfitting

To prevent overfitting, we also deployed two additional strategies to test the performance of tissue-weighted DHS. In the first one, we used the 'representative SNVs' so that any two disease-associated SNVs are not in the same LD block. In the second one, we deployed a chromosome held-out strategy so that the SNVs in the test and train set are on different chromosomes (see **Method**). These two strategies ensure that the SNVs in the test and train set are separated or in different chromosome to reduce overfitting. We observe that in any of these two settings, we can still observe a significant increase with the tissue-weighted model, which is consistent with our previous finding, even though the amount of the improvement is in a lesser degree in some diseases (See See Appendix Figure A.18- A.21).

A.2 Supplemental tables

Score	Wins	Losses	Ties	Wins (agg)	Losses (agg)	Ties (agg)
GenoCanyon	335	87	22	4	0	0
LINSIGHT	303	124	17	3	1	0
eigen	247	177	20	2	2	0
GWAVA	168	256	20	1	3	0
CADD	15	424	5	0	4	0

Table A.1: *Relative performance of organism-level variant scores, measured by AUROC.* Wins, Losses, Ties refers to significantly better (or worse, or tied) performance across all possible pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term, while the last three columns represent results of aggregate comparisons across terms. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were ties).

Score	Wins	Losses	Ties	Wins (agg)	Losses (agg)	Ties (agg)
DHS	138	54	30	2	10	0
Fitcons2	80	111	31	0	1	1
Genoskyline	69	122	31	0	1	1

Table A.2: *Relative performance of disease-specific variant scores, measured by AUROC.* Wins, Losses, Ties refers to significantly better (or worse, or tied) performance across all possible pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term, while the last three columns represent results of aggregate comparisons across terms. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were ties).

Score/Method	By disease term			Aggregated		
	Wins	Losses	Ties	Wins	Losses	Ties
DHS	375	127	53	4	0	1
GenoCanyon	375	144	36	4	0	1
LINSIGHT	342	184	29	3	2	0
eigen	273	250	32	2	3	0
GWAVA	186	338	31	1	4	0
CADD	19	527	9	0	5	0

Table A.3: *DHS outperforms organism-level variant scores, measured by AUROC.* Wins, Losses, Ties refer to significantly better (or worse, or tied) performance across all possible score pairings (see **Methods**). The first three columns summarize separate comparisons for each disease term (for each row there are two other methods and 111 terms, i.e., 555 comparisons), while the last three columns represent results of comparisons aggregated over terms. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were reported as ties).

Score/Method	Performance vs. DHS			
	Wins	Losses	Ties	Winning percent
LINSIGHT	17	84	10	20
GenoCanyon	7	92	12	12
GWAVA	12	93	6	14
eigen	4	98	8	8
CADD	4	107	2	4

Table A.4: *Disease-specific variant prioritization outperforms organism-level approaches, measured by AUPR.* Wins losses and ties of organism-level scores against tissue-weighted DHS scores (performance measured by average precision, Wilcoxon signed-ranks test for determining significance). Winning percent was calculated as number of wins plus half the number of ties, divided by the number of comparisons, and rounded to the nearest integer. Rows have been ordered by winning percent.

Score/Method	Performance vs. DHS			
	Wins	Losses	Ties	Winning percent
GenoCanyon	40	57	14	36
LINSIGHT	39	60	12	35
eigen	26	73	12	23
GWAVA	18	82	11	16
CADD	4	103	4	4

Table A.5: *Disease-specific variant prioritization outperforms organism-level approaches, measured by AUROC.* Wins losses and ties of organism-level scores against tissue-weighted DHS scores (performance measured by average precision, Wilcoxon signed-ranks test for determining significance). Winning percent was calculated as number of wins plus half the number of ties, divided by the number of comparisons, and rounded to the nearest integer. Rows have been ordered by winning percent.

Score	Wins	Losses	Ties	Winning percent
GenoCanyon	31	25	2	55
DHS	27	26	5	51
DIVAN	24	31	3	44

Table A.6: *DHS tissue-weighted disease-specific scoring outperforms DIVAN.* Across 30 disease terms, this table summarizes all pairwise comparison for DHS tissue-weighted, GenoCanyon and DIVAN using a specifically created test dataset. Wins, losses, ties refer to significantly better (or worse, or tied) performance. Average precision was used as the performance metric, and the Wilcoxon singed-ranks test to determine wins and losses (p-values less than 0.05 were ties). Winning percent = #Wins/(#Wins+#Losses).

Rank	ID	Tissue name	Group	epimap FDR
Systemic scleroderma				
1	E116	GM12878 Lymphoblastoid Cells	blood	0.04
2	E032	Primary B cells from peripheral blood	blood	0.16
3	E041	Primary T helper cells PMA-I stimulated	blood	0.08
4	E123	K562 Leukemia Cells	blood	1.00
5	E030	Primary neutrophils from peripheral blood	blood	0.86
Sclerosing cholangitis				
1	E116	GM12878 Lymphoblastoid Cells	blood	<0.001
2	E061	Foreskin Melanocyte Primary Cells skin03	skin	0.18
3	E102	Rectal Mucosa Donor 31	gi_rectum	<0.001
4	E041	Primary T helper cells PMA-I stimulated	blood	<0.001
5	E029	Primary monocytes from peripheral blood	blood	<0.001
Colorectal adenoma				
1	E102	Rectal Mucosa Donor 31	gi_rectum	0.05
2	E110	Stomach Mucosa	gi_stomach	0.008
3	E057	Foreskin Keratinocyte Primary Cells skin02	skin	0.12
4	E101	Rectal Mucosa Donor 29	gi_rectum	0.004
5	E028	Breast variant Human Mammary Epithelial Cells (vHMEC)	breast	0.20
Atrial fibrillation				
1	E083	Fetal Heart	heart	<0.001
2	E108	Skeletal Muscle Female	muscle	0.01
3	E107	Skeletal Muscle Male	muscle	0.009
4	E088	Fetal Lung	lung	0.002
5	E120	HSMM Skeletal Muscle Myoblasts Cells	muscle	0.18
Cutaneous melanoma				
1	E061	Foreskin Melanocyte Primary Cells skin03	skin	0.08
2	E059	Foreskin Melanocyte Primary Cells skin01	skin	0.08
3	E117	HeLa-S3 Cervical Carcinoma Cell Line	cervix	0.40
4	E041	Primary T helper cells PMA-I stimulated	blood	0.50
5	E122	HUVEC Umbilical Vein Endothelial Primary Cells	vascular	0.87

Table A.7: *Top-ranked tissues for five diseases.* For five diseases when show the top-five tissues with the largest tissue weights in the corresponding model we derive. The first column is the tissue rank, the second the tissue's roadmap ID, the third the tissue name, the fourth the tissue group, and the fifth listst the adjusted p-value in an enrichment analysis performed by epimap [84].

A.3 Supplemental data legends

The supplemental data can be found at: `http://d-scholarship.pitt.edu/46026/2/sup_data_qianqian_dissertation.zip`

A.3.1 Phenotypes used in this study

`Filename: sup_data_disease-terms.csv.gz`

The first column denotes the EFO name of disease phenotypes used. Column #2 is the EFO ID. Column #3 shows the number of SNVs associated with the term (coding and non-coding). Columns #4 shows the number of non-coding SNVS used in the study before aggregation and #5 shows the number of non-coding SNVs used after aggregation. Non-EUR 1KG SNVs and SNVs in the HLA region have been removed in column #4 and #5.

A.3.2 Disease-associated SNVs used in this study

`Filename: sup_data_disease-snvs.csv.gz`

The first column denotes the SNV ID. Column #2 is the rsID. Column #3 is the phenotype. Columns #4 and 5 are the chromosome and the specific location (hg19 coordinates). Column #6 is the LD block cluster id where this SNV resides(SNVs in the same LD block will have the same cluster id), and column #7 indicates whether this SNV is selected as the representative SNV for the block (1 as selected, 0 as not selected). SNVs associated with multiple diseases appear in more than one row.

A.3.3 Control SNVs used in this study

`Filename: sup_data_control-snvs.csv.gz`

For each disease-associated SNV, this table lists ∼10 randomly-selected control SNVs by four different methods (see **Methods**). The first column denotes the SNV ID. Column #2 is the rs ID. Column #3 is the phenotype. Column #4 and 5 are the chromosome and the specific location (hg19) of that SNV. Column #6 is the matching strategy (i.e. snpsnap, snpsnap_tss, tss, random) and column #7 is the SNV ID of the corresponding disease-associated SNV.

A.3.4 Pairwise comparisons of organism-level scores for each disease term

`Filename: sup_data_pairwise-org-individual.csv.gz`

For each combination of organism-level scores we report p-values for a Wilcoxon signed-ranks test for each individual disease (see **Methods**). Column #1 is the score name. Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the disease term for which the comparison was performed. Column #6 is the curve type we used for the area under the curve performance metric (ROC or PR). Column #7 is the p-value of the test. Column #8 is the score with the higher median.

A.3.5 Pairwise comparisons of organism-level scores, aggregated across diseases

`Filename: sup_data_pairwise-org-aggregated.csv.gz`

For each combination of organism-level scores we report p-values for a Wilcoxon signed-ranks test, aggregated across 111 diseases (see **Methods**). Column #1 is the score name. Column #2 is the median performance across 111 diseases. Column #3 is the second score name. Column #4 is the median performance for the second score across 111 diseases. Column #5 is the curve type for the area under the curve performance metric (ROC or PR). Column #6 is the p-value of the test. Column #7 is the score with the higher median.

A.3.6 Pairwise comparison of tissue-weighted scores vs. tissue-mean scores for each disease term

`Filename: sup_data_pairwise-tis-individual.csv.gz`

For each score we report p-values for a Wilcoxon signed-ranks test between the tissue-mean and tissue-weighted version for each individual disease. Column #1 is the score name.

Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the disease term for which the comparison was performed. Column #6 is the curve type we used for the area under the curve performance metric (ROC or average precision). Column #7 is the p-value of the test. Column #8 is the score with the higher median.

A.3.7 Pairwise comparison of tissue-weighted scores vs. tissue-mean scores, aggregated across diseases

`Filename: sup_data_pairwise-tis-aggregated.csv.gz`

For each score we report p-values for a Wilcoxon signed-ranks test between the tissue-mean and tissue-weighted version, aggregated across all diseases. Column #1 is the score name. Column #2 is the median performance across all diseases for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the curve type for the area under the curve performance metric (ROC or PR). Column #6 the p-value of the test. Column #7 is the score with the higher median.

A.3.8 Pairwise comparison of three tissue-weighted scores for each disease term

`Filename: sup_data_pairwise-tis-weighted-individual.csv.gz`

For each combination of Tissue-weighted scores we report p-values for a Wilcoxon signed-ranks test for each individual disease (see **Methods**). Column #1 is the score name. Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the disease term for which the comparison was performed. Column #6 is the curve type we used for the area under the curve performance metric (ROC or PR). Column #7 is the p-value of the test. Column #8 is the score with the higher median.

A.3.9 Pairwise comparison of three tissue-weighted scores, aggregated across diseases

`Filename: sup_data_pairwise-tis-weighted-aggregated.csv.gz`

For each combination of Tissue-weighted scores we report p-values for a Wilcoxon signed-ranks test, aggregated across all diseases (see **Methods**). Column #1 is the score name. Column #2 is the median performance across all diseases for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the curve type for the area under the curve performance metric (ROC or PR). Column #6 is the p-value of the test. Column #7 is the score with the higher median.

A.3.10 Pairwise comparison of tissue-weighted-DHS vs five organism-level scores for each disease term

`Filename: sup_data_pairwise-tis-vs-org-individual.csv.gz`

We report p-values for a Wilcoxon signed-ranks test between the Tissue-weighted-DHS and five organism-level scores for each individual disease. Column #1 is the score name. Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the disease term for which the comparison was performed. Column #6 is the curve type we used for the area under the curve performance metric (ROC or PR). Column #7 is the p-value of the test. Column #8 is the score with the higher median.

A.3.11 Pairwise comparison of tissue-weighted-DHS and five organism-level scores aggregated

`Filename: sup_data_pairwise-tis-org-aggregated.csv.gz`

We report p-values for a Wilcoxon signed-ranks test between the Tissue-weighted-DHS and five organism-level scores, aggregated across all diseases. Column #1 is the score name. Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column

#5 is the curve type we used for the area under the curve performance metric (ROC or PR). Column #6 is the p-value of the test. Column #7 is the score with the higher median.

A.3.12 Mapping of mesh terms to EFO terms

`Filename: sup_data_mapping-efo-mesh.csv.gz`

The first and second columns are the mesh term id and mesh term label used by DIVAN. The third and fourth columns are the EFO ID and EFO label that is mapped to the mesh terms. (Note: there are two MeSH terms that are matched to more than 1 EFO term.)

A.3.13 Training and test SNVs used to compare Tissue-weighted with DIVAN (including disease and matched control SNVs)

`Filename: sup_data_divan-snvs.csv.gz`

Column #1 denotes the SNV ID. Column #2 is the rs ID of the SNV. Column #3 is the phenotype. Column #4 and #5 are the chromosome and location of the SNV. Column #6 indicates whether the variant is a disease-associated or a control variant. Column #7 is the SNV ID of the corresponding disease-associated SNV. Column #8 indicates whether the variant is in training or test set.

A.3.14 Pairwise comparison of DIVAN vs. GenoCanyon vs Tissue-weighted-DHS for each disease term

`Filename: sup_data_pairwise-divan-individual.csv.gz`

For each combination of DIVAN vs. GenoCanyon vs. Tissue-weighted-DHS we report p-values for a Wilcoxon signed-ranks test for each individual disease (see **Methods**). Column #1 is the score name. Column #2 is the median performance across bootstrap runs for that score. Column #3 is the second score name. Column #4 is the median performance for the second score. Column #5 is the disease term for which the comparison was performed. Column #6 is the curve type we used for the area under the curve performance metric

(ROC or PR). Column #7 is the p-value of the test. Column #8 is the score with the higher median.

A.3.15 GenoCanyon vs DIVAN in our study and in Chen study (the DIVAN study)

`Filename: sup_data_perf-divan-our-vs-chen.csv.gz`

Column #1 is the disease names of 27 overlapping diseases. Column #2 indicates whether GenoCanyon is better than DIVAN in our study. Column #3 indicates whether GenoCanyon is better than DIVAN as published by DIVAN.

A.3.16 Tissue-weighted prediction scores for SNVs across 111 diseases

`Filename: sup_data_prediction-scores-dhs-weighted.csv.gz`

Column #1 is the the SNV_ID (chr:position). (If a SNV is annotated to multiple phenotypes, there will be multiple entries.) Column #2 is the phenotype that is annotated to the SNVs. Column #3 indicates whether this SNV is a disease-associated variant or a control variant. Column #4-6 are Tissue-weighted prediction scores in Genoskyline, DHS and Fitcons2

A.3.17 Beta coefficients of the logistic regression models in 111 diseases (using DHS score)

`Filename: sup_data_beta-coefficients-mean-dhs.csv.gz`

Column #1 is the phenotypes. Column #2-128 are the mean of the coefficients of 127 tissues.

A.3.18 Standard deviation of the beta coefficients in A.3.1

`Filename: sup_data_beta-coefficients-sd-dhs.csv.gz`

Column #1 is the phenotypes. Column #2-128 are the standard deviation of the coefficients in 127 tissues.

A.3.19 Disease-disease similarities derived from the logistic regression model (DHS)

Filename: `sup_data_beta-model-similarity-dhs.csv.gz`

column #1 and column #2 are the names of the disease pairs. Column #3 is the weighted disease-disease similarity derived from the model.

A.3.20 Clusters assigned to 111 diseases

Filename: `sup_data_cluster-id-name.csv.gz`

Column #1 is the disease name. Column #2 is the cluster id. Column #3 is the cluster name.

A.3.21 Term frequency in 7 disease clusters

Filename: `sup_data_cluster_term_frequency.csv.gz`

Column #1 is the term name. Column #2 is the term id. Column #3 is the term frequency of a term in the cluster. Column #4 is cluster id. Term frequency means the fraction of diseases in this cluster that is a descendant of this term. For example, immune system disease with a term frequency 0.588 in cluster immune-1 means that 58.8% of diseases in immune-1 cluster is a immune system disease.

A.3.22 Top five tissues in 7 disease clusters

Filename: `sup_data_top-five-tissues.csv.gz`

Column #1 and #2 are the cluster id and name. Column #3-5 are the tissue id, tissue name and tissue anatomy.

A.3.23 ID, name and group of standard epigenomes

Filename: `sup_data_standard-epigenomes`

Column #1 is the ID of the standard epigenomes (e.g. E043). Column #2 is the group name (e.g. Blood & T-cell). Column #3 is the standardized epigenome name. Column #4 is the anatomy(e.g. blood). Column #5 is the type (e.g. PrimaryCell).

A.3.24 Genetic correlation of the disease pairs

Filename: `sup_data_genetic-correlation.csv.gz`

column #1 and column #2 are the name of the disease pairs. Column #3 is the genetic correlation derived from the GWAS ATLAS

A.3.25 Performance of tissue-weighted (DHS) in different held-out strategies

Filename: `sup_data_perf-chrom-heldout.csv.gz`

Column #1 is the disease name. Column #2-#10 are the performance of Tissue-weighted (DHS) and Tissue-mean (DHS) measured in different held-out strategies. CV-B: cross-validation, baseline; CV-LR: cross-validation logistic regression; CV-LR (SD): standard deviation of CV-LR; random-B: randomly sampled test set, baseline; random-B (SD): random-B standard deviation; random-LR: randomly sample test set, logistic regression; random-LR (SD): random-LR standard deviation; chr-B: test set held out by chromosome, baseline; chr-B (SD): chr-B standard deviation.

A.4 Supplemental figures

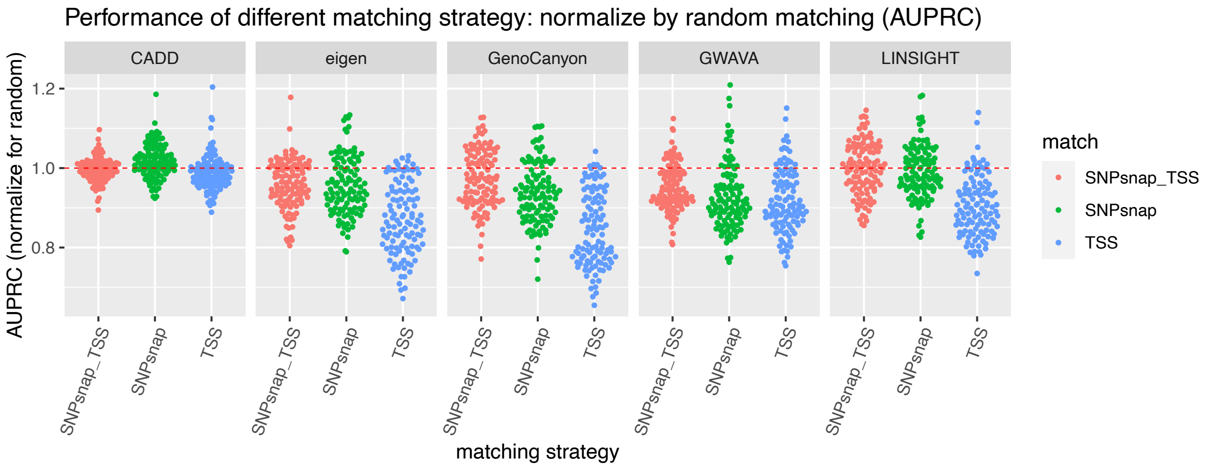

Figure A.1: *Performance of different matching strategies, measured by area under the PR curve.* X-axis delineates three different matching strategies (i.e. snpsnap-tss, snpsnap, tss). Y axis shows the performance in terms of area under precision recall curve, normalized by random matching. Each point represents a specific disease term. Horizontal lines spanning the dataset denotes the scenario that the normalized performance equals to 1.

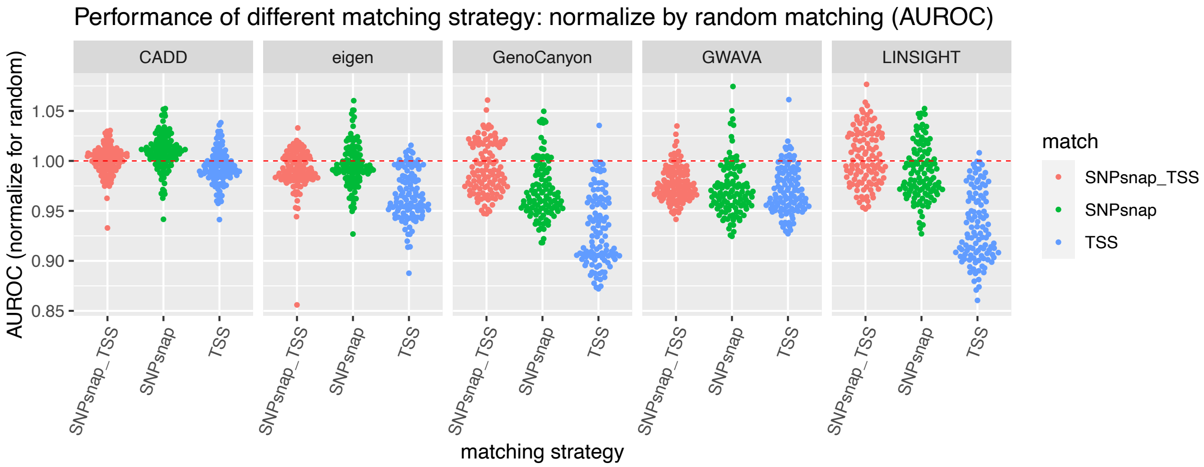

Figure A.2: *Performance of different matching strategies, measured by area under the ROC curve.* X-axis delineates three different matching strategies (i.e. snpsnap-tss, snpsnap, tss). Y axis shows the performance in terms of area under receiver operating characteristic curve, normalized by random matching. Each point represents a specific disease term. Horizontal lines spanning the dataset denotes the scenario that the normalized performance equals to 1.

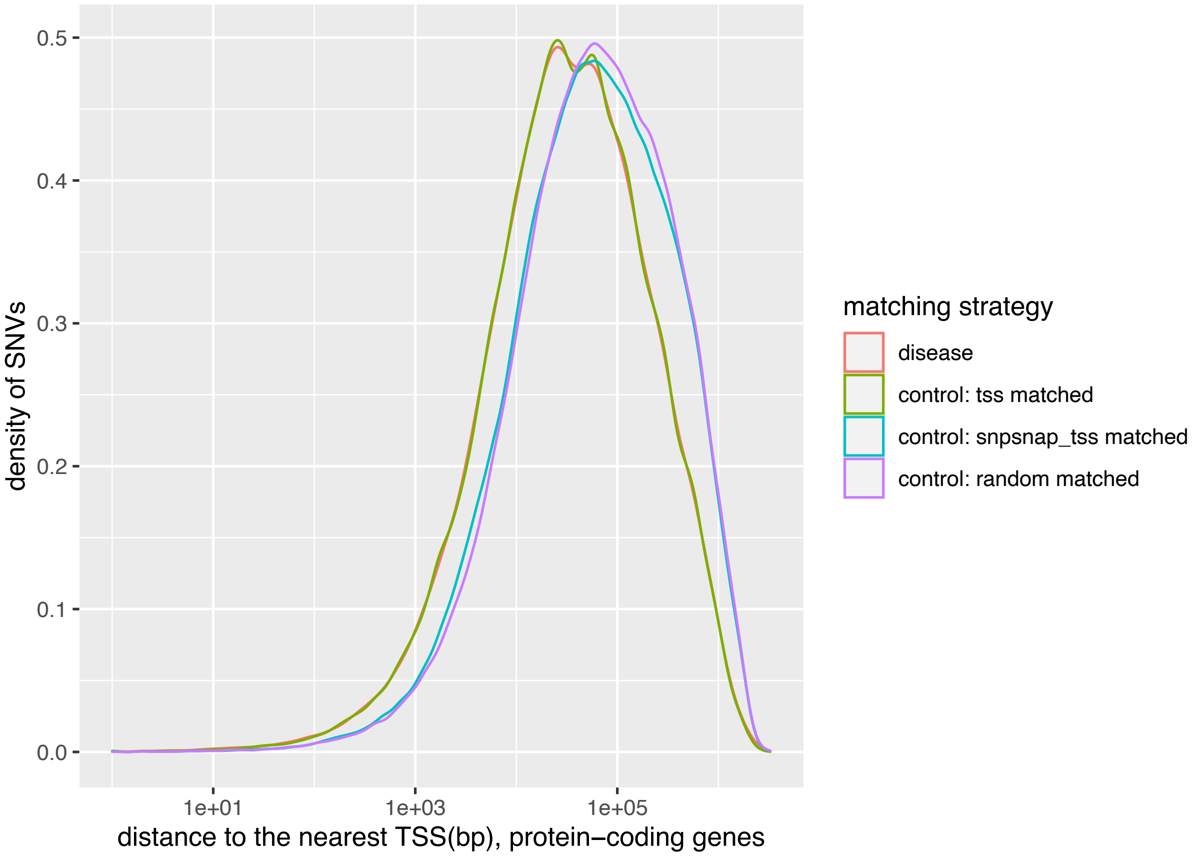

Figure A.3: *A density plot showing the distribution of distance to nearest TSS (protein coding genes) in disease SNVs and three different control SNVs.* X-axis shows the distance to the nearest TSS of the protein-coding genes and is log 10 scaled. Y axis shows the density of SNVs.

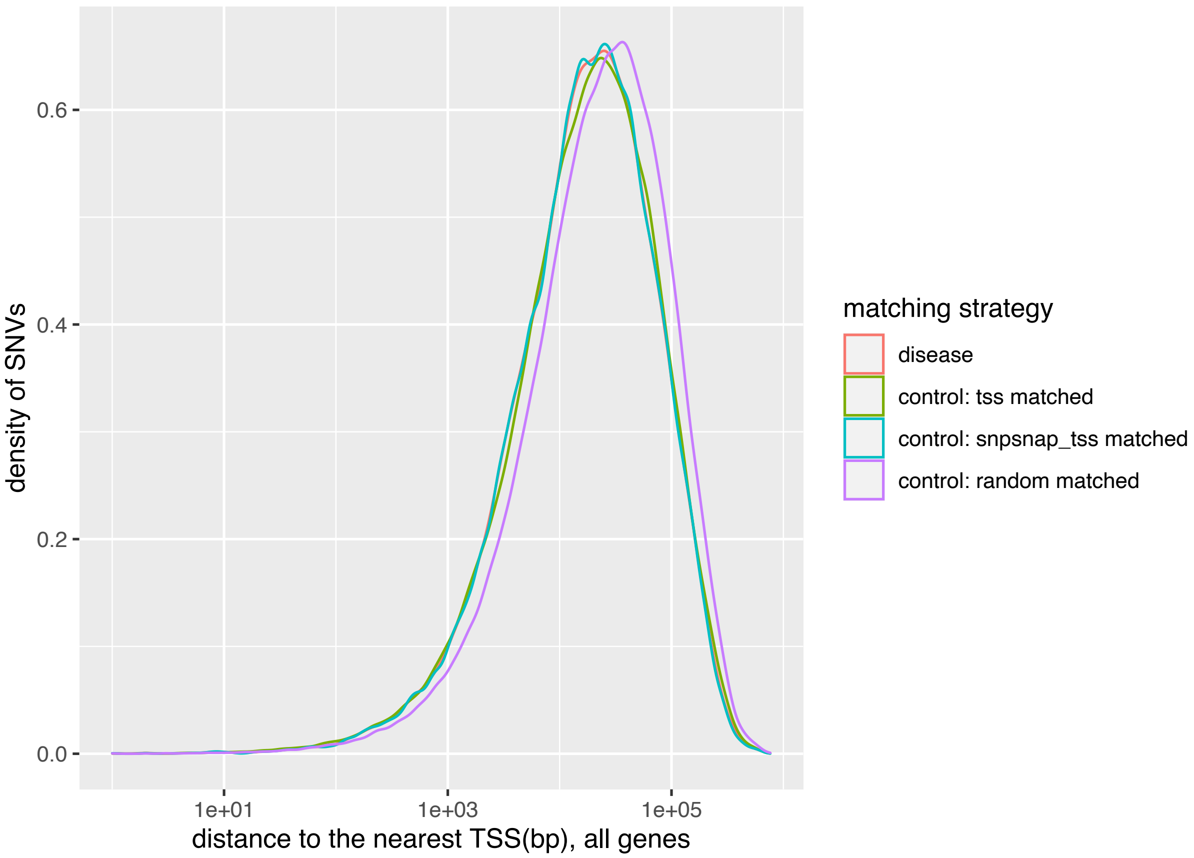

Figure A.4: *A density plot showing the distribution of distance to nearest TSS (all genes) in disease SNVs and three different control SNVs.* X-axis shows the distance to the nearest TSS of the protein-coding genes and is log 10 scaled. Y axis shows the density of SNVs.

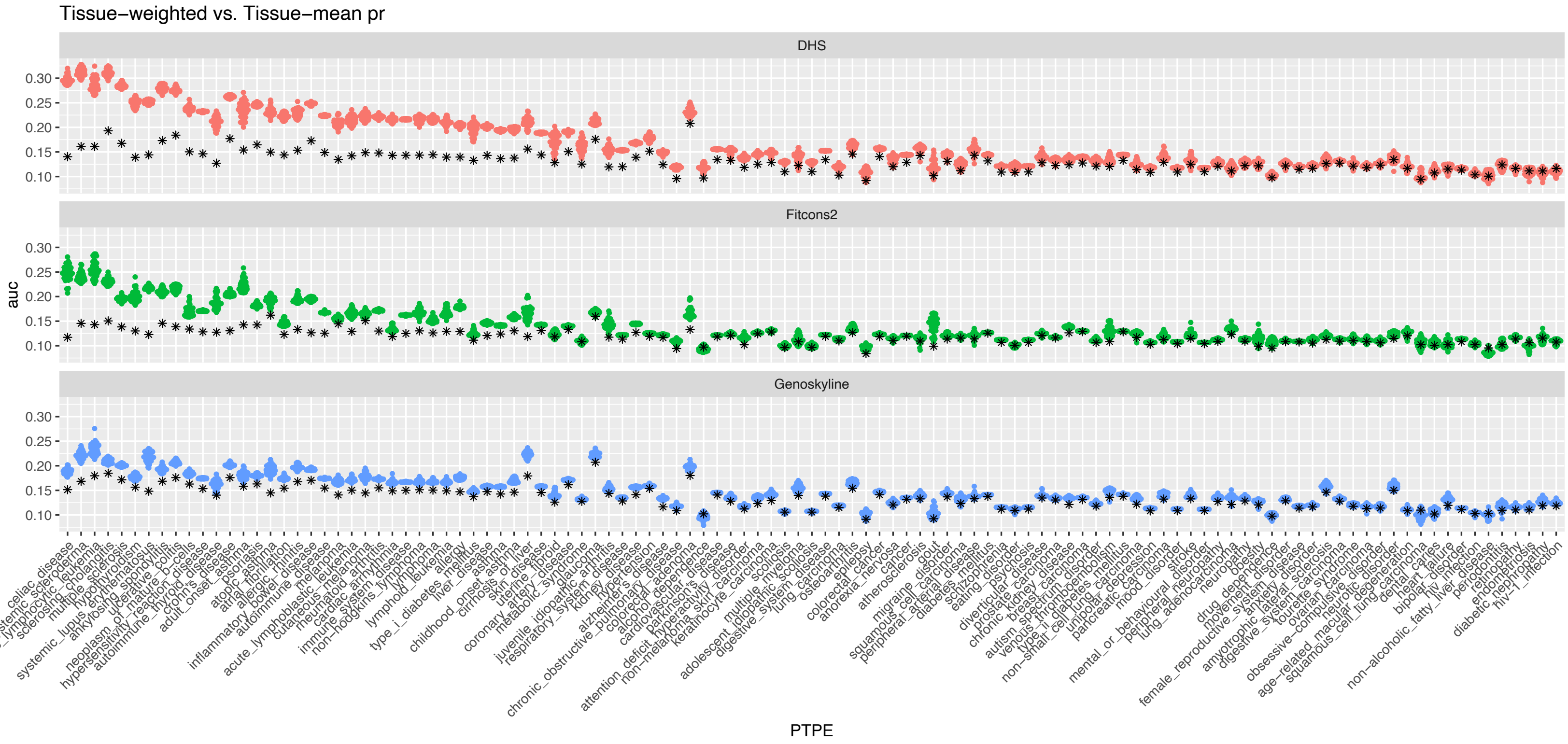

Figure A.5: *Tissue-weighted performance compared with Tissue-Mean in 111 diseases.* X-axis delineates different diseases and y-axis is the performance in terms of area under the precision recall curve. The star represents the Tissue-Mean and the colored dots are Tissue-Weighted with 30 replicates.

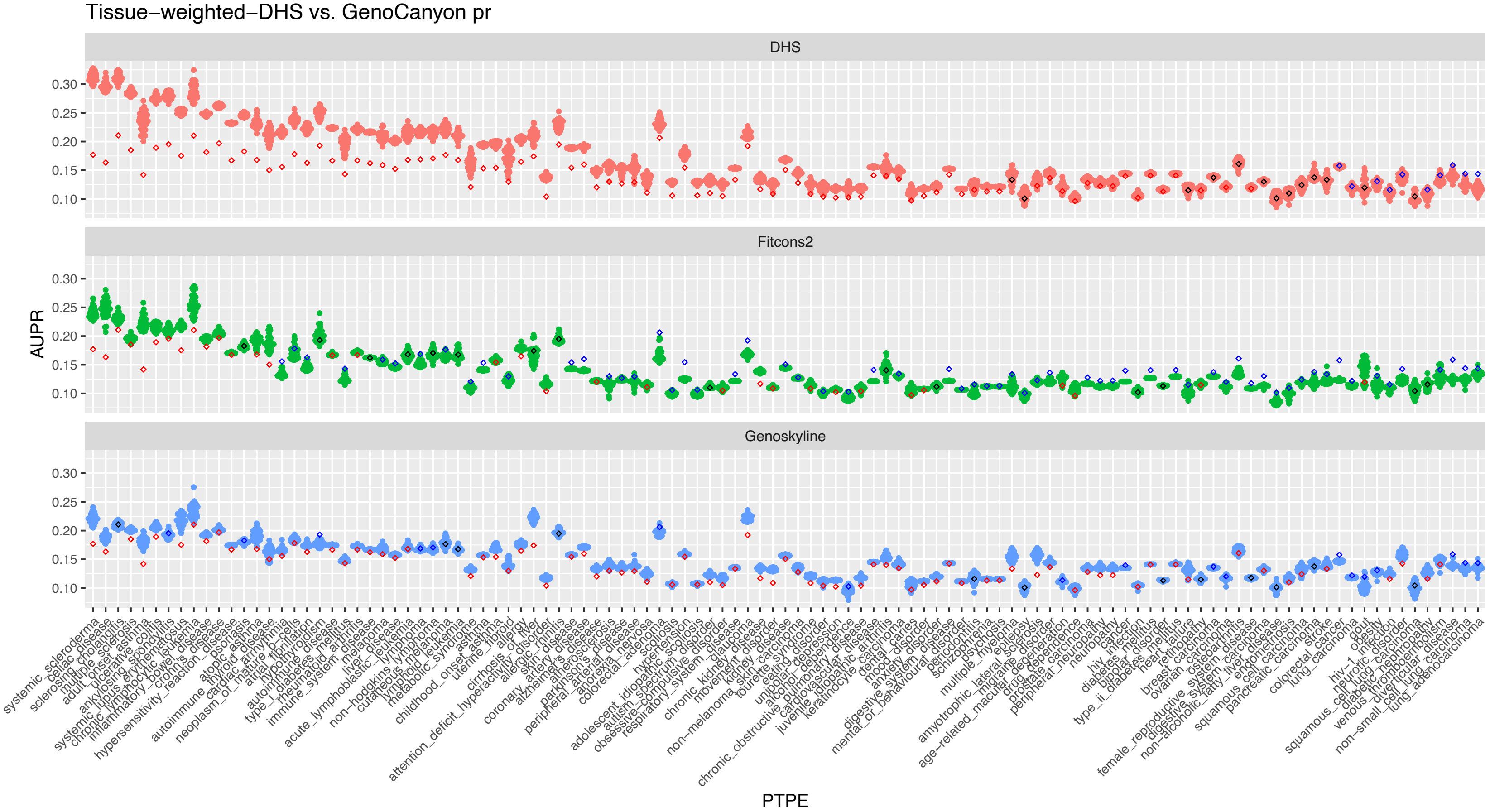

Figure A.6: *Tissue-weighted-DHS performance compared with GenoCanyon in 111 diseases.* X-axis delineates different diseases and y-axis is the performance in terms of area under the precision recall curve. The diamond represents the GenoCanyon (red: Tissue-Weighted better; black: comparable performance; blue: GenoCanyon better) and the colored dots are Tissue-Weighted with 30 replicates.

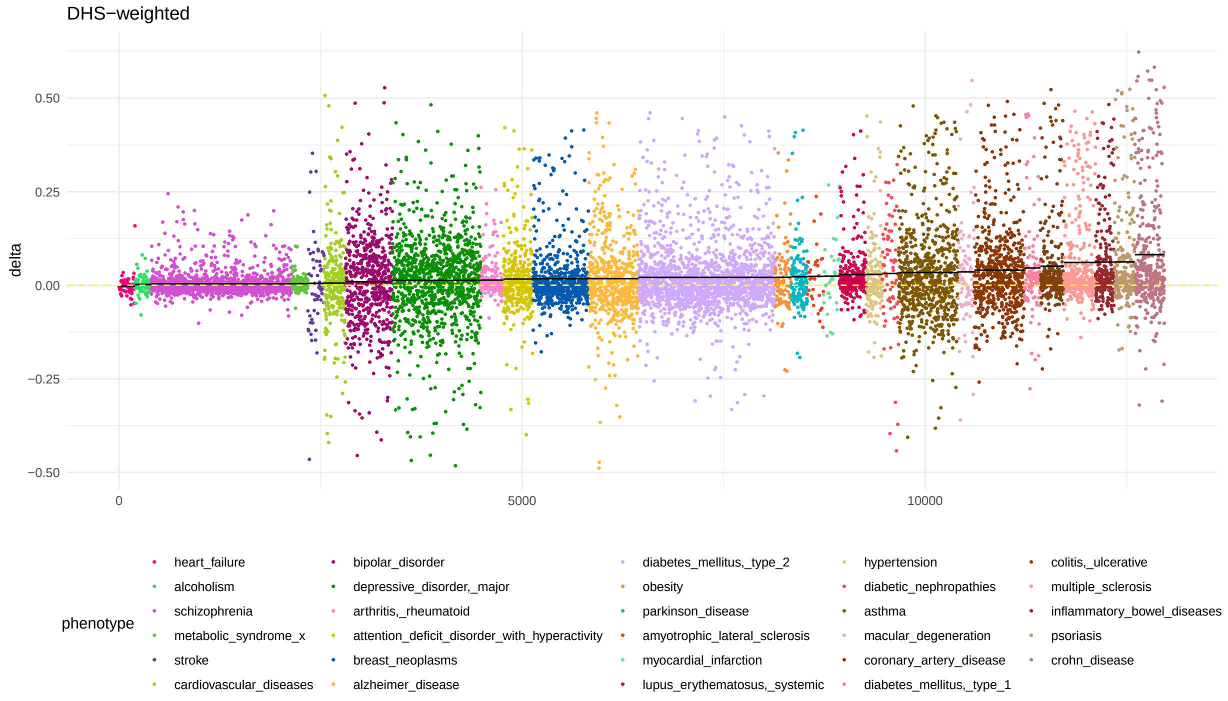

Figure A.7: *DHS-weighted variant score difference on DIVAN test dataset.* Y-axis delineates the delta value, where it is calculated as the difference in variant scores in disease-associated and the mean of 10 matched control variants. Each dot represents a disease-associated variant (along with 10 matched control variants). The diseases are color coded.

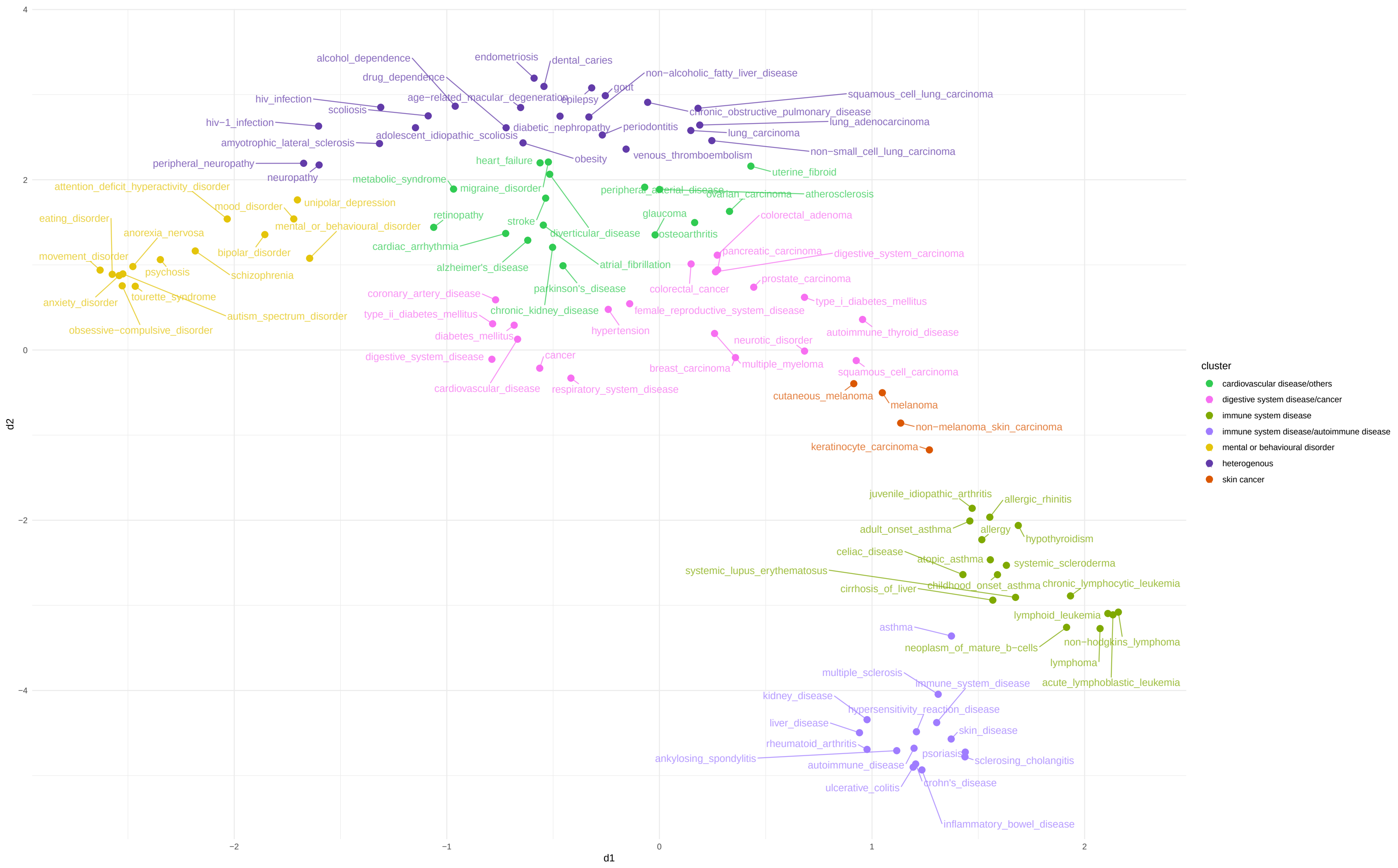

Figure A.8: *Umap plot shows 7 clusters of 111 diseases.* Hierarchical clustering was used to group diseases into 7 clusters.

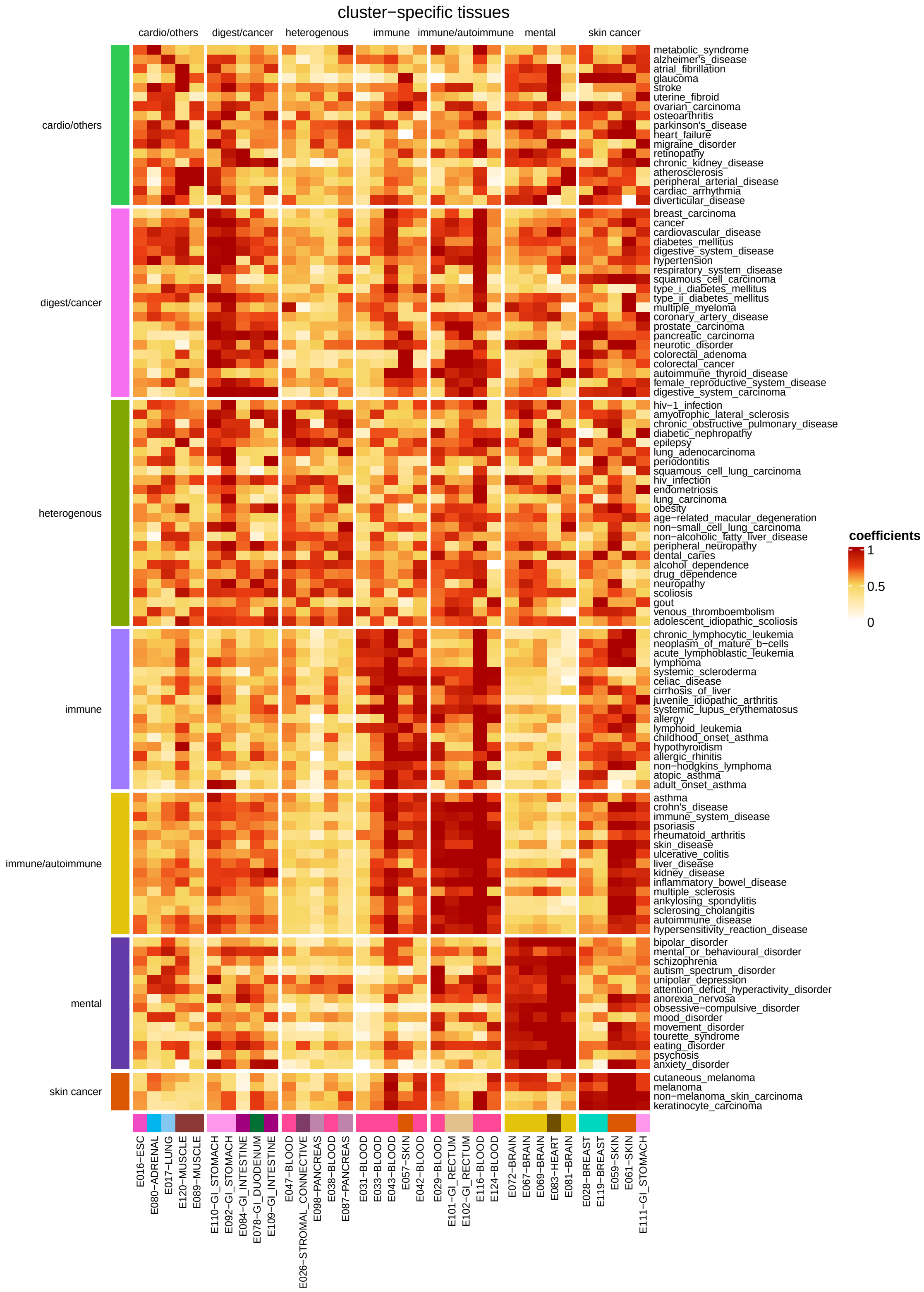

Figure A.9: *Heatmap plot of coefficients of 111 diseases.* Coefficients are regularized by each disease.

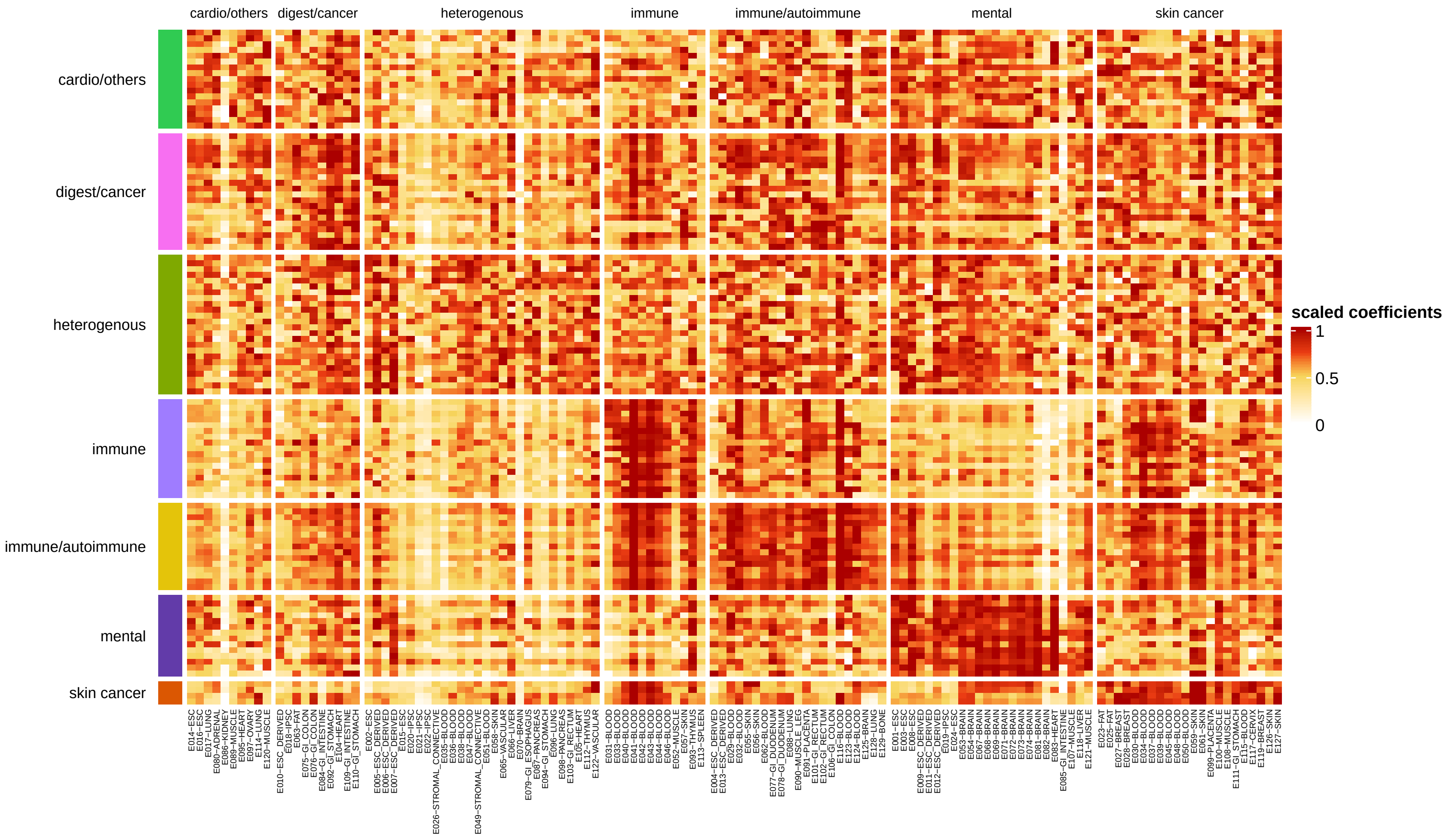

Figure A.10: *Heatmap plot of coefficients of 111 diseases on 5 cluster-specific tissues.* Coefficients are regularized by each disease. Tissue names are shown by ’Tissue name-group’ from 127 standard epigenomes.

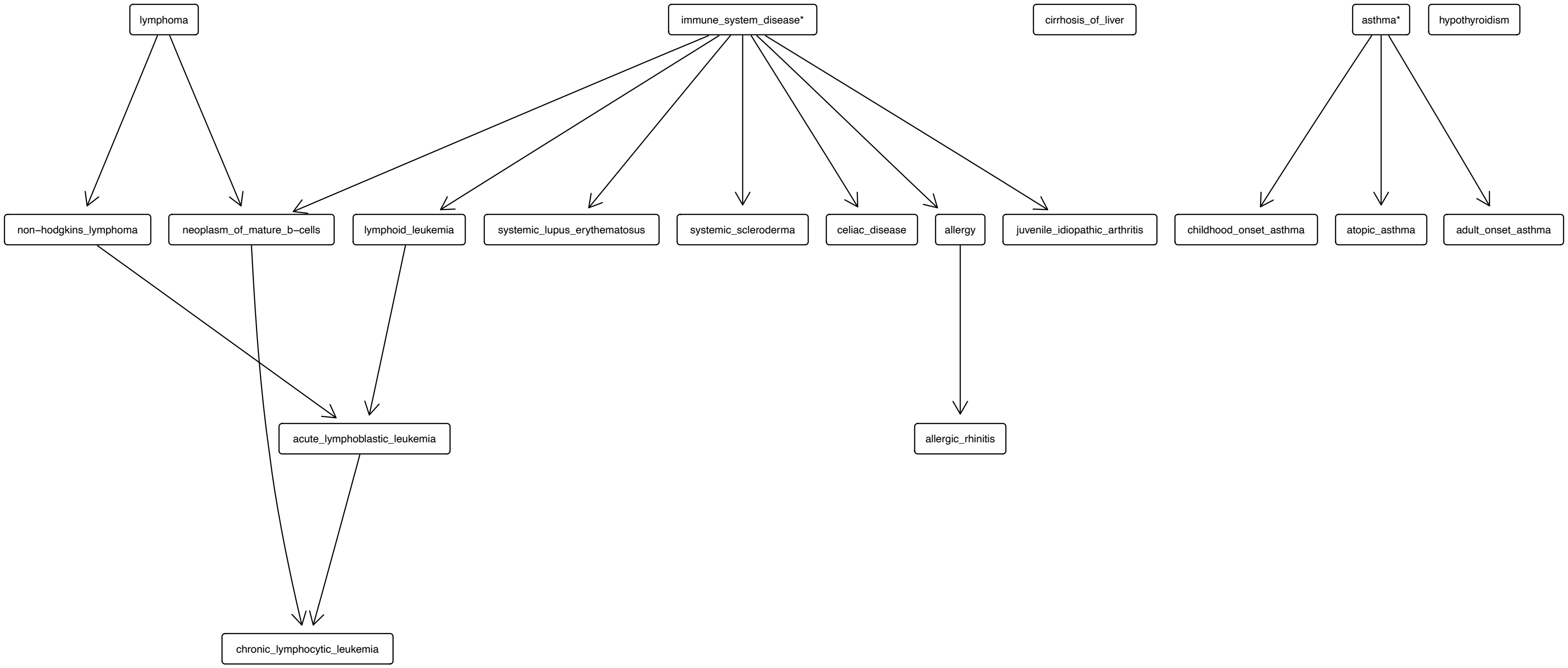

Figure A.11: *Disease relationships for immune1 cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

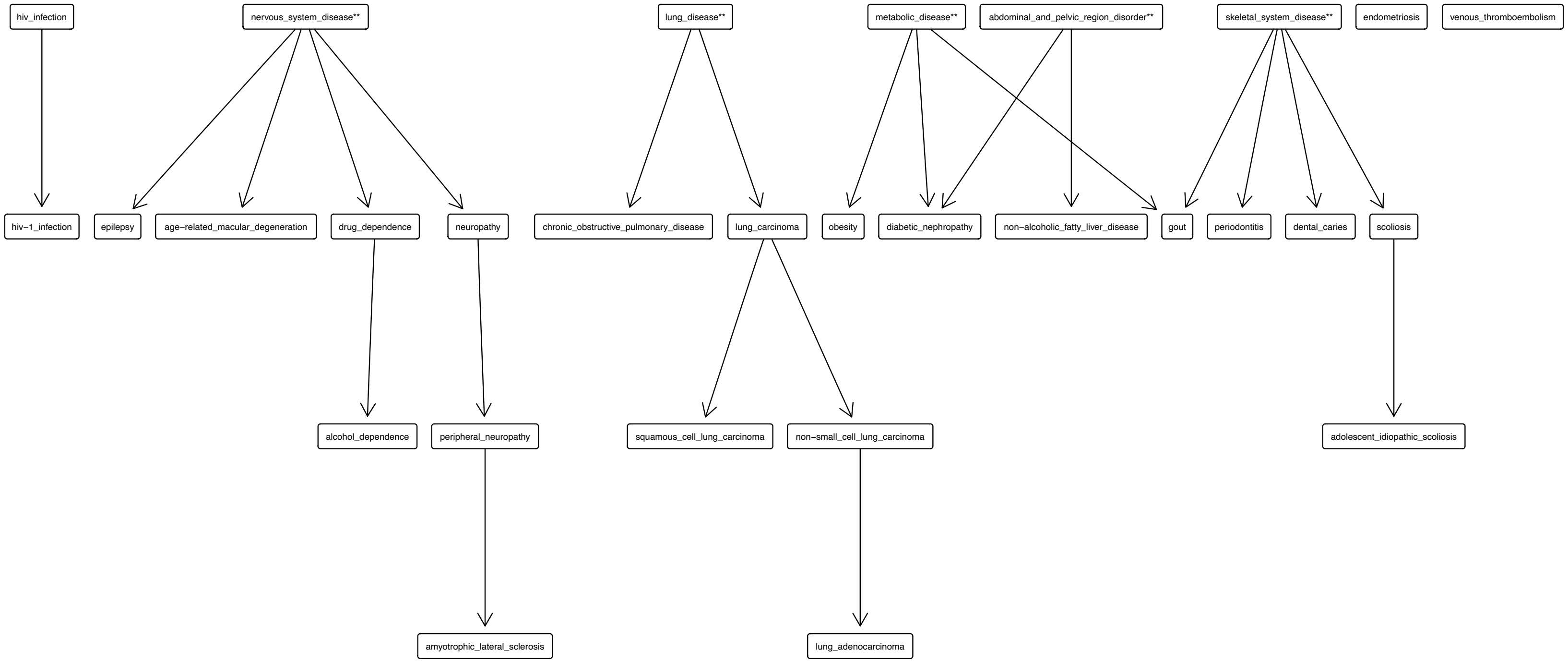

Figure A.12: *Disease relationships for others cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

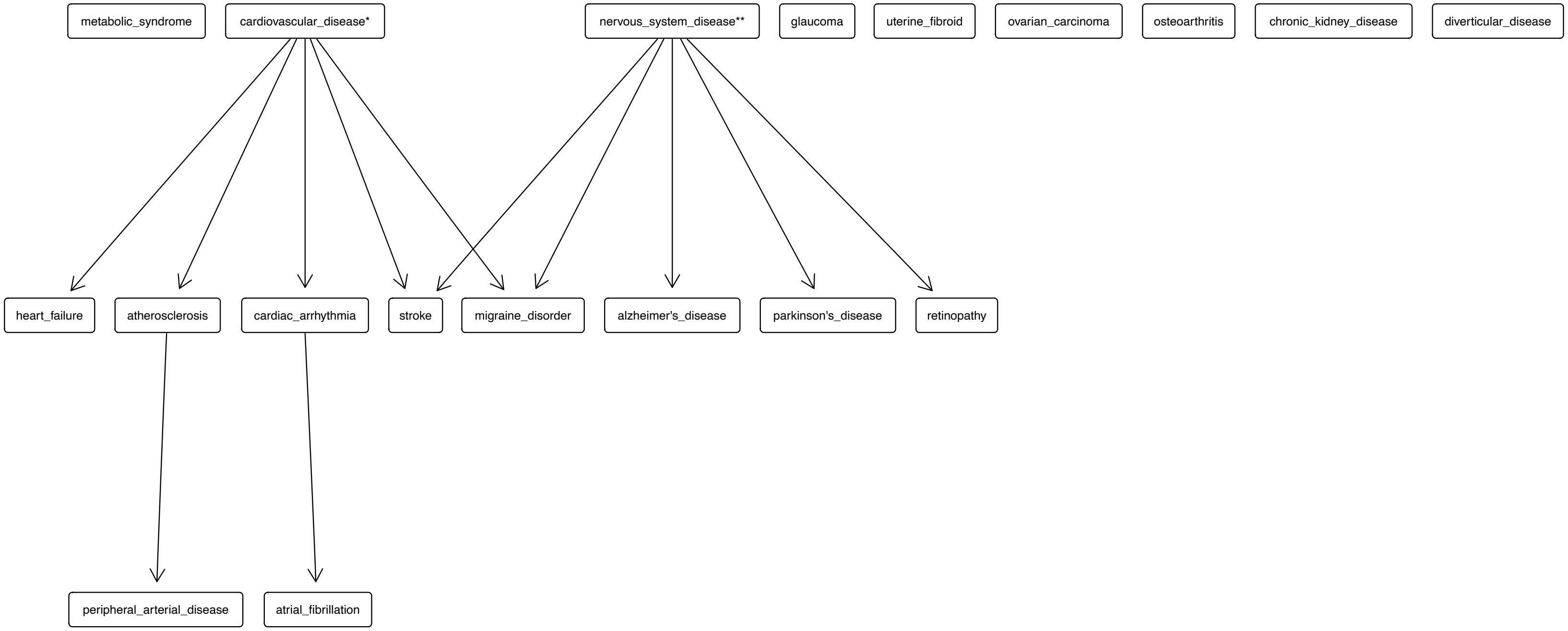

Figure A.13: *Disease relationships for cardiovasular disease and others cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

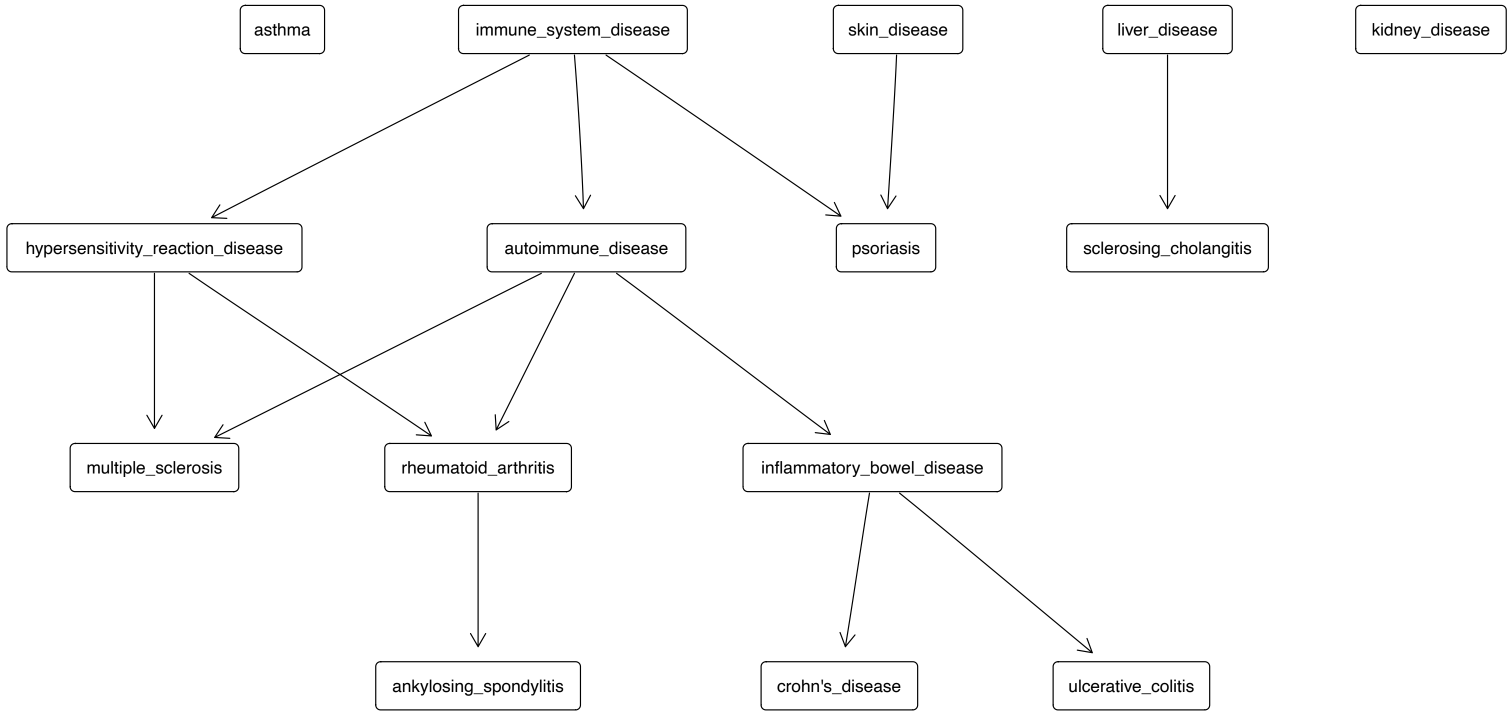

Figure A.14: *Disease relationships for immune2 cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

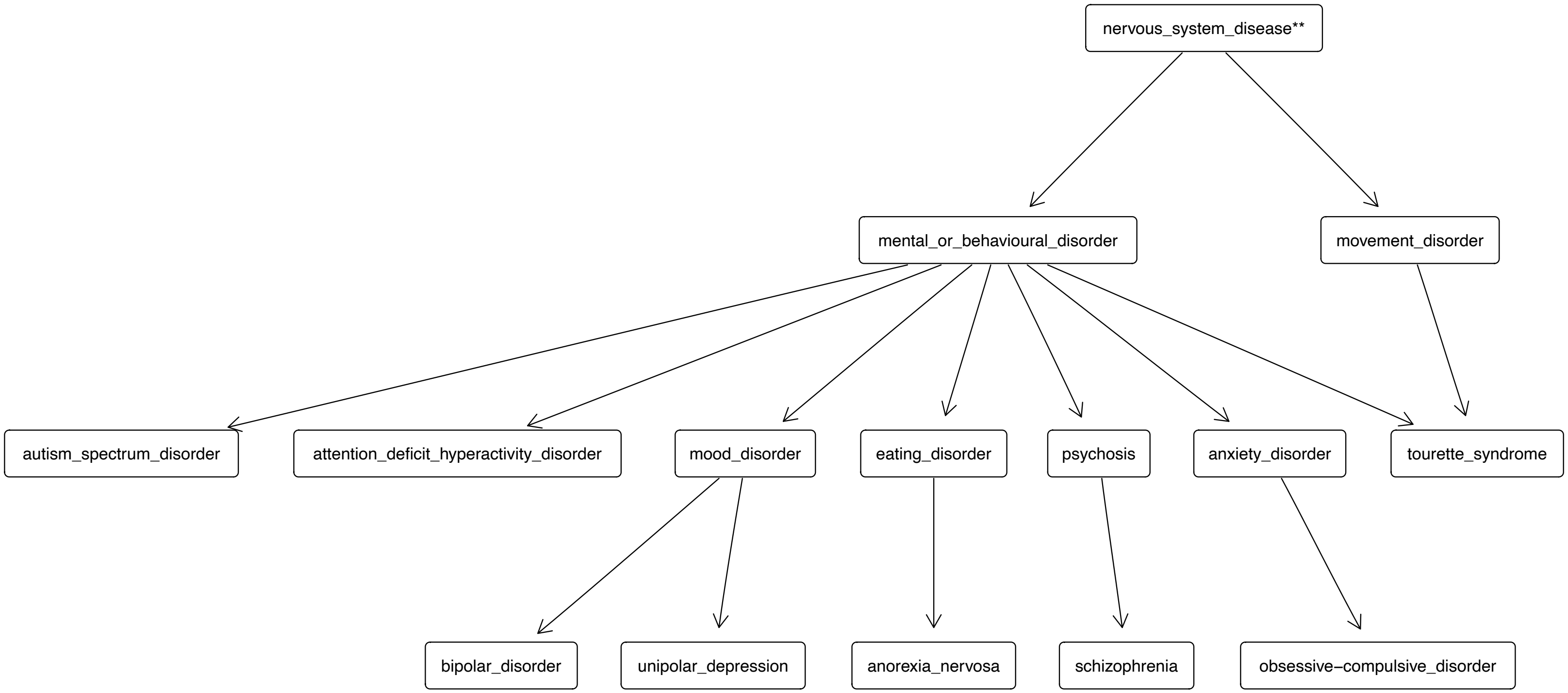

Figure A.15: *Disease relationships for mental or behavioural disorder cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

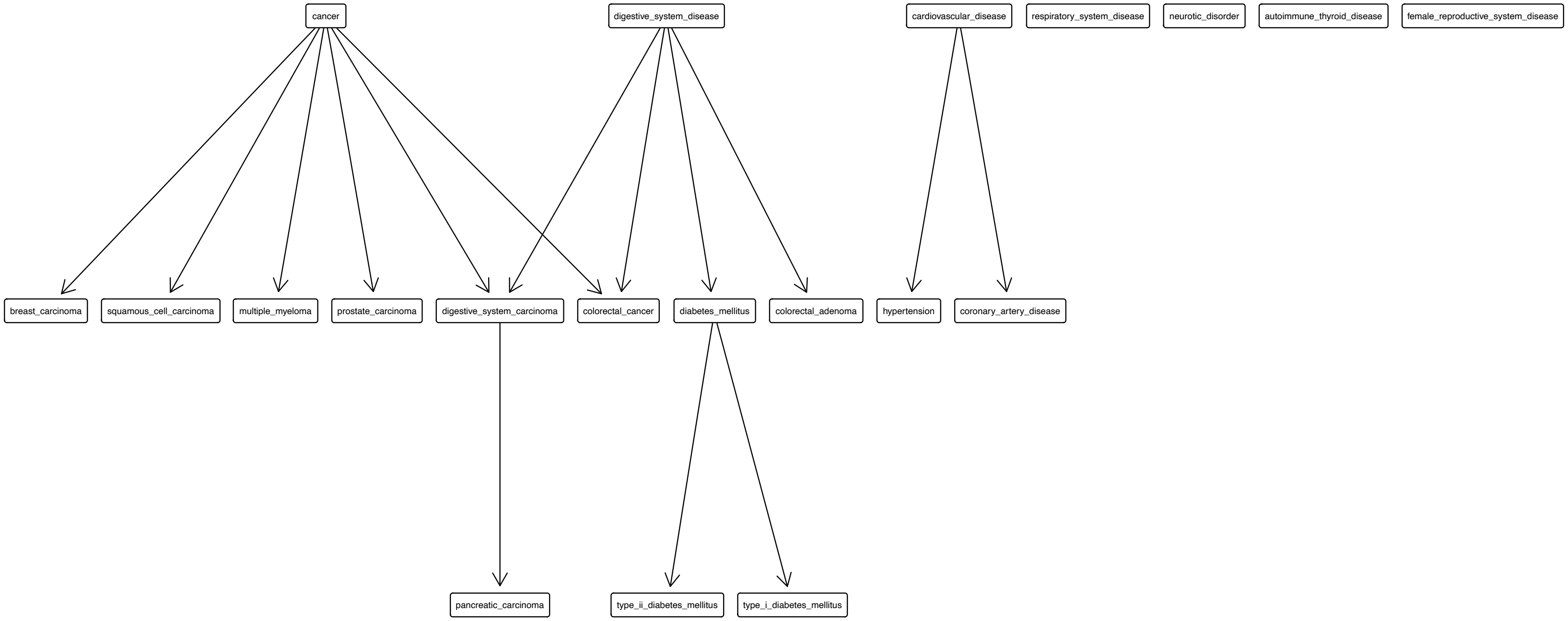

Figure A.16: *Disease relationships for digestive and cancer cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

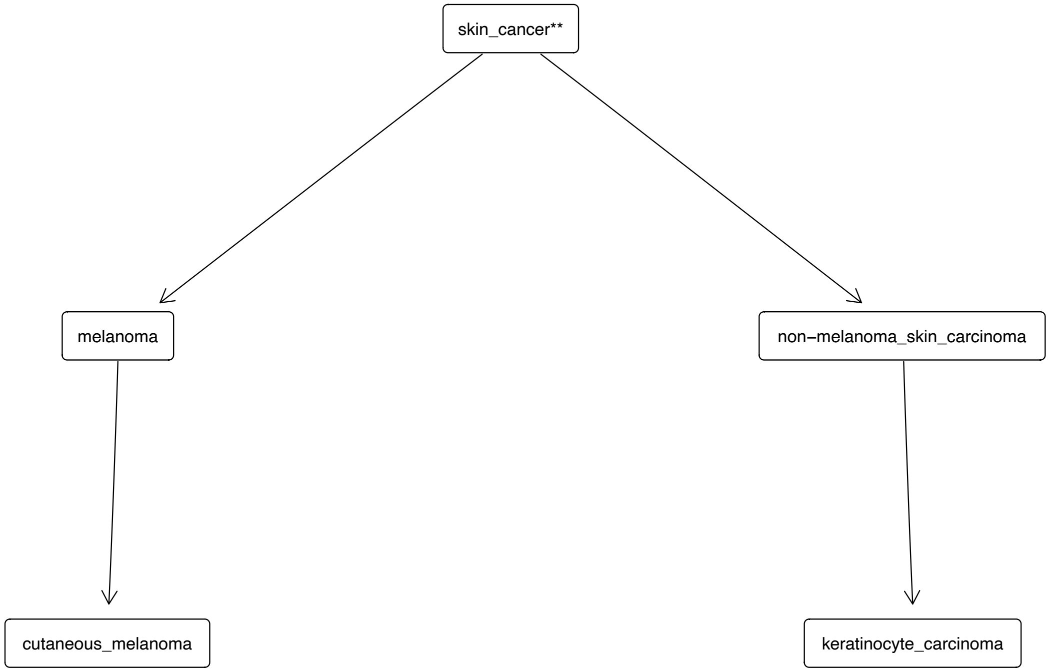

Figure A.17: *Disease relationships for skin cancer cluster.* The diseases placed at the top are more general than the diseases at the bottom. Arrow points from a more general term to a more specific term. A disease marked with one star indicates that it is not in this cluster but among the 111 diseases we studied. Diseases with two stars indicate that they are not among the 111 diseases.

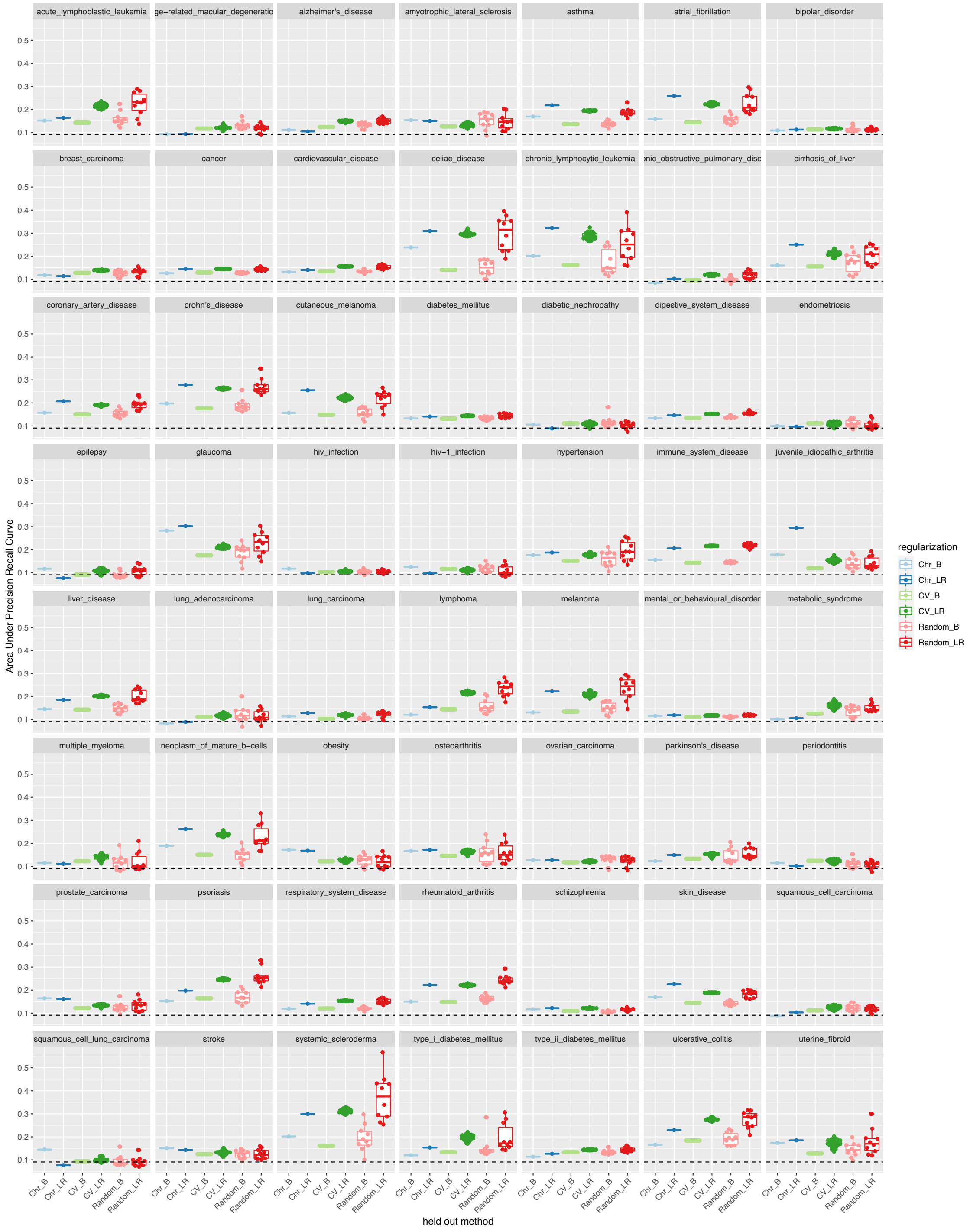

Figure A.18: *Performance of tissue-weighted (DHS) in different held-out strategies.* Chr-B: test set held out by chromosome, baseline; Chr-LR: test set held out by chromosome, logistic regression; CV-B: cross-validation, baseline; CV-LR: cross-validation logistic regression; random-B: randomly sampled test set, baseline; random-LR: randomly sample test set, logistic regression.

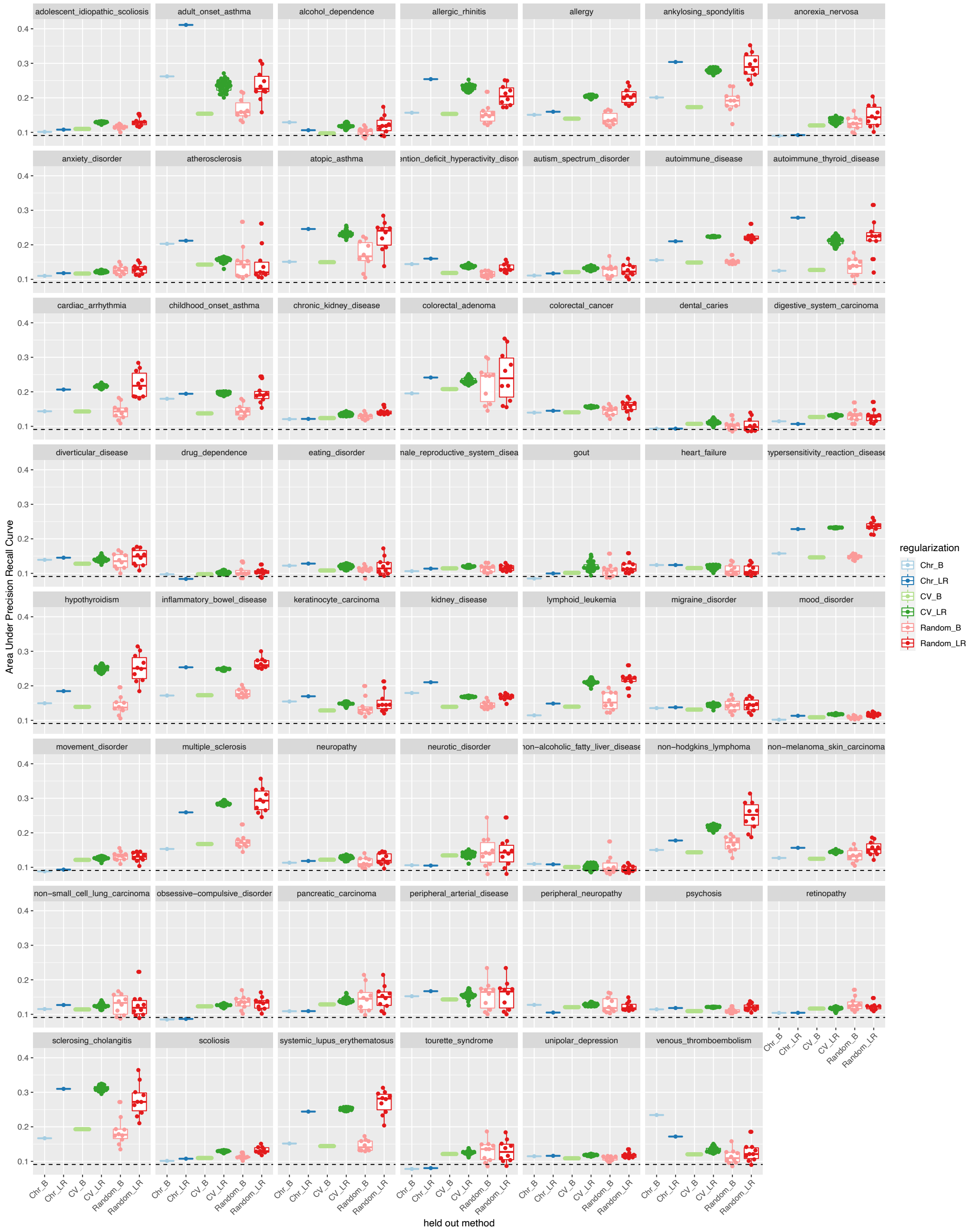

Figure A.19: *Performance of tissue-weighted (DHS) in different held-out strategies, continued.* Chr-B: test set held out by chromosome, baseline; Chr-LR: test set held out by chromosome, logistic regression; CV-B: cross-validation, baseline; CV-LR: cross-validation logistic regression; random-B: randomly sampled test set, baseline; random-LR: randomly sample test set, logistic regression.

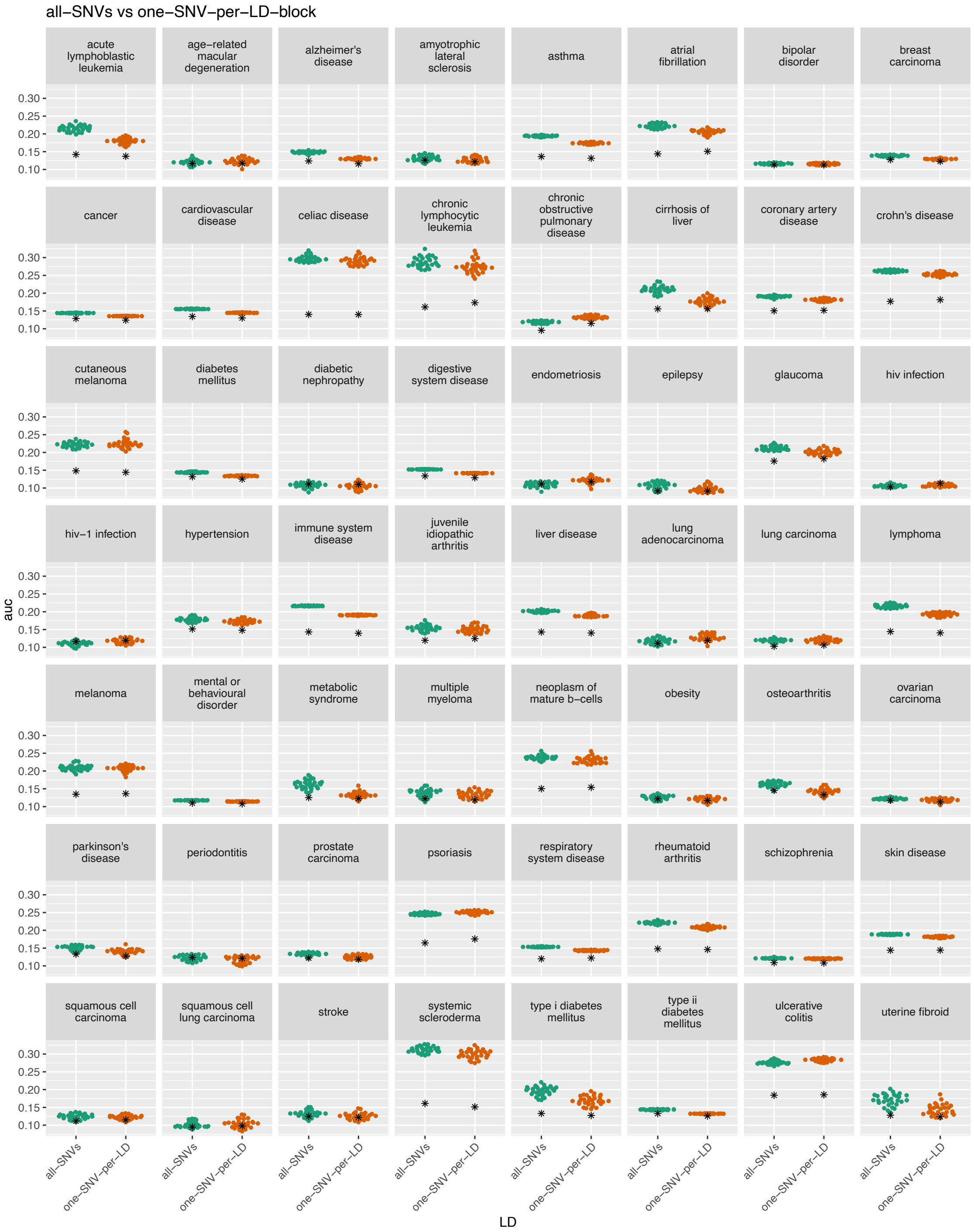

Figure A.20: *Performance of tissue-weighted (DHS) in all SNVs or representative SNVs (one SNV per LD block).* Colored dots represent the performance of tissue-weighted (DHS) in all SNVs or representative SNVs. Stars represent the baseline performance (tissue-mean DHS) in all SNVs or representative SNVs.

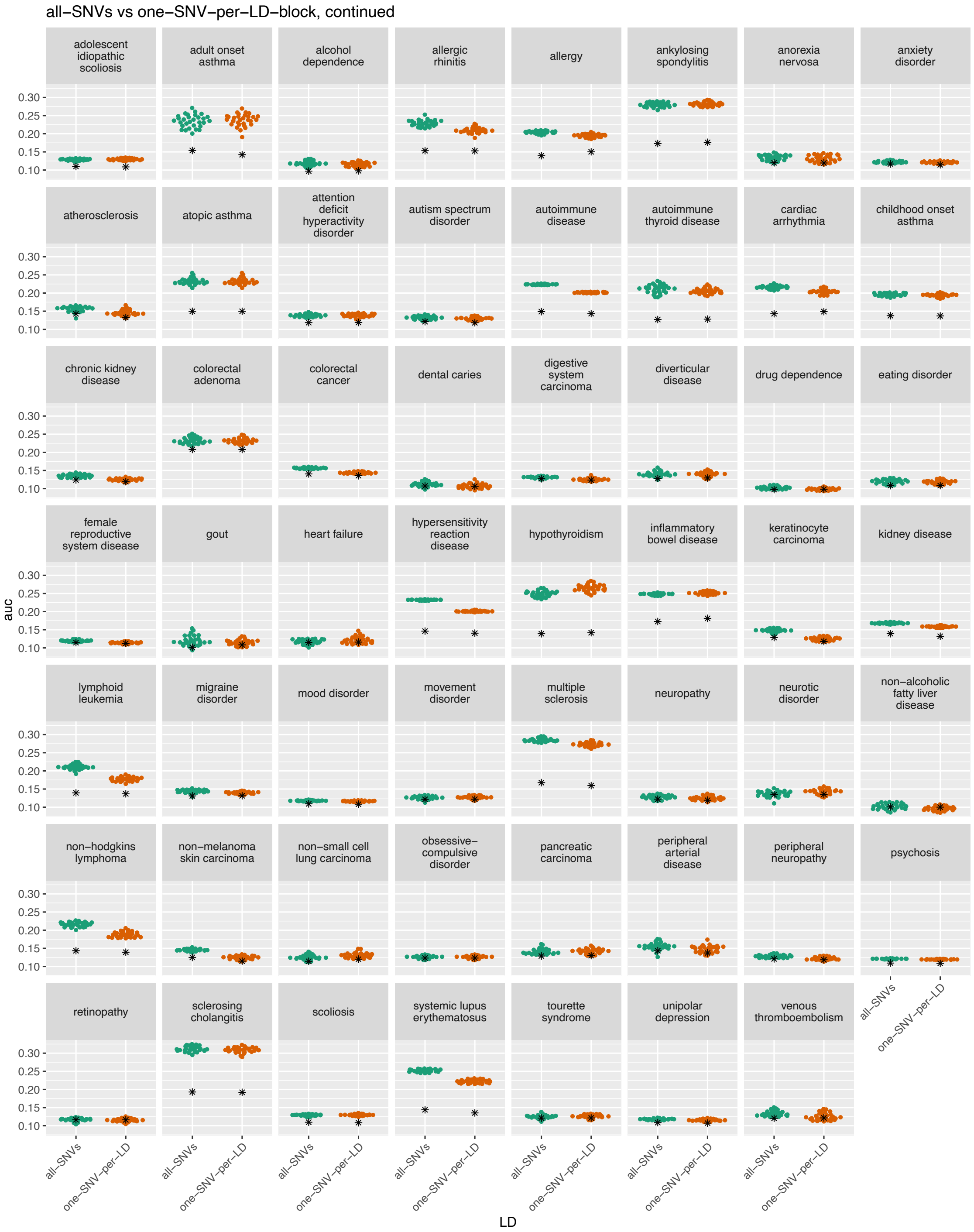

Figure A.21: *Performance of tissue-weighted (DHS) in all SNVs or representative SNVs (one SNV per LD block), continued.* Colored dots represent the performance of tissue-weighted (DHS) in all SNVs or representative SNVs. Stars represent the baseline performance (tissue-mean DHS) in all SNVs or representative SNVs.

Appendix B Supplemental materials for aim2

B.1 Supplemental data legend

The supplemental data can be found at: `http://d-scholarship.pitt.edu/46026/2/sup_data_qianqian_dissertation.zip`

B.1.1 Relative performance of the information sharing model for all disease pairs

`Filename: sup_dat_auc_all.csv.gz`

The first column denotes the EFO name of Disease 1 (D_1). Column #2 is the EFO name of Disease 2 (D_2). Column #3 shows the performance measured by average precision where only SNVs in D_1 are used to train the model (without the information-sharing model). Columns #4 shows the performance absolute increased (or decreased) measured by average precision using the SNVs in D_1D_2 (using the information sharing model). Column #5 is the relative weight assigned. Column #6 is the relative performance of the information-sharing model using D_1D_2. Column #7 is the model similarity between D_1 and D_2. Column #8 is the number of SNVs in D_1, #9 is the number of SNVs in D_2 and #10 is the number of SNVs in D_2 but excluding overlapping SNVs with D1.

B.1.2 Wilcox sign rank test p-value and corrected p-value for top 10 disease pairs

`Filename: sup_dat_top_pvalue.csv.gz`

The first column denotes the EFO name of Disease 1 (D_1). Column #2 is the EFO name of Disease 2 (D_2). Column #3 is the relative performance of the information-sharing model using D_1D_2. Column #4 Wilcox sign rank test p-value Column #5 is p value using FDR correction

B.2 Supplemental figures and tables

Disease name	Included in aim 2
acute_lymphoblastic_leukemia	FALSE
adolescent_idiopathic_scoliosis	TRUE
adult_onset_asthma	TRUE
age-related_macular_degeneration	TRUE
alcohol_dependence	TRUE
allergic_rhinitis	TRUE
allergy	FALSE
alzheimer's_disease	TRUE
amyotrophic_lateral_sclerosis	TRUE
ankylosing_spondylitis	TRUE
anorexia_nervosa	TRUE
anxiety_disorder	FALSE
asthma	FALSE
atherosclerosis	FALSE
atopic_asthma	TRUE
atrial_fibrillation	TRUE
attention_deficit_hyperactivity_disorder	TRUE
autism_spectrum_disorder	TRUE
autoimmune_disease	FALSE
autoimmune_thyroid_disease	TRUE
bipolar_disorder	TRUE
breast_carcinoma	TRUE
cancer	FALSE
cardiac_arrhythmia	FALSE
cardiovascular_disease	FALSE
celiac_disease	TRUE
childhood_onset_asthma	TRUE
chronic_kidney_disease	FALSE
chronic_lymphocytic_leukemia	TRUE
chronic_obstructive_pulmonary_disease	TRUE
cirrhosis_of_liver	TRUE

Table B.1: *Diseases used in aim 2, total of 68.*

Disease name	Included in Aim 2
colorectal_adenoma	TRUE
colorectal_cancer	TRUE
coronary_artery_disease	TRUE
crohn's_disease	TRUE
cutaneous_melanoma	TRUE
dental_caries	TRUE
diabetes_mellitus	FALSE
diabetic_nephropathy	TRUE
digestive_system_carcinoma	FALSE
digestive_system_disease	FALSE
diverticular_disease	TRUE
drug_dependence	FALSE
eating_disorder	FALSE
endometriosis	TRUE
epilepsy	TRUE
female_reproductive_system_disease	FALSE
glaucoma	TRUE
gout	TRUE
heart_failure	TRUE
hiv_infection	FALSE
hiv-1_infection	TRUE
hypersensitivity_reaction_disease	FALSE
hypertension	TRUE
hypothyroidism	TRUE
immune_system_disease	FALSE
inflammatory_bowel_disease	FALSE
juvenile_idiopathic_arthritis	TRUE
keratinocyte_carcinoma	FALSE
kidney_disease	FALSE
liver_disease	FALSE
lung_adenocarcinoma	TRUE
lung_carcinoma	FALSE
lymphoid_leukemia	FALSE
lymphoma	FALSE
melanoma	FALSE
mental_or_behavioural_disorder	FALSE
metabolic_syndrome	TRUE
migraine_disorder	TRUE

Table B.2: *Diseases used in aim 2, total of 68, continued.*

Disease name	Included in Aim 2
mood_disorder	FALSE
movement_disorder	FALSE
multiple_myeloma	TRUE
multiple_sclerosis	TRUE
neoplasm_of_mature_b-cells	FALSE
neuropathy	FALSE
neurotic_disorder	TRUE
non-alcoholic_fatty_liver_disease	TRUE
non-hodgkins_lymphoma	FALSE
non-melanoma_skin_carcinoma	FALSE
non-small_cell_lung_carcinoma	FALSE
obesity	TRUE
obsessive-compulsive_disorder	TRUE
osteoarthritis	TRUE
ovarian_carcinoma	TRUE
pancreatic_carcinoma	TRUE
parkinson's_disease	TRUE
periodontitis	TRUE
peripheral_arterial_disease	TRUE
peripheral_neuropathy	FALSE
prostate_carcinoma	TRUE
psoriasis	TRUE
psychosis	FALSE
respiratory_system_disease	FALSE
retinopathy	FALSE
rheumatoid_arthritis	FALSE
schizophrenia	TRUE
sclerosing_cholangitis	TRUE
scoliosis	FALSE
skin_disease	FALSE
squamous_cell_carcinoma	FALSE
squamous_cell_lung_carcinoma	TRUE
stroke	TRUE
systemic_lupus_erythematosus	TRUE
systemic_scleroderma	TRUE
tourette_syndrome	TRUE
type_i_diabetes_mellitus	TRUE
type_ii_diabetes_mellitus	TRUE
ulcerative_colitis	TRUE
unipolar_depression	TRUE
uterine_fibroid	TRUE
venous_thromboembolism	TRUE

Table B.3: *Diseases used in aim 2, total of 68, continued.*

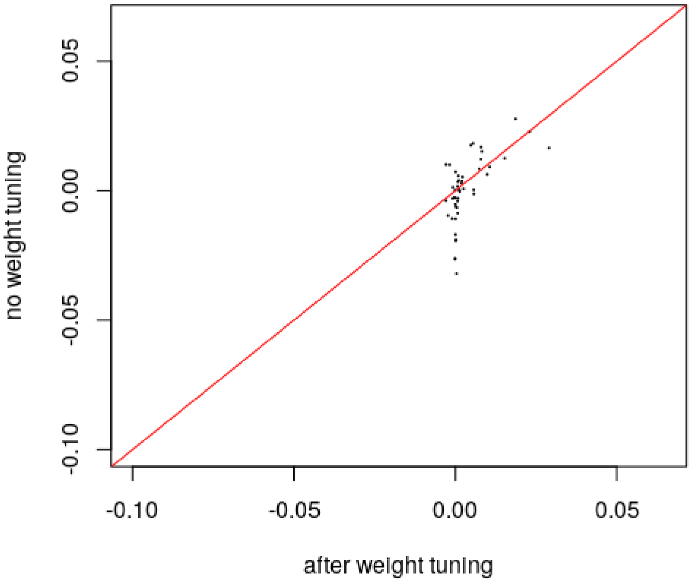

Figure B.1: *Before and after weight tuning.* We randomly pick 50 disease pairs. The x-axis is the relative performance of the disease pairs using our weight tuning strategy, while the y-axis is the relative performance of the disease pairs without weight tuning (we assign the same overall weight to D1 and D2).

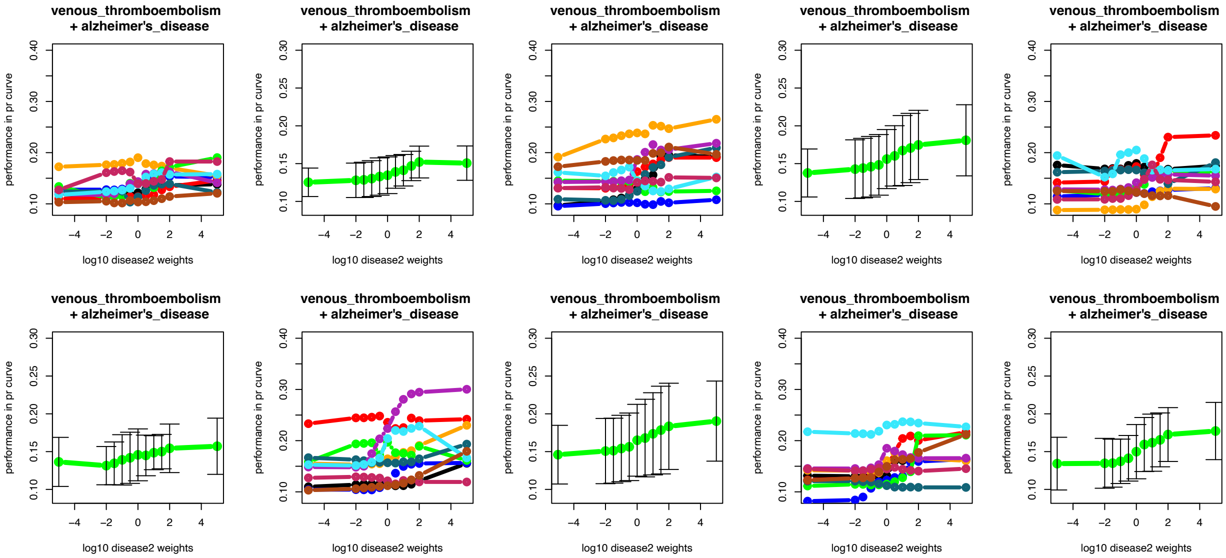

Figure B.2: *Performance in AUPR curve in the inner loop for the example disease pair: venous thromboembolism and Alzheimer's disease.* The X-axis is the log 10 weight w, and the y-axis is the average precision of the validation set.

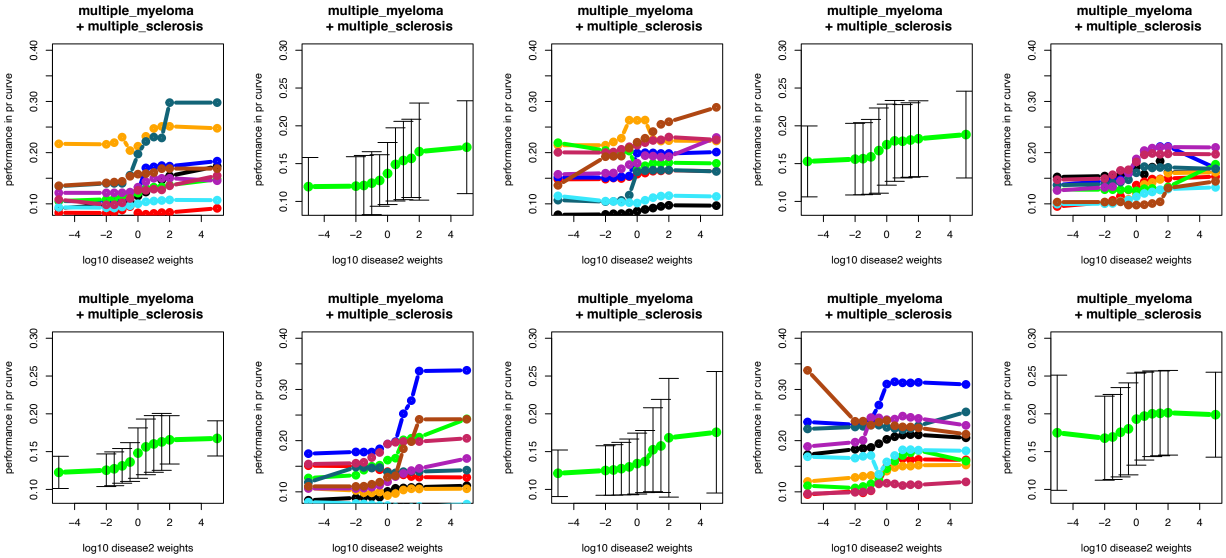

Figure B.3: *Performance in AUPR curve in the inner loop for the example disease pair: multiple myeloma and multiple sclerosis.* X axis is the log 10 weight w, y axis is average precision of the validation set.

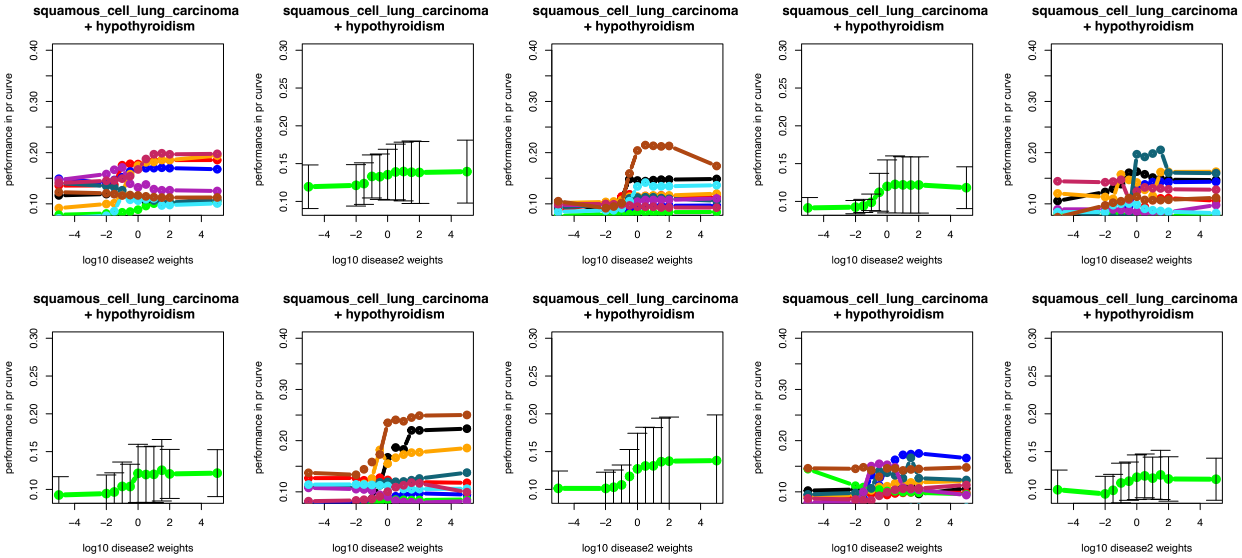

Figure B.4: *Performance in AUPR curve in the inner loop for the example disease pair: squamous cell lung carcinoma and hypothyroidism.* X-axis is the log 10 weight w, y-axis is the average precision of the validation set.

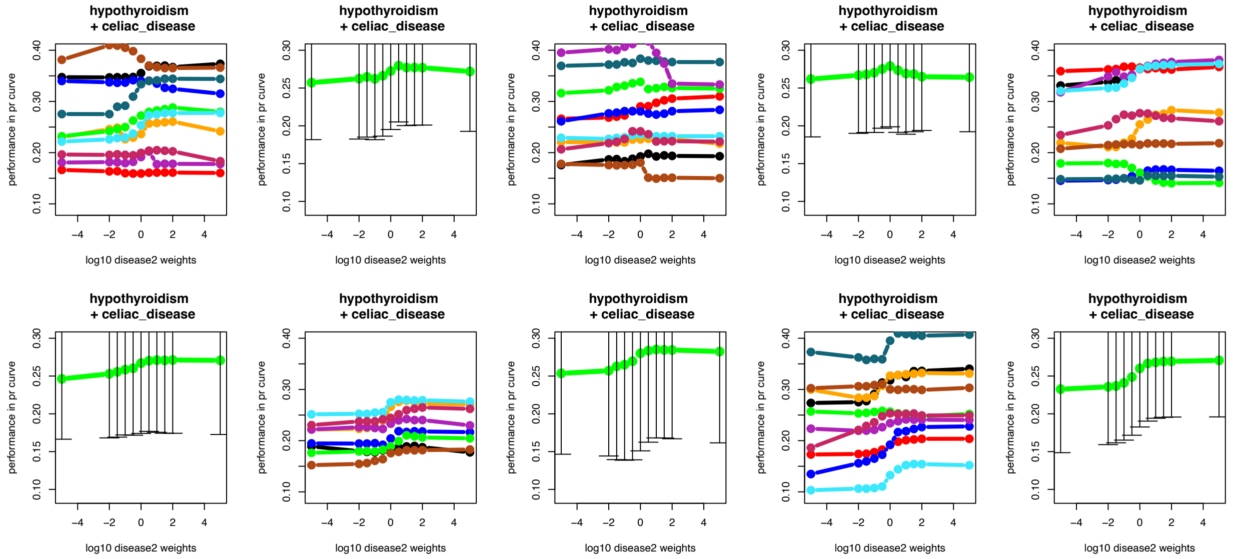

Figure B.5: *Performance in AUPR curve in innerloop for the example disease pair: hypothyroidism and celiac disease.* The X-axis is the log 10 weight w, and the y-axis is the average precision of the validation set.

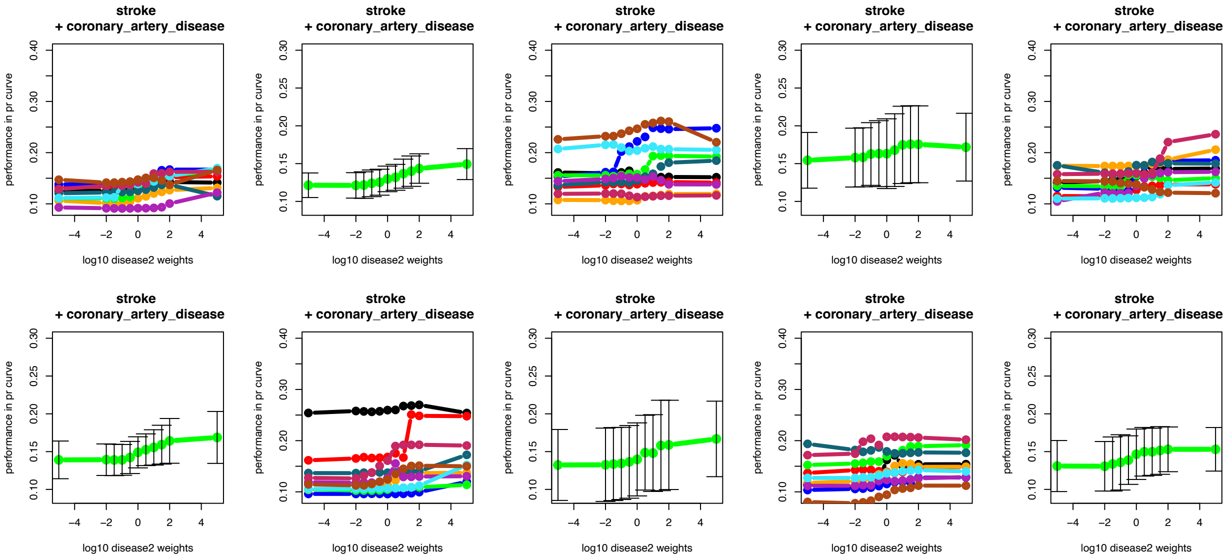

Figure B.6: *Performance in AUPR curve in the inner loop for the example disease pair: stroke and coronary artery disease.* The X-axis is the log 10 weight w, the y-axis is the average precision of the validation set.

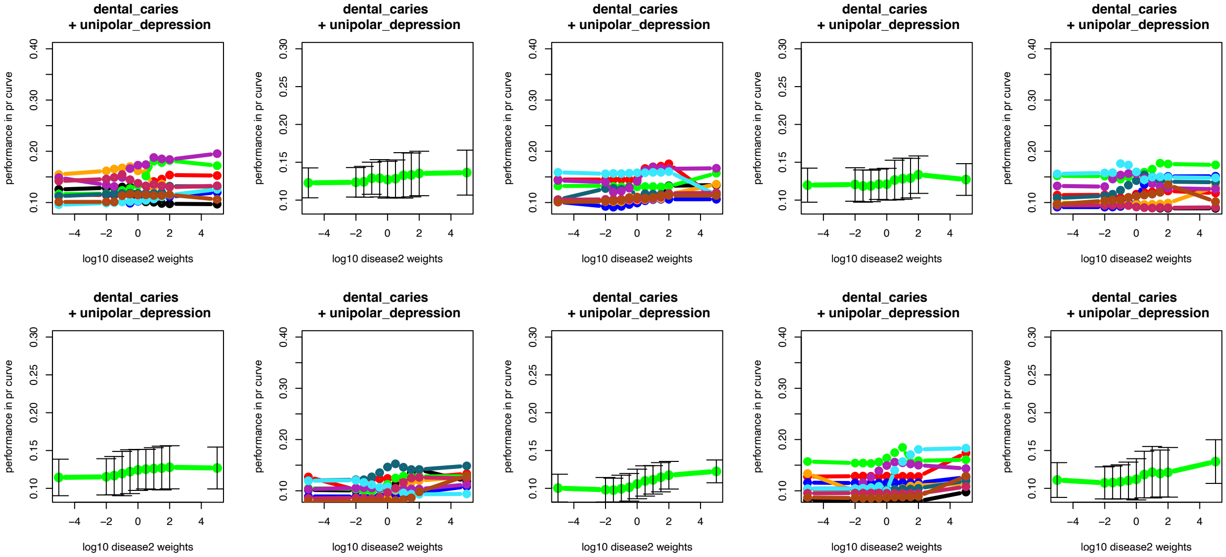

Figure B.7: *Performance in AUPR curve in the inner loop for the example disease pair: dental caries and unipolar depression.* X axis is the log 10 weight w, y-axis is the average precision of the validation set.

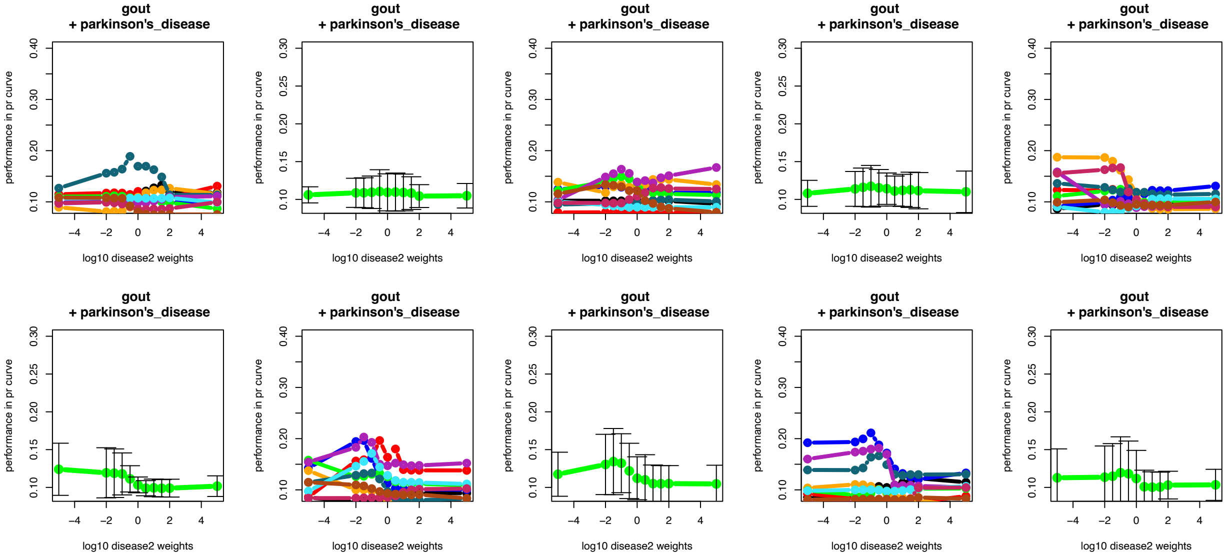

Figure B.8: *Performance in AUPR curve in the inner loop for the example disease pair: gout and Parkinson's disease.* The X-axis is the log 10 weight w, the y-axis is the average precision of the validation set.

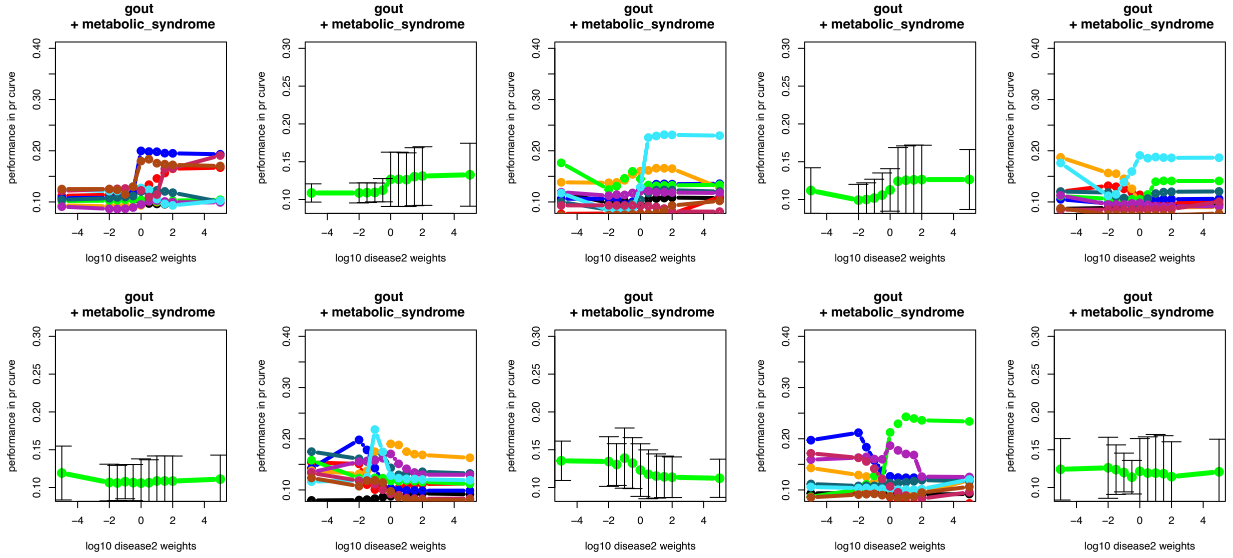

Figure B.9: *Performance in AUPR curve in the inner loop for the example disease pair: gout and metabolic syndrome.* The X-axis is the log 10 weight w, y-axis is the average precision of the validation set.

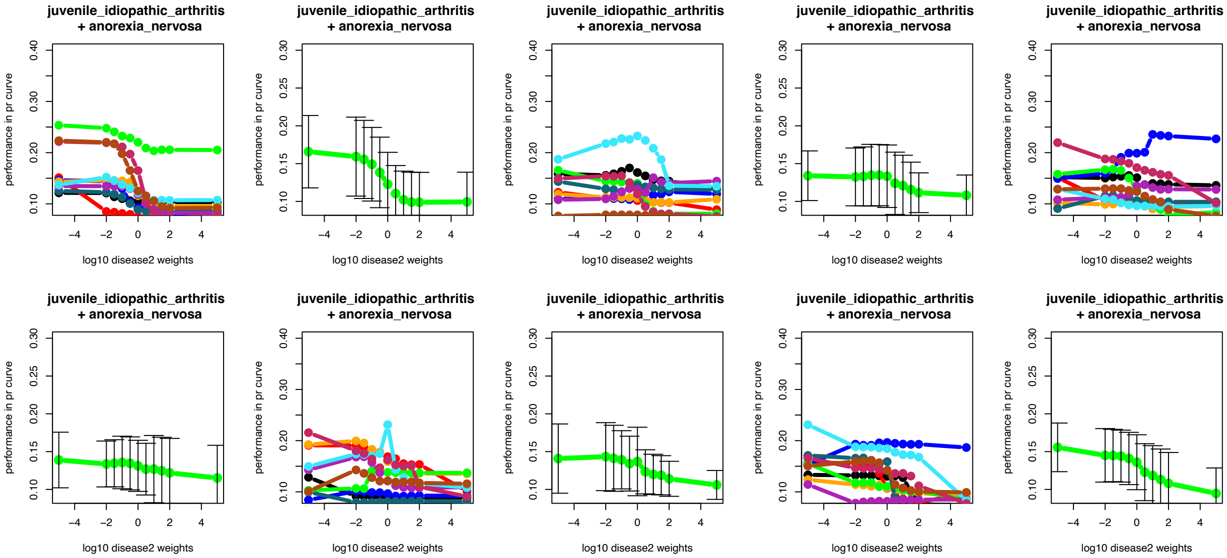

Figure B.10: *Performance in AUPR curve in the inner loop for the juvenile idiopathic arthritis and anorexia nervosa.* The X-axis is the log 10 weight w, y-axis is the average precision of the validation set.

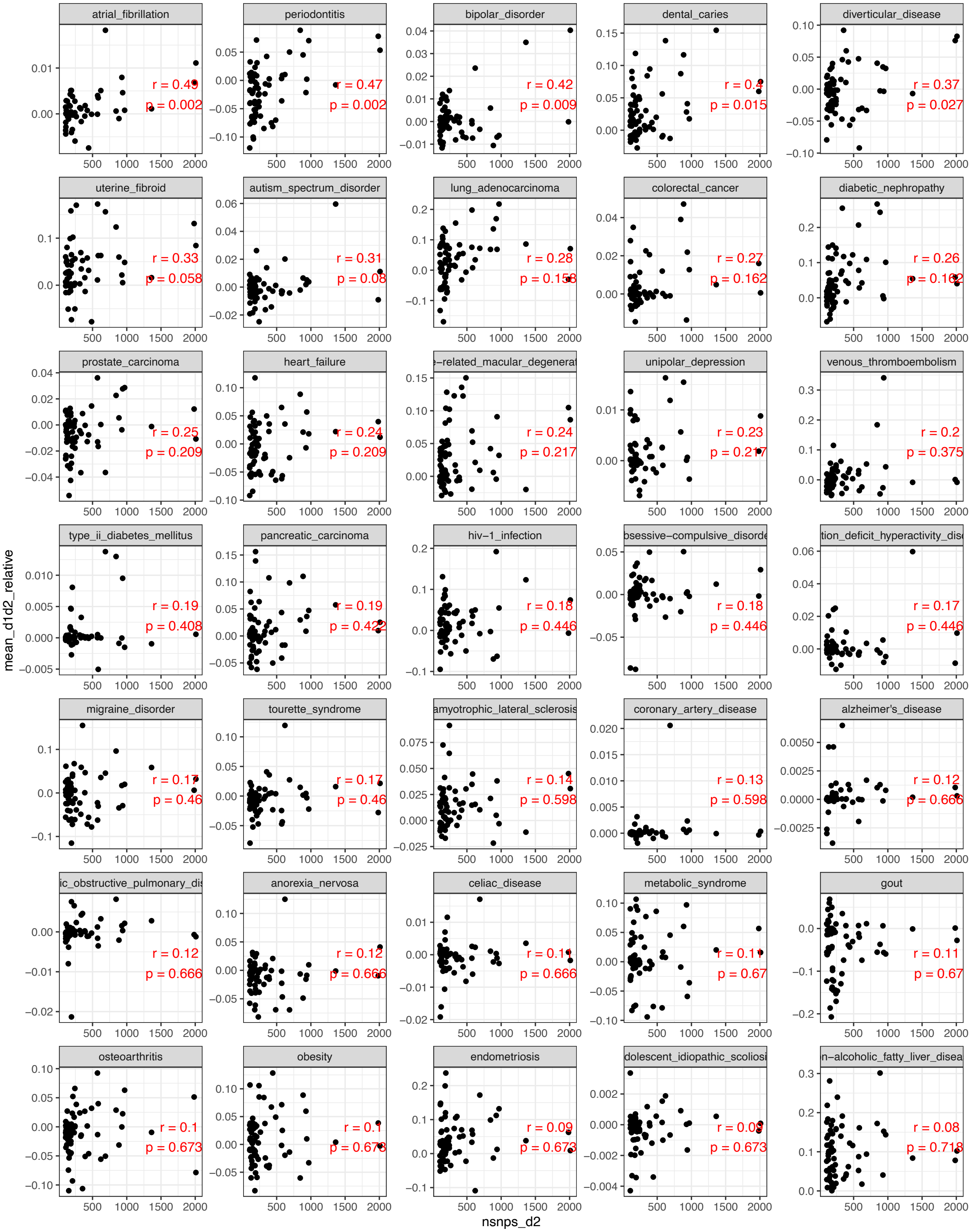

Figure B.11: *Number of SNVs in D_2 and relative performance.*

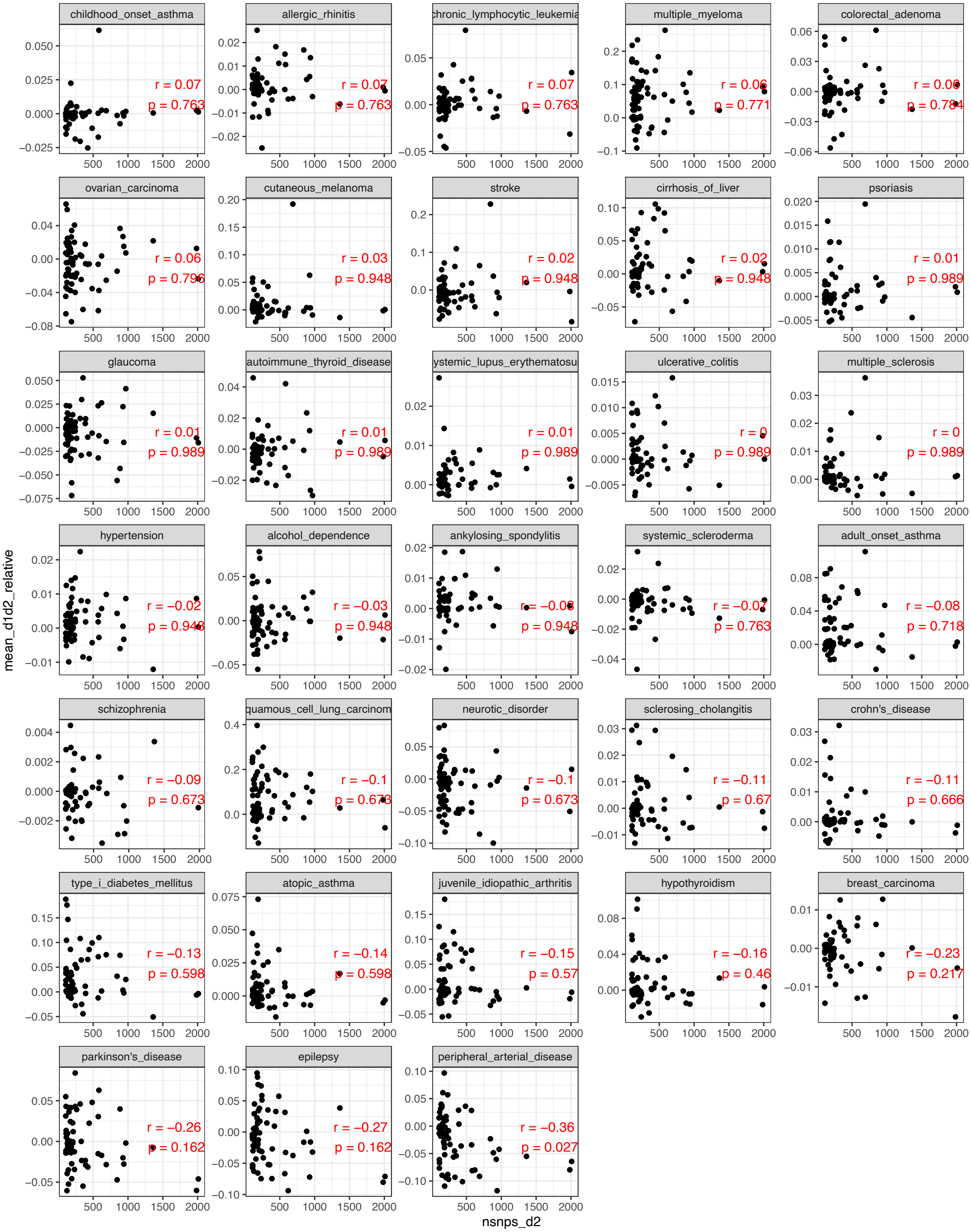

Figure B.12: *Number of SNVs in D_2 and relative performance, continued.*

d1	corre	pval	fdr_adj_p_values
atrial_fibrillation	0.489	0.000	0.002
periodontitis	0.471	0.000	0.002
bipolar_disorder	0.421	0.000	0.009
dental_caries	0.398	0.001	0.015
diverticular_disease	0.371	0.002	0.027
uterine_fibroid	0.333	0.006	0.058
autism_spectrum_disorder	0.315	0.009	0.080
lung_adenocarcinoma	0.282	0.021	0.158
colorectal_cancer	0.268	0.028	0.162
diabetic_nephropathy	0.264	0.031	0.162
prostate_carcinoma	0.247	0.044	0.209
heart_failure	0.245	0.046	0.209
age-related_macular_degeneration	0.236	0.054	0.217
unipolar_depression	0.233	0.057	0.217
venous_thromboembolism	0.200	0.105	0.375
type_ii_diabetes_mellitus	0.192	0.120	0.408
pancreatic_carcinoma	0.187	0.130	0.422
hiv-1_infection	0.180	0.145	0.446
obsessive-compulsive_disorder	0.177	0.151	0.446
attention_deficit_hyperactivity_disorder	0.175	0.157	0.446
migraine_disorder	0.170	0.170	0.460
tourette_syndrome	0.167	0.178	0.460
amyotrophic_lateral_sclerosis	0.139	0.262	0.598

Table B.4: *Correlation of the number of SNVs in D_2 vs relative performance.*

d1	corre	pval	fdr_adj_p_values
coronary_artery_disease	0.134	0.281	0.598
alzheimer's_disease	0.120	0.333	0.666
chronic_obstructive_pulmonary_disease	0.119	0.338	0.666
anorexia_nervosa	0.116	0.351	0.666
celiac_disease	0.115	0.356	0.666
metabolic_syndrome	0.108	0.386	0.670
gout	0.107	0.391	0.670
osteoarthritis	0.097	0.436	0.673
obesity	0.096	0.442	0.673
endometriosis	0.092	0.459	0.673
adolescent_idiopathic_scoliosis	0.091	0.465	0.673
non-alcoholic_fatty_liver_disease	0.082	0.508	0.718
childhood_onset_asthma	0.071	0.571	0.763
allergic_rhinitis	0.070	0.575	0.763
chronic_lymphocytic_leukemia	0.066	0.595	0.763
multiple_myeloma	0.063	0.613	0.771
colorectal_adenoma	0.059	0.634	0.784
ovarian_carcinoma	0.055	0.656	0.796
cutaneous_melanoma	0.032	0.799	0.948
stroke	0.022	0.860	0.948
cirrhosis_of_liver	0.021	0.864	0.948
psoriasis	0.010	0.939	0.989
glaucoma	0.008	0.947	0.989

Table B.5: *Correlation of the number of SNVs in D_2 vs relative performance, continued.*

d1	corre	pval	fdr_adj_p_values
autoimmune_thyroid_disease	0.007	0.957	0.989
systemic_lupus_erythematosus	0.006	0.964	0.989
ulcerative_colitis	0.002	0.984	0.989
multiple_sclerosis	-0.002	0.989	0.989
hypertension	-0.022	0.861	0.948
alcohol_dependence	-0.025	0.839	0.948
ankylosing_spondylitis	-0.026	0.832	0.948
systemic_scleroderma	-0.068	0.585	0.763
adult_onset_asthma	-0.081	0.517	0.718
schizophrenia	-0.092	0.461	0.673
squamous_cell_lung_carcinoma	-0.097	0.434	0.673
neurotic_disorder	-0.102	0.410	0.673
sclerosing_cholangitis	-0.106	0.394	0.670
crohn's_disease	-0.113	0.362	0.666
type_i_diabetes_mellitus	-0.135	0.277	0.598
atopic_asthma	-0.136	0.272	0.598
juvenile_idiopathic_arthritis	-0.147	0.235	0.570
hypothyroidism	-0.165	0.183	0.460
breast_carcinoma	-0.234	0.056	0.217
parkinson's_disease	-0.265	0.030	0.162
epilepsy	-0.270	0.027	0.162
peripheral_arterial_disease	-0.365	0.002	0.027

Table B.6: *Correlation of the number of SNVs in D_2 vs relative performance, continued.*

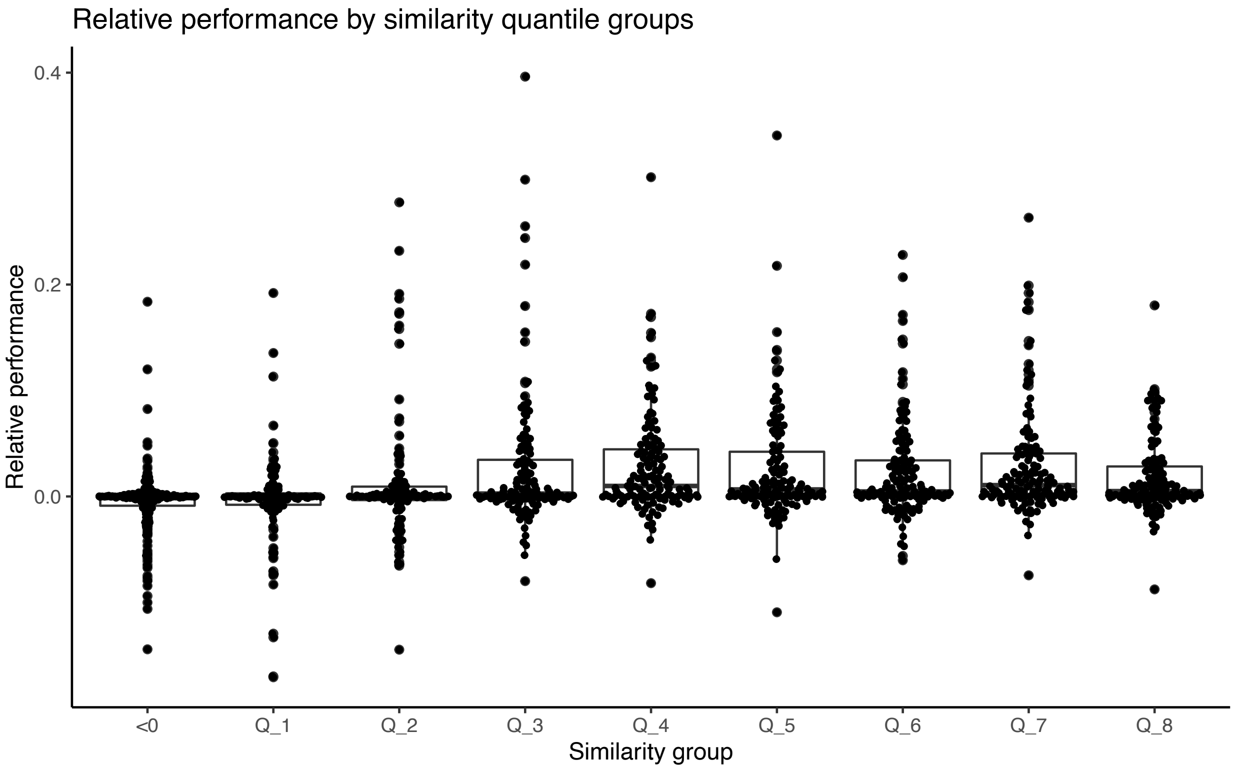

Figure B.13: *Relative performance by similarity quantile groups.* X axis is the similarity quantile groups of disease pairs. Y axis is the weight we assign for each disease pair through the information sharing model.

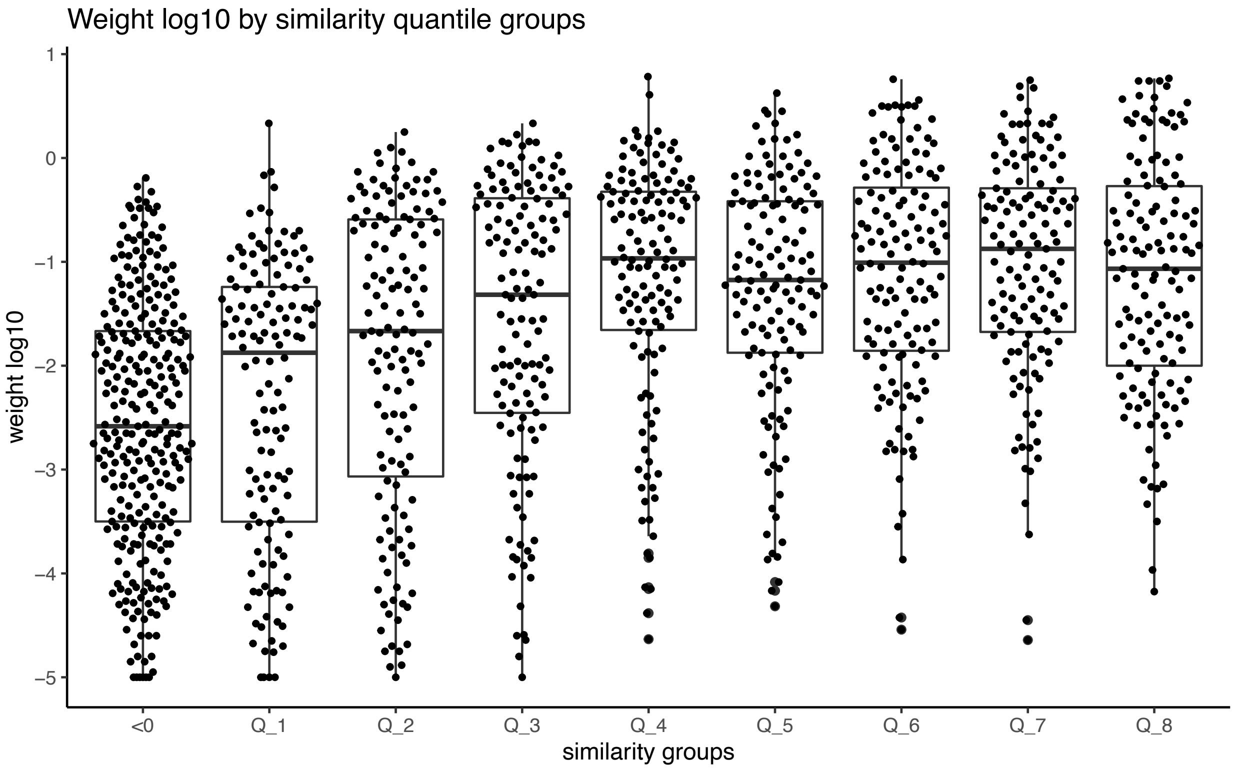

Figure B.14: *Log relative weight w by similarity quantile groups.* X axis is the similarity quantile groups of disease pairs. Y axis is the log relative weight for each disease pair through the information sharing model.

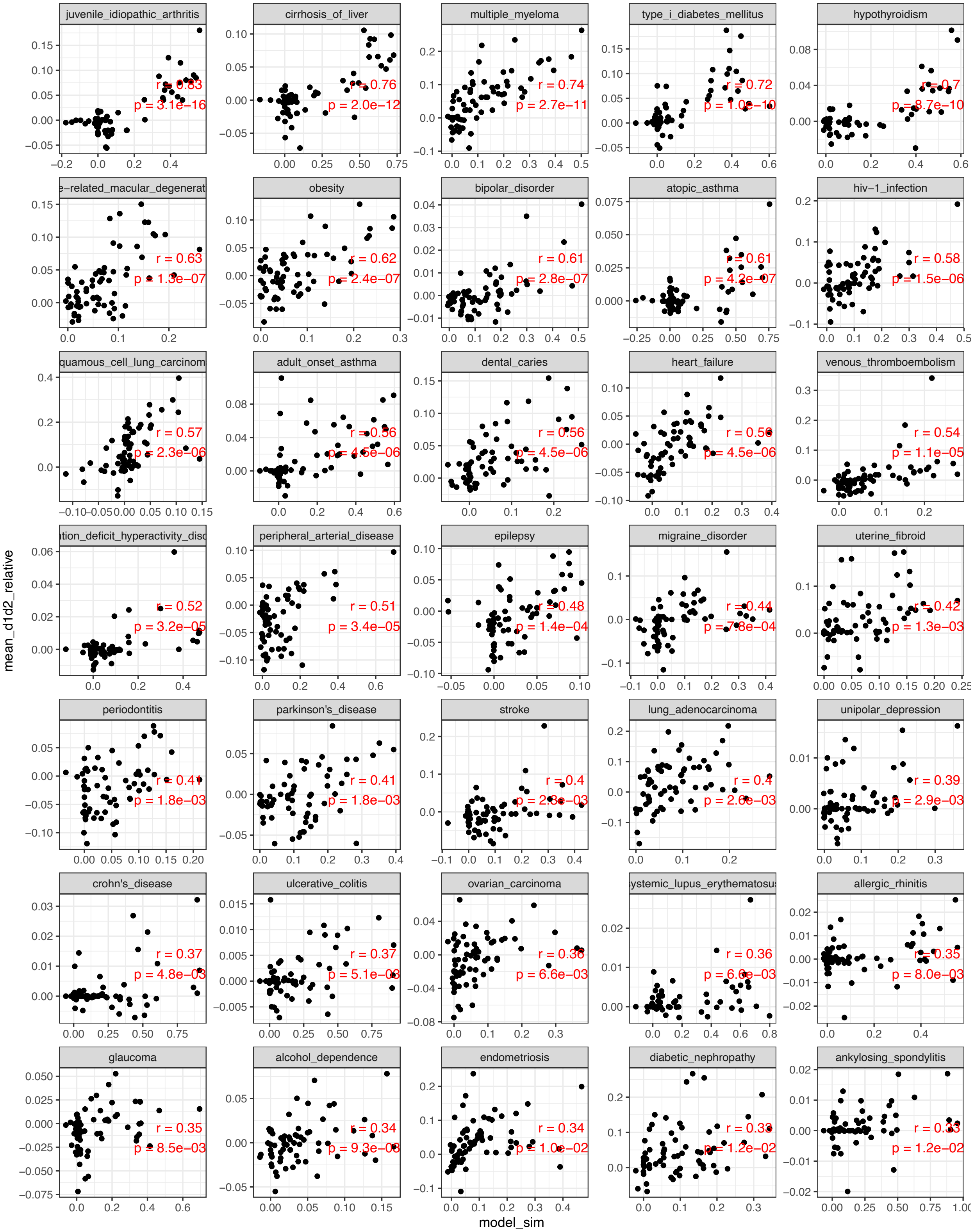

Figure B.15: *Model similarity and relative performance.*

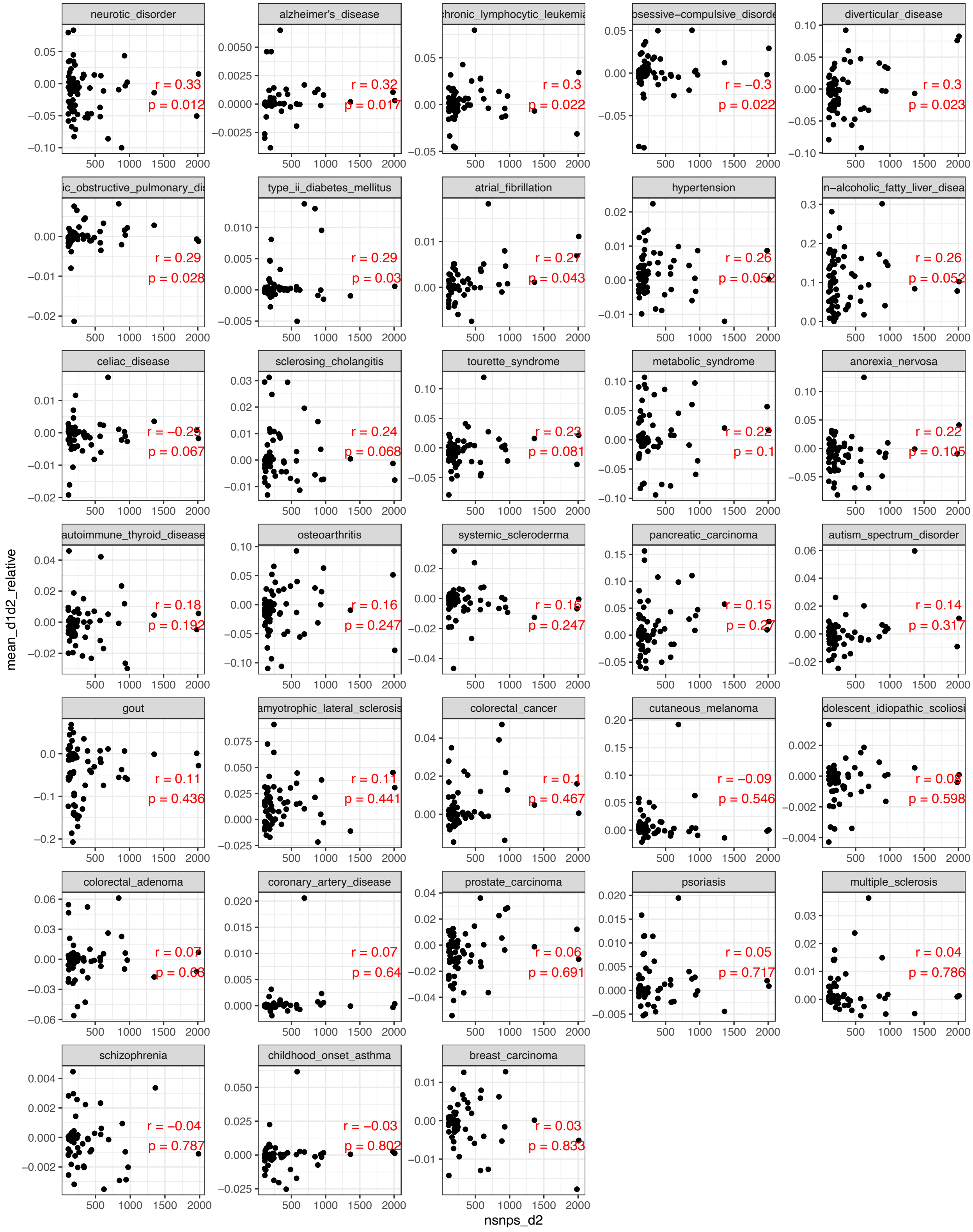

Figure B.16: *Model similarity and relative performance, continued.*

d1	corre	pval	fdr_adj_p_values
juvenile_idiopathic_arthritis	0.829	4.6e-18	3.13e-16
cirrhosis_of_liver	0.764	5.77e-14	1.96e-12
multiple_myeloma	0.737	1.18e-12	2.67e-11
type_i_diabetes_mellitus	0.721	6.04e-12	1.03e-10
hypothyroidism	0.696	6.38e-11	8.68e-10
age-related_macular_degeneration	0.629	1.16e-08	1.31e-07
obesity	0.618	2.46e-08	2.39e-07
bipolar_disorder	0.614	3.32e-08	2.82e-07
atopic_asthma	0.606	5.53e-08	4.18e-07
hiv-1_infection	0.583	2.21e-07	1.5e-06
squamous_cell_lung_carcinoma	0.575	3.69e-07	2.28e-06
adult_onset_asthma	0.560	8.28e-07	4.54e-06
dental_caries	0.558	9.34e-07	4.54e-06
heart_failure	0.559	8.68e-07	4.54e-06
venous_thromboembolism	0.540	2.4e-06	1.09e-05
attention_deficit_hyperactivity_disorder	0.517	7.61e-06	3.23e-05
peripheral_arterial_disease	0.514	8.51e-06	3.41e-05
epilepsy	0.481	3.71e-05	0.00014
migraine_disorder	0.437	0.000219	0.000783
uterine_fibroid	0.422	0.000381	0.0013
periodontitis	0.411	0.000545	0.00177
parkinson's_disease	0.409	0.000585	0.00181
stroke	0.401	0.000772	0.00228

Table B.7: *Correlation of model similarity vs relative performance.*

d1	corre	pval	fdr_adj_p_values
lung_adenocarcinoma	0.396	0.001	0.003
unipolar_depression	0.391	0.001	0.003
crohn's_disease	0.374	0.002	0.005
ulcerative_colitis	0.370	0.002	0.005
ovarian_carcinoma	0.360	0.003	0.007
systemic_lupus_erythematosus	0.360	0.003	0.007
allergic_rhinitis	0.352	0.004	0.008
glaucoma	0.348	0.004	0.008
alcohol_dependence	0.344	0.004	0.009
endometriosis	0.339	0.005	0.010
diabetic_nephropathy	0.334	0.006	0.012
ankylosing_spondylitis	0.331	0.006	0.012
neurotic_disorder	0.331	0.006	0.012
alzheimer's_disease	0.316	0.009	0.017
chronic_lymphocytic_leukemia	0.303	0.013	0.022
obsessive-compulsive_disorder	-0.303	0.013	0.022
diverticular_disease	0.300	0.014	0.023
chronic_obstructive_pulmonary_disease	0.291	0.017	0.028
type_ii_diabetes_mellitus	0.286	0.019	0.030
atrial_fibrillation	0.270	0.027	0.043
hypertension	0.260	0.034	0.052
non-alcoholic_fatty_liver_disease	0.259	0.034	0.052
celiac_disease	-0.246	0.045	0.067

Table B.8: *Correlation of model similarity vs relative performance, continued.*

d1	corre	pval	fdr_adj_p_values
sclerosing_cholangitis	0.243	0.047	0.068
tourette_syndrome	0.234	0.057	0.081
metabolic_syndrome	0.221	0.072	0.100
anorexia_nervosa	0.217	0.077	0.105
autoimmune_thyroid_disease	0.180	0.144	0.192
osteoarthritis	0.162	0.189	0.247
systemic_scleroderma	0.161	0.192	0.247
pancreatic_carcinoma	0.154	0.214	0.270
autism_spectrum_disorder	0.140	0.257	0.317
gout	0.114	0.359	0.436
amyotrophic_lateral_sclerosis	0.111	0.370	0.441
colorectal_cancer	0.105	0.398	0.467
cutaneous_melanoma	-0.089	0.474	0.546
adolescent_idiopathic_scoliosis	0.079	0.527	0.598
colorectal_adenoma	0.072	0.565	0.630
coronary_artery_disease	0.068	0.583	0.640
prostate_carcinoma	0.058	0.640	0.691
psoriasis	0.052	0.675	0.717
multiple_sclerosis	0.039	0.751	0.786
schizophrenia	-0.037	0.764	0.787
childhood_onset_asthma	-0.033	0.790	0.802
breast_carcinoma	0.026	0.833	0.833

Table B.9: *Correlation of model similarity vs relative performance, continued.*

D_1	Top informative D_2	impr	weight*
squamous_cell_lung_carcinoma	hypothyroidism	40%	0.14
venous_thromboembolism	alzheimer's_disease	34%	0.62
non-alcoholic_fatty_liver_disease	adolescent_idiopathic_scoliosis	30%	0.21
diabetic_nephropathy	coronary_artery_disease	27%	0.44
multiple_myeloma	multiple_sclerosis	26%	0.69
endometriosis	amyotrophic_lateral_sclerosis	24%	0.45
stroke	coronary_artery_disease	23%	0.76
lung_adenocarcinoma	breast_carcinoma	22%	0.45
cutaneous_melanoma	chronic_obstructive_pulmonary_disease	19%	-0.13
hiv-1_infection	bipolar_disorder	19%	0.45
type_i_diabetes_mellitus	chronic_lymphocytic_leukemia	19%	0.10
juvenile_idiopathic_arthritis	celiac_disease	18%	0.43
uterine_fibroid	crohn's_disease	17%	-1.05
pancreatic_carcinoma	alcohol_dependence	16%	-0.45
migraine_disorder	atrial_fibrillation	16%	0.18
dental_caries	unipolar_depression	15%	0.26
age-related_macular_degeneration	systemic_lupus_erythematosus	15%	-0.22
obesity	psoriasis	13%	-0.38
anorexia_nervosa	attention_deficit_hyperactivity_disorder	13%	0.39
tourette_syndrome	attention_deficit_hyperactivity_disorder	12%	0.17
heart_failure	cutaneous_melanoma	12%	0.31
adult_onset_asthma	chronic_obstructive_pulmonary_disease	11%	-0.03
metabolic_syndrome	amyotrophic_lateral_sclerosis	11%	-1.32
cirrhosis_of_liver	psoriasis	11%	0.82

* All weight values are presented in logarithmic scale (log10).

Table B.10: *Top informative diseases D_2 enhancing model performance of diseases D_1 with less than 300 SNVs.*

D_1	D_2 Best Improvement	impr	weight*
hypothyroidism	celiac_disease	10%	0.58
peripheral_arterial_disease	glaucoma	10%	0.02
epilepsy	hypothyroidism	9%	-0.32
osteoarthritis	colorectal_cancer	9%	0.33
diverticular_disease	atrial_fibrillation	9%	0.05
amyotrophic_lateral_sclerosis	parkinson's_disease	9%	0.08
periodontitis	coronary_artery_disease	9%	-0.30
parkinson's_disease	sclerosing_cholangitis	8%	-0.28
neurotic_disorder	anorexia_nervosa	8%	0.08
chronic_lymphocytic_leukemia	systemic_lupus_erythematosus	8%	-0.53
alcohol_dependence	metabolic_syndrome	8%	-0.53
atopic_asthma	celiac_disease	7%	0.53
gout	epilepsy	7%	-1.84
ovarian_carcinoma	uterine_fibroid	7%	-0.11
colorectal_adenoma	coronary_artery_disease	6%	-0.30
glaucoma	atrial_fibrillation	5%	0.34
obsessive-compulsive_disorder	adolescent_idiopathic_scoliosis	5%	-1.58
autoimmune_thyroid_disease	adult_onset_asthma	5%	-1.84
systemic_scleroderma	celiac_disease	3%	-0.17
sclerosing_cholangitis	hypothyroidism	3%	-1.41
allergic_rhinitis	hypothyroidism	3%	-0.67
celiac_disease	chronic_obstructive_pulmonary_disease	2%	-1.31

* All weight values are presented in logarithmic scale (log10).

Table B.11: *Top informative diseases D_2 enhancing model performance of diseases D_1 with less than 300 SNVs, continued.*

Appendix C Supplemental materials for aim 3

Clinvar Database		
Overall number of variants:	1,572,399	
Variants overlaping AENMD default tx set:	1,035,485	
	number	percentage
Unique PTC variants	107,171	100
NMD triggering	66,178	61.7
NMD escaping canonical	16,474	15.4
NMD escaping non-canonical	22,085	20.6
Transcript dependent predictions	2,434	2.3
NMD escaping overall	40,993	38.3
Unique PTC benign variants	965	100
NMD triggering	340	35.2
NMD escaping canonical	446	46.2
NMD escaping non-canonical	168	17.4
Transcript dependent predictions	11	1.1
NMD escaping overall	625	64.8
Unique PTC uncertain variants	12,311	100
NMD triggering	5,924	48.1
NMD escaping canonical	4,399	35.7
NMD escaping non-canonical	1,818	14.8
Transcript dependent predictions	170	1.4
NMD escaping overall	6,387	51.9

Table C.1: *Clinvar database.*

Clinvar Database		
	number	percentage
Unique PTC pathogenic variants	91,984	100
NMD triggering	58,649	63.8
NMD escaping canonical	11,380	12.4
NMD escaping non-canonical	19,754	21.5
Transcript dependent predictions	2,201	2.4
NMD escaping overall	33,335	36.2
Unique PTC other variants	1,911	100
NMD triggering	1,265	66.2
NMD escaping canonical	249	13.0
NMD escaping non-canonical	345	18.1
Transcript dependent predictions	52	2.7
NMD escaping overall	646	33.8
Unique PTC variant-transcript pairs	203,462	100
NMD triggering	132,585	65.2
NMD escaping canonical	31,755	15.6
NMD escaping non-canonical	39,122	19.2
NMD escaping overall	70,877	34.8

Table C.2: *Clinvar database, continued.*

gnomAD Database		
Overall number of variants:	14,951,900	
Variants overlaping AENMD default tx set:	7,375,683	
	number	percentage
Unique PTC variants	300,034	100.0
NMD triggering	151,440	50.5
NMD escaping canonical	92,524	30.8
NMD escaping non-canonical	50,418	16.8
Variants with transcript dependent predictions	5,652	1.9
NMD escaping overall	148,594	49.5
PTC variant-transcript pairs	477,142	100
NMD triggering	258,834	54.2
NMD escaping canonical	136,991	28.7
NMD escaping non-canonical	81,317	17.0
NMD escaping overall	218,308	45.8

Table C.3: *GnomAD database.*

aenmd and VEP Annotation of Clinvar - Comparison		
Total number of annotations by AENMD:	203,462	
Total number of annotations by VEP:	77,001	
	number	percentage
Overlapping PTC variant-transcript pairs annotated:	75,839	100
AENMD annotations:		
NMD triggering	62,060	81.8
NMD escaping	13,779	18.2
VEP annotations:		
NMD triggering	62,384	82.3
NMD escaping	13,455	17.7
Called NMD triggering by both	61,288	80.8
Called NMD escaping by both	12,683	16.7
Called NMD triggering by AENMD, not VEP	1,096	1.4
Called NMD triggering by VEP, not AENMD	773	1.0

Table C.4: *VEP comparison.*

Bibliography

[1] Shannon Stefl, Hafumi Nishi, Marharyta Petukh, Anna R. Panchenko, and Emil Alexov. Molecular mechanisms of disease-causing missense mutations. *Journal of Molecular Biology*, 425(21):3919–3936, 2013.

[2] Zhe Zhang, Maria A. Miteva, Lin Wang, and Emil Alexov. Analyzing effects of naturally occurring missense mutations. *Computational and Mathematical Methods in Medicine*, 2012:805827, 2012.

[3] J. D. French and S. L. Edwards. The role of noncoding variants in heritable disease. *Trends in Genetics*, 36(11):880–891, 2020.

[4] Lucas D. Ward and Manolis Kellis. Interpreting noncoding genetic variation in complex traits and human disease. *Nature biotechnology*, 30(11):1095–1106, 2012.

[5] Phil H. Lee, Christian Lee, Xihao Li, Brian Wee, Tushar Dwivedi, and Mark Daly. Principles and methods of in-silico prioritization of non-coding regulatory variants. *Human genetics*, 137(1):15–30, 2018.

[6] Martin Kircher, Daniela M. Witten, Preti Jain, Brian J. O'Roak, Gregory M. Cooper, and Jay Shendure. A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, 46(3):310–315, 2014.

[7] Qiongshi Lu, Ryan Lee Powles, Qian Wang, Beixin Julie He, and Hongyu Zhao. Integrative tissue-specific functional annotations in the human genome provide novel insights on many complex traits and improve signal prioritization in genome wide association studies. *PLoS genetics*, 12(4):e1005947, 2016.

[8] Fran Supek, Ben Lehner, and Rik G. H. Lindeboom. To nmd or not to nmd: Nonsense-mediated mrna decay in cancer and other genetic diseases. *Trends in Genetics*, 37(7):657–668, 2021.

[9] Jake N. Miller and David A. Pearce. Nonsense-mediated decay in genetic disease: Friend or foe? *Mutation Research/Reviews in Mutation Research*, 762:52–64, 2014.

[10] Mehrdad Khajavi, Ken Inoue, and James R. Lupski. Nonsense-mediated mrna decay modulates clinical outcome of genetic disease. *European Journal of Human Genetics*, 14(10):1074–1081, 2006.

[11] Sarah E Hunt, Benjamin Moore, Ridwan M Amode, Irina M Armean, Diana Lemos, Aleena Mushtaq, Andrew Parton, Helen Schuilenburg, Michał Szpak, and Anja Thormann. Annotating and prioritizing genomic variants using the ensembl variant effect predictor—a tutorial. *Human mutation*, 43(8):986–997, 2022.

[12] Rik G. H. Lindeboom, Michiel Vermeulen, Ben Lehner, and Fran Supek. The impact of nonsense-mediated mrna decay on genetic disease, gene editing and cancer immunotherapy. *Nature Genetics*, 51(11):1645–1651, 2019.

[13] J. D. French and S. L. Edwards. The role of noncoding variants in heritable disease. *Trends in Genetics*, 36(11):880–891, 2020.

[14] M. Schipper and D. Posthuma. Demystifying non-coding gwas variants: an overview of computational tools and methods. *Hum Mol Genet*, 31(R1):R73–r83, 2022.

[15] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller, V. Amin, J. W. Whitaker, M. D. Schultz, L. D. Ward, A. Sarkar, G. Quon, R. S. Sandstrom, M. L. Eaton, Y. C. Wu, A. R. Pfenning, X. Wang, M. Claussnitzer, Y. Liu, C. Coarfa, R. A. Harris, N. Shoresh, C. B. Epstein, E. Gjoneska, D. Leung, W. Xie, R. D. Hawkins, R. Lister, C. Hong, P. Gascard, A. J. Mungall, R. Moore, E. Chuah, A. Tam, T. K. Canfield, R. S. Hansen, R. Kaul, P. J. Sabo, M. S. Bansal, A. Carles, J. R. Dixon, K. H. Farh, S. Feizi, R. Karlic, A. R. Kim, A. Kulkarni, D. Li, R. Lowdon, G. Elliott, T. R. Mercer, S. J. Neph, V. Onuchic, P. Polak, N. Rajagopal, P. Ray, R. C. Sallari, K. T. Siebenthall, N. A. Sinnott-Armstrong, M. Stevens, R. E. Thurman, J. Wu, B. Zhang, X. Zhou, A. E. Beaudet, L. A. Boyer, P. L. De Jager, P. J. Farnham, S. J. Fisher, D. Haussler, S. J. Jones, W. Li, M. A. Marra, M. T. McManus, S. Sunyaev, J. A. Thomson, T. D. Tlsty, L. H. Tsai, W. Wang, R. A. Waterland, M. Q. Zhang, L. H. Chadwick, B. E. Bernstein, J. F. Costello, J. R. Ecker, M. Hirst, A. Meissner, A. Milosavljevic, B. Ren, J. A. Stamatoyannopoulos, T. Wang, and M. Kellis. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539):317–30, 2015.

[16] Jill E. Moore, Michael J. Purcaro, Henry E. Pratt, Charles B. Epstein, Noam Shoresh, Jessika Adrian, Trupti Kawli, Carrie A. Davis, Alexander Dobin, Rajinder Kaul, Jessica Halow, Eric L. Van Nostrand, Peter Freese, David U. Gorkin, Yin Shen, Yupeng He, Mark Mackiewicz, Florencia Pauli-Behn, Brian A. Williams, Ali Mortazavi, Cheryl A. Keller, Xiao-Ou Zhang, Shaimae I. Elhajjajy, Jack Huey, Diane E. Dickel, Valentina Snetkova, Xintao Wei, Xiaofeng Wang, Juan Carlos Rivera-Mulia, Joel

Rozowsky, Jing Zhang, Surya B. Chhetri, Jialing Zhang, Alec Victorsen, Kevin P. White, Axel Visel, Gene W. Yeo, Christopher B. Burge, Eric Lécuyer, David M. Gilbert, Job Dekker, John Rinn, Eric M. Mendenhall, Joseph R. Ecker, Manolis Kellis, Robert J. Klein, William S. Noble, Anshul Kundaje, Roderic Guigó, Peggy J. Farnham, J. Michael Cherry, Richard M. Myers, Bing Ren, Brenton R. Graveley, Mark B. Gerstein, Len A. Pennacchio, Michael P. Snyder, Bradley E. Bernstein, Barbara Wold, Ross C. Hardison, Thomas R. Gingeras, John A. Stamatoyannopoulos, Zhiping Weng, and Encode Project Consortium The. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583(7818):699–710, 2020.

[17] Pavel P. Kuksa, Emily Greenfest-Allen, Jeffrey Cifello, Matei Ionita, Hui Wang, Heather Nicaretta, Po-Liang Cheng, Wan-Ping Lee, Li-San Wang, and Yuk Yee Leung. Scalable approaches for functional analyses of whole-genome sequencing non-coding variants. *Human Molecular Genetics*, 31(R1):R62–R72, 2022.

[18] C. Yuen RK, D. Merico, M. Bookman, L. Howe J, B. Thiruvahindrapuram, R. V. Patel, J. Whitney, N. Deflaux, J. Bingham, Z. Wang, G. Pellecchia, J. A. Buchanan, S. Walker, C. R. Marshall, M. Uddin, M. Zarrei, E. Deneault, L. D'Abate, A. J. Chan, S. Koyanagi, T. Paton, S. L. Pereira, N. Hoang, W. Engchuan, E. J. Higginbotham, K. Ho, S. Lamoureux, W. Li, J. R. MacDonald, T. Nalpathamkalam, W. W. Sung, F. J. Tsoi, J. Wei, L. Xu, A. M. Tasse, E. Kirby, W. Van Etten, S. Twigger, W. Roberts, I. Drmic, S. Jilderda, B. M. Modi, B. Kellam, M. Szego, C. Cytrynbaum, R. Weksberg, L. Zwaigenbaum, M. Woodbury-Smith, J. Brian, L. Senman, A. Iaboni, K. Doyle-Thomas, A. Thompson, C. Chrysler, J. Leef, T. Savion-Lemieux, I. M. Smith, X. Liu, R. Nicolson, V. Seifer, A. Fedele, E. H. Cook, S. Dager, A. Estes, L. Gallagher, B. A. Malow, J. R. Parr, S. J. Spence, J. Vorstman, B. J. Frey, J. T. Robinson, L. J. Strug, B. A. Fernandez, M. Elsabbagh, M. T. Carter, J. Hallmayer, B. M. Knoppers, E. Anagnostou, P. Szatmari, R. H. Ring, D. Glazer, M. T. Pletcher, and S. W. Scherer. Whole genome sequencing resource identifies 18 new candidate genes for autism spectrum disorder. *Nat Neurosci*, 20(4):602–611, 2017.

[19] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[20] Eugene V. Davydov, David L. Goode, Marina Sirota, Gregory M. Cooper, Arend Sidow, and Serafim Batzoglou. Identifying a high fraction of the human genome to be under selective constraint using gerp++. *PLOS Computational Biology*, 6(12):e1001025, 2010.

[21] Adam Siepel, Gill Bejerano, Jakob S. Pedersen, Angie S. Hinrichs, Minmei Hou, Kate Rosenbloom, Hiram Clawson, John Spieth, LaDeana W. Hillier, Stephen Richards, George M. Weinstock, Richard K. Wilson, Richard A. Gibbs, W. James Kent, Webb Miller, and David Haussler. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Research*, 15(8):1034–1050, 2005.

[22] Manuel Garber, Mitchell Guttman, Michele Clamp, Michael C. Zody, Nir Friedman, and Xiaohui Xie. Identifying novel constrained elements by exploiting biased substitution patterns. *Bioinformatics*, 25(12):i54–i62, 2009.

[23] Hong Sun and Guangjun Yu. New insights into the pathogenicity of non-synonymous variants through multi-level analysis. *Scientific Reports*, 9(1):1667, 2019.

[24] Iuliana Ionita-Laza, Kenneth McCallum, Bin Xu, and Joseph D. Buxbaum. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nature Genetics*, 48(2):214–220, 2016.

[25] Ian Dunham, Anshul Kundaje, Shelley F. Aldred, Patrick J. Collins, Carrie A. Davis, Francis Doyle, Charles B. Epstein, Seth Frietze, Jennifer Harrow, Rajinder Kaul, Jainab Khatun, Bryan R. Lajoie, Stephen G. Landt, Bum-Kyu Lee, Florencia Pauli, Kate R. Rosenbloom, Peter Sabo, Alexias Safi, Amartya Sanyal, Noam Shoresh, Jeremy M. Simon, Lingyun Song, Nathan D. Trinklein, Robert C. Altshuler, Ewan Birney, James B. Brown, Chao Cheng, Sarah Djebali, Xianjun Dong, Ian Dunham, Jason Ernst, Terrence S. Furey, Mark Gerstein, Belinda Giardine, Melissa Greven, Ross C. Hardison, Robert S. Harris, Javier Herrero, Michael M. Hoffman, Sowmya Iyer, Manolis Kellis, Jainab Khatun, Pouya Kheradpour, Anshul Kundaje, Timo Lassmann, Qunhua Li, Xinying Lin, Georgi K. Marinov, Angelika Merkel, Ali Mortazavi, Stephen C. J. Parker, Timothy E. Reddy, Joel Rozowsky, Felix Schlesinger, Robert E. Thurman, Jie Wang, Lucas D. Ward, Troy W. Whitfield, Steven P. Wilder, Weisheng Wu, Hualin S. Xi, Kevin Y. Yip, Jiali Zhuang, Bradley E. Bernstein, Ewan Birney, Ian Dunham, Eric D. Green, Chris Gunter, Michael Snyder, Michael J. Pazin, Rebecca F. Lowdon, Laura A. L. Dillon, Leslie B. Adams, Caroline J. Kelly, Julia Zhang, Judith R. Wexler, Eric D. Green, Peter J. Good, Elise A. Feingold, Bradley E. Bernstein, Ewan Birney, Gregory E. Crawford, Job Dekker, Laura Elnitski, Peggy J. Farnham, Mark Gerstein, Morgan C. Giddings, Thomas R. Gingeras, Eric D. Green, Roderic Guigó, Ross C. Hardison, Timothy J. Hubbard, Manolis Kellis, W. James Kent, Jason D. Lieb, Elliott H. Margulies, Richard M. Myers, Michael Snyder, John A. Stamatoyannopoulos, Scott A. Tenenbaum, et al. An integrated encyclopedia of dna elements in the human genome. *Nature*, 489(7414):57–74, 2012.

[26] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki,

Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen Bagger, Mette Jørgensen, Peter Refsing Andersen, Nicolas Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J. Mungall, Terrence F. Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O. Daub, Peter Heutink, David A. Hume, Torben Heick Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, Alistair R. R. Forrest, Piero Carninci, Michael Rehli, Albin Sandelin, and Fantom Consortium The. An atlas of active enhancers across human cell types and tissues. *Nature*, 507(7493):455–461, 2014.

[27] GTEx Consortium The, François Aguet, Shankara Anand, Kristin G. Ardlie, Stacey Gabriel, Gad A. Getz, Aaron Graubert, Kane Hadley, Robert E. Handsaker, Katherine H. Huang, Seva Kashin, Xiao Li, Daniel G. MacArthur, Samuel R. Meier, Jared L. Nedzel, Duyen T. Nguyen, Ayellet V. Segrè, Ellen Todres, Brunilda Balliu, Alvaro N. Barbeira, Alexis Battle, Rodrigo Bonazzola, Andrew Brown, Christopher D. Brown, Stephane E. Castel, Donald F. Conrad, Daniel J. Cotter, Nancy Cox, Sayantan Das, Olivia M. de Goede, Emmanouil T. Dermitzakis, Jonah Einson, Barbara E. Engelhardt, Eleazar Eskin, Tiffany Y. Eulalio, Nicole M. Ferraro, Elise D. Flynn, Laure Fresard, Eric R. Gamazon, Diego Garrido-Martín, Nicole R. Gay, Michael J. Gloudemans, Roderic Guigó, Andrew R. Hame, Yuan He, Paul J. Hoffman, Farhad Hormozdiari, Lei Hou, Hae Kyung Im, Brian Jo, Silva Kasela, Manolis Kellis, Sarah Kim-Hellmuth, Alan Kwong, Tuuli Lappalainen, Xin Li, Yanyu Liang, Serghei Mangul, Pejman Mohammadi, Stephen B. Montgomery, Manuel Muñoz-Aguirre, Daniel C. Nachun, Andrew B. Nobel, Meritxell Oliva, YoSon Park, Yongjin Park, Princy Parsana, Abhiram S. Rao, Ferran Reverter, John M. Rouhana, Chiara Sabatti, Ashis Saha, Matthew Stephens, Barbara E. Stranger, Benjamin J. Strober, Nicole A. Teran, Ana Viñuela, Gao Wang, Xiaoquan Wen, Fred Wright, Valentin Wucher, Yuxin Zou, Pedro G. Ferreira, Gen Li, Marta Melé, Esti Yeger-Lotem, Mary E. Barcus, Debra Bradbury, Tanya Krubit, Jeffrey A. McLean, Liqun Qi, Karna Robinson, Nancy V. Roche, Anna M. Smith, Leslie Sobin, David E. Tabor, Anita Undale, Jason Bridge, Lori E. Brigham, Barbara A. Foster, et al. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[28] Ekta Khurana, Yao Fu, Dimple Chakravarty, Francesca Demichelis, Mark A. Rubin, and Mark Gerstein. Role of non-coding sequence variants in cancer. *Nature Reviews Genetics*, 17(2):93–108, 2016.

[29] Fyodor D Urnov. Chromatin remodeling as a guide to transcriptional regulatory networks in mammals. *Journal of cellular biochemistry*, 88(4):684–694, 2003.

[30] Wen-Hua Qi, Chao-chao Yan, Wu-Jiao Li, Xue-Mei Jiang, Guang-Zhou Li, Xiu-Yue Zhang, Ting-Zhang Hu, Jing Li, and Bi-Song Yue. Distinct patterns of simple sequence

repeats and gc distribution in intragenic and intergenic regions of primate genomes. *Aging (Albany NY)*, 8(11):2635, 2016.

[31] Stephen CJ Parker, Elliott H Margulies, and Thomas D Tullius. The relationship between fine scale dna structure, gc content, and functional elements in 1 *Genome Informatics*, 20:199–211, 2008.

[32] Carina F. Mugal, Peter F. Arndt, Lena Holm, and Hans Ellegren. Evolutionary consequences of dna methylation on the gc content in vertebrate genomes. *G3 Genes—Genomes—Genetics*, 5(3):441–447, 2015.

[33] Graham R. S. Ritchie, Ian Dunham, Eleftheria Zeggini, and Paul Flicek. Functional annotation of noncoding sequence variants. *Nature Methods*, 11(3):294–296, 2014.

[34] Qiongshi Lu, Yiming Hu, Jiehuan Sun, Yuwei Cheng, Kei-Hoi Cheung, and Hongyu Zhao. A statistical framework to predict functional non-coding regions in the human genome through integrated analysis of annotation data. *Scientific Reports*, 5(1):10576, 2015.

[35] Yi-Fei Huang, Brad Gulko, and Adam Siepel. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nature Genetics*, 49(4):618–624, 2017.

[36] Qiongshi Lu, Ryan L Powles, Sarah Abdallah, Derek Ou, Qian Wang, Yiming Hu, Yisi Lu, Wei Liu, Boyang Li, and Shubhabrata Mukherjee. Systematic tissue-specific functional annotation of the human genome highlights immune-related dna elements for late-onset alzheimer's disease. *PLoS genetics*, 13(7):e1006933, 2017.

[37] Brad Gulko and Adam Siepel. An evolutionary framework for measuring epigenomic information and estimating cell-type-specific fitness consequences. *Nature Genetics*, 51(2):335–342, 2019.

[38] Felix Richter, Sarah U. Morton, Seong Won Kim, Alexander Kitaygorodsky, Lauren K. Wasson, Kathleen M. Chen, Jian Zhou, Hongjian Qi, Nihir Patel, Steven R. DePalma, Michael Parfenov, Jason Homsy, Joshua M. Gorham, Kathryn B. Manheimer, Matthew Velinder, Andrew Farrell, Gabor Marth, Eric E. Schadt, Jonathan R. Kaltman, Jane W. Newburger, Alessandro Giardini, Elizabeth Goldmuntz, Martina Brueckner, Richard Kim, George A. Porter, Daniel Bernstein, Wendy K. Chung, Deepak Srivastava, Martin Tristani-Firouzi, Olga G. Troyanskaya, Diane E. Dickel,

Yufeng Shen, Jonathan G. Seidman, Christine E. Seidman, and Bruce D. Gelb. Genomic analyses implicate noncoding de novo variants in congenital heart disease. *Nature Genetics*, 52(8):769–777, 2020.

[39] Ali Yousefian-Jazi, Min Kyung Sung, Taeyeop Lee, Yoon-Ho Hong, Jung Kyoon Choi, and Jinwook Choi. Functional fine-mapping of noncoding risk variants in amyotrophic lateral sclerosis utilizing convolutional neural network. *Scientific Reports*, 10(1):12872, 2020.

[40] Ali Yousefian-Jazi, Jieun Jung, Jung Kyoon Choi, and Jinwook Choi. Functional annotation of noncoding causal variants in autoimmune diseases. *Genomics*, 112(2):1208–1213, 2020.

[41] Long Gao, Yasin Uzun, Peng Gao, Bing He, Xiaoke Ma, Jiahui Wang, Shizhong Han, and Kai Tan. Identifying noncoding risk variants using disease-relevant gene regulatory networks. *Nature Communications*, 9(1):702, 2018.

[42] Corneliu A. Bodea, Adele A. Mitchell, Alex Bloemendal, Aaron G. Day-Williams, Heiko Runz, and Shamil R. Sunyaev. Pines: phenotype-informed tissue weighting improves prediction of pathogenic noncoding variants. *Genome Biology*, 19(1):173, 2018.

[43] Li Chen, Peng Jin, and Zhaohui S. Qin. Divan: accurate identification of non-coding disease-specific risk variants using multi-omics profiles. *Genome Biology*, 17(1):252, 2016.

[44] Chris Cotsapas, Benjamin F. Voight, Elizabeth Rossin, Kasper Lage, Benjamin M. Neale, Chris Wallace, Gonçalo R. Abecasis, Jeffrey C. Barrett, Timothy Behrens, Judy Cho, Philip L. De Jager, James T. Elder, Robert R. Graham, Peter Gregersen, Lars Klareskog, Katherine A. Siminovitch, David A. van Heel, Cisca Wijmenga, Jane Worthington, John A. Todd, David A. Hafler, Stephen S. Rich, Mark J. Daly, and FOCiS Network of Consortia on behalf of the. Pervasive sharing of genetic effects in autoimmune disease. *PLOS Genetics*, 7(8):e1002254, 2011.

[45] Yun R. Li, Jin Li, Sihai D. Zhao, Jonathan P. Bradfield, Frank D. Mentch, S. Melkorka Maggadottir, Cuiping Hou, Debra J. Abrams, Diana Chang, Feng Gao, Yiran Guo, Zhi Wei, John J. Connolly, Christopher J. Cardinale, Marina Bakay, Joseph T. Glessner, Dong Li, Charlly Kao, Kelly A. Thomas, Haijun Qiu, Rosetta M. Chiavacci, Cecilia E. Kim, Fengxiang Wang, James Snyder, Marylyn D. Richie, Berit Flatø, Øystein Førre, Lee A. Denson, Susan D. Thompson, Mara L. Becker, Stephen L. Guthery, Anna Latiano, Elena Perez, Elena Resnick, Richard K. Russell, David C. Wilson,

Mark S. Silverberg, Vito Annese, Benedicte A. Lie, Marilynn Punaro, Marla C. Dubinsky, Dimitri S. Monos, Caterina Strisciuglio, Annamaria Staiano, Erasmo Miele, Subra Kugathasan, Justine A. Ellis, Jane E. Munro, Kathleen E. Sullivan, Carol A. Wise, Helen Chapel, Charlotte Cunningham-Rundles, Struan F. A. Grant, Jordan S. Orange, Patrick M. A. Sleiman, Edward M. Behrens, Anne M. Griffiths, Jack Satsangi, Terri H. Finkel, Alon Keinan, Eline T. Luning Prak, Constantin Polychronakos, Robert N. Baldassano, Hongzhe Li, Brendan J. Keating, and Hakon Hakonarson. Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases. *Nature Medicine*, 21(9):1018–1027, 2015.

[46] Thomas S. Wingo, Yue Liu, Ekaterina S. Gerasimov, Selina M. Vattathil, Meghan E. Wynne, Jiaqi Liu, Adriana Lori, Victor Faundez, David A. Bennett, Nicholas T. Seyfried, Allan I. Levey, and Aliza P. Wingo. Shared mechanisms across the major psychiatric and neurodegenerative diseases. *Nature Communications*, 13(1):4314, 2022.

[47] Cato Romero, Josefin Werme, Philip R. Jansen, Joel Gelernter, Murray B. Stein, Daniel Levey, Renato Polimanti, Christiaan de Leeuw, Danielle Posthuma, Mats Nagel, and Sophie van der Sluis. Exploring the genetic overlap between twelve psychiatric disorders. *Nature Genetics*, 54(12):1795–1802, 2022.

[48] Zixin Hu, Rong Jiao, Panpan Wang, Yun Zhu, Jinying Zhao, Phil De Jager, David A. Bennett, Li Jin, and Momiao Xiong. Shared causal paths underlying alzheimer's dementia and type 2 diabetes. *Scientific Reports*, 10(1):4107, 2020.

[49] Khyati Mittal and Deepshikha Pande Katare. Shared links between type 2 diabetes mellitus and alzheimer's disease: A review. *Diabetes Metabolic Syndrome: Clinical Research Reviews*, 10(2, Supplement 1):S144–S149, 2016.

[50] Manojkumar Kumaran and Bharanidharan Devarajan. eyevarp: a computational framework for the identification of pathogenic variants specific to eye disease. *Genetics in Medicine*, page 100862, 2023.

[51] X. F. Chen, M. R. Guo, Y. Y. Duan, F. Jiang, H. Wu, S. S. Dong, X. R. Zhou, H. N. Thynn, C. C. Liu, L. Zhang, Y. Guo, and T. L. Yang. Multiomics dissection of molecular regulatory mechanisms underlying autoimmune-associated noncoding snps. *JCI Insight*, 5(17), 2020.

[52] Z. Cao, Y. Huang, R. Duan, P. Jin, Z. S. Qin, and S. Zhang. Disease category-specific annotation of variants using an ensemble learning framework. *Brief Bioinform*, 23(1), 2022.

[53] Saverio Brogna and Jikai Wen. Nonsense-mediated mrna decay (nmd) mechanisms. *Nature Structural Molecular Biology*, 16(2):107–113, 2009.

[54] Michael C. Dyle, Divya Kolakada, Michael A. Cortazar, and Sujatha Jagannathan. How to get away with nonsense: Mechanisms and consequences of escape from nonsense-mediated rna decay. *WIREs RNA*, 11(1):e1560, 2020.

[55] Lionel Ballut, Brice Marchadier, Aurélie Baguet, Catherine Tomasetto, Bertrand Séraphin, and Hervé Le Hir. The exon junction core complex is locked onto rna by inhibition of eif4aiii atpase activity. *Nature Structural Molecular Biology*, 12(10):861–869, 2005.

[56] Christoph Schweingruber, Simone C. Rufener, David Zünd, Akio Yamashita, and Oliver Mühlemann. Nonsense-mediated mrna decay — mechanisms of substrate mrna recognition and degradation in mammalian cells. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1829(6):612–623, 2013.

[57] Manuel A. Rivas, Matti Pirinen, Donald F. Conrad, Monkol Lek, Emily K. Tsang, Konrad J. Karczewski, Julian B. Maller, Kimberly R. Kukurba, David S. DeLuca, Menachem Fromer, Pedro G. Ferreira, Kevin S. Smith, Rui Zhang, Fengmei Zhao, Eric Banks, Ryan Poplin, Douglas M. Ruderfer, Shaun M. Purcell, Taru Tukiainen, Eric V. Minikel, Peter D. Stenson, David N. Cooper, Katharine H. Huang, Timothy J. Sullivan, Jared Nedzel, GTEx Consortium The, Consortium The Geuvadis, Carlos D. Bustamante, Jin Billy Li, Mark J. Daly, Roderic Guigo, Peter Donnelly, Kristin Ardlie, Michael Sammeth, Emmanouil T. Dermitzakis, Mark I. McCarthy, Stephen B. Montgomery, Tuuli Lappalainen, Daniel G. MacArthur, Ayellet V. Segre, Taylor R. Young, Ellen T. Gelfand, Casandra A. Trowbridge, Lucas D. Ward, Pouya Kheradpour, Benjamin Iriarte, Yan Meng, Cameron D. Palmer, Tonu Esko, Wendy Winckler, Joel Hirschhorn, Manolis Kellis, Gad Getz, Andrey A. Shablin, Gen Li, Yi-Hui Zhou, Andrew B. Nobel, Ivan Rusyn, Fred A. Wright, Alexis Battle, Sara Mostafavi, Marta Mele, Ferran Reverter, Jakob Goldmann, Daphne Koller, Eric R. Gamazon, Hae Kyung Im, Anuar Konkashbaev, Dan L. Nicolae, Nancy J. Cox, Timothe Flutre, Xiaoquan Wen, Matthew Stephens, Jonathan K. Pritchard, Zhidong Tu, Bin Zhang, Tao Huang, Quan Long, Luan Lin, Jialiang Yang, Jun Zhu, Jun Liu, Amanda Brown, Bernadette Mestichelli, Denee Tidwell, Edmund Lo, Mike Salvatore, Saboor Shad, Jeffrey A. Thomas, John T. Lonsdale, Roswell Christopher Choi, Ellen Karasik, Kimberly Ramsey, Michael T. Moser, Barbara A. Foster, Bryan M. Gillard, John Syron, Johnelle Fleming, Harold Magazine, Rick Hasz, et al. Effect of predicted protein-truncating genetic variants on the human transcriptome. *Science*, 348(6235):666–669, 2015.

[58] Daniel G MacArthur, Suganthi Balasubramanian, Adam Frankish, Ni Huang, James Morris, Klaudia Walter, Luke Jostins, Lukas Habegger, Joseph K Pickrell, and Stephen B Montgomery. A systematic survey of loss-of-function variants in human protein-coding genes. *Science*, 335(6070):823–828, 2012.

[59] Tuuli Lappalainen, Michael Sammeth, Marc R Friedländer, Peter AC 't Hoen, Jean Monlong, Manuel A Rivas, Mar Gonzalez-Porta, Natalja Kurbatova, Thasso Griebel, and Pedro G Ferreira. Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, 501(7468):506–511, 2013.

[60] Rik G. H. Lindeboom, Fran Supek, and Ben Lehner. The rules and impact of nonsense-mediated mrna decay in human cancers. *Nature Genetics*, 48(10):1112–1118, 2016.

[61] L. W. Harries, Coralie Bingham, Christine Bellanne-Chantelot, A. T. Hattersley, and Sian Ellard. The position of premature termination codons in the hepatocyte nuclear factor 1 beta gene determines susceptibility to nonsense-mediated decay. *Human Genetics*, 118(2):214–224, 2005.

[62] Monique Buisson, Olga Anczuków, Almoutassem B. Zetoune, Mark D. Ware, and Sylvie Mazoyer. The 185delag mutation (c.68_69delag) in the brca1 gene triggers translation reinitiation at a downstream aug codon. *Human Mutation*, 27(10):1024–1029, 2006.

[63] Jing Zhang and Lynne E. Maquat. Evidence that translation reinitiation abrogates nonsense-mediated mrna decay in mammalian cells. *The EMBO Journal*, 16(4):826–833, 1997.

[64] Lynne E. Maquat and Xiaojie Li. Mammalian heat shock p70 and histone h4 transcripts, which derive from naturally intronless genes, are immune to nonsense-mediated decay. *RNA*, 7(3):445–456, 2001.

[65] Katja S. Brocke, Gabriele Neu-Yilik, Niels H. Gehring, Matthias W. Hentze, and Andreas E. Kulozik. The human intronless melanocortin 4-receptor gene is nmd insensitive. *Human Molecular Genetics*, 11(3):331–335, 2002.

[66] Z. Coban-Akdemir, J. J. White, X. Song, S. N. Jhangiani, J. M. Fatih, T. Gambin, Y. Bayram, I. K. Chinn, E. Karaca, J. Punetha, C. Poli, E. Boerwinkle, C. A. Shaw, J. S. Orange, R. A. Gibbs, T. Lappalainen, J. R. Lupski, and C. M. B. Carvalho. Identifying genes whose mutant transcripts cause dominant disease traits by potential gain-of-function alleles. *Am J Hum Genet*, 103(2):171–187, 2018.

[67] Suganthi Balasubramanian, Yao Fu, Mayur Pawashe, Patrick McGillivray, Mike Jin, Jeremy Liu, Konrad J. Karczewski, Daniel G. MacArthur, and Mark Gerstein. Using aloft to determine the impact of putative loss-of-function variants in protein-coding genes. *Nature Communications*, 8(1):382, 2017.

[68] Pablo Cingolani, Adrian Platts, Le Lily Wang, Melissa Coon, Tung Nguyen, Luan Wang, Susan J Land, Xiangyi Lu, and Douglas M Ruden. A program for annotating and predicting the effects of single nucleotide polymorphisms, snpeff: Snps in the genome of drosophila melanogaster strain w1118; iso-2; iso-3. *fly*, 6(2):80–92, 2012.

[69] Dan G Blazer and Lyla M Hernandez. Genes, behavior, and the social environment: Moving beyond the nature/nurture debate. 2006.

[70] William Kermode, Dianne De Santis, Linh Truong, Erika Della Mina, Sam Salman, Grace Thompson, David Nolan, Richard Loh, Dominic Mallon, Andrew McLean-Tooke, Mina John, Stuart G. Tangye, Michael O'Sullivan, and Lloyd J. D'Orsogna. A novel targeted amplicon next-generation sequencing gene panel for the diagnosis of common variable immunodeficiency has a high diagnosticxa0;yield: Results from the perth cvid cohort study. *The Journal of Molecular Diagnostics*, 24(6):586–599, 2022.

[71] Derek Hong and Lilia M. Iakoucheva. Therapeutic strategies for autism: targeting three levels of the central dogma of molecular biology. *Translational Psychiatry*, 13(1):58, 2023.

[72] Ákos Zsembery, Wolfgang Jessner, Gerlinde Sitter, Carlo Spirlí, Mario Strazzabosco, and Jürg Graf. Correction of cftr malfunction and stimulation of ca2+-activated cl channels restore hco secretion in cystic fibrosis bile ductular cells. *Hepatology*, 35(1):95–104, 2002.

[73] MICHAEL WILSCHANSKI, CHAGIT FAMINI, HANNAH BLAU, JOSEPH RIVLIN, ARIEH AUGARTEN, AVRAHAM AVITAL, BATSHEVA KEREM, and EITAN KEREM. A pilot study of the effect of gentamicin on nasal potential difference measurements in cystic fibrosis patients carrying stop mutations. *American journal of respiratory and critical care medicine*, 161(3):860–865, 2000.

[74] Kathryn R Wagner, Sherifa Hamed, Donald W Hadley, Andrea L Gropman, Aaron H Burstein, Diana M Escolar, Eric P Hoffman, and Kenneth H Fischbeck. Gentamicin treatment of duchenne and becker muscular dystrophy due to nonsense mutations. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 49(6):706–711, 2001.

[75] David M Bedwell, Anisa Kaenjak, Dale J Benos, Zsuzsa Bebok, James K Bubien, Jeong Hong, Albert Tousson, JP Clancy, and Eric J Sorscher. Suppression of a cftr premature stop mutation in a bronchial epithelial cell line. *Nature medicine*, 3(11):1280–1284, 1997.

[76] Jill A Holbrook, Gabriele Neu-Yilik, Matthias W Hentze, and Andreas E Kulozik. Nmd and human disease. In *Madame Curie Bioscience Database*. Landes Bioscience, Austin, TX, 2000-2013.

[77] Liang Qianqian, Abraham Abin, A. Capra John, and Kostka Dennis. Disease-specific prioritization of non-coding gwas variants based on chromatin accessibility. *medRxiv*, page 2023.10.17.23297164, 2023.

[78] A. Buniello, J. A. L. MacArthur, M. Cerezo, L. W. Harris, J. Hayhurst, C. Malangone, A. McMahon, J. Morales, E. Mountjoy, E. Sollis, D. Suveges, O. Vrousgou, P. L. Whetzel, R. Amode, J. A. Guillen, H. S. Riat, S. J. Trevanion, P. Hall, H. Junkins, P. Flicek, T. Burdett, L. A. Hindorff, F. Cunningham, and H. Parkinson. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res*, 47(D1):D1005–d1012, 2019.

[79] Joon-Yong An, Kevin Lin, Lingxue Zhu, Donna M. Werling, Shan Dong, Harrison Brand, Harold Z. Wang, Xuefang Zhao, Grace B. Schwartz, Ryan L. Collins, Benjamin B. Currall, Claudia Dastmalchi, Jeanselle Dea, Clif Duhn, Michael C. Gilson, Lambertus Klei, Lindsay Liang, Eirene Markenscoff-Papadimitriou, Sirisha Pochareddy, Nadav Ahituv, Joseph D. Buxbaum, Hilary Coon, Mark J. Daly, Young Shin Kim, Gabor T. Marth, Benjamin M. Neale, Aaron R. Quinlan, John L. Rubenstein, Nenad Sestan, Matthew W. State, A. Jeremy Willsey, Michael E. Talkowski, Bernie Devlin, Kathryn Roeder, and Stephan J. Sanders. Genome-wide de novo risk score implicates promoter variation in autism spectrum disorder. *Science (New York, N.Y.)*, 362(6420):eaat6576, 2018.

[80] Ernest Turro, William J. Astle, Karyn Megy, Stefan Gräf, Daniel Greene, Olga Shamardina, Hana Lango Allen, Alba Sanchis-Juan, Mattia Frontini, Chantal Thys, Jonathan Stephens, Rutendo Mapeta, Oliver S. Burren, Kate Downes, Matthias Haimel, Salih Tuna, Sri V. V. Deevi, Timothy J. Aitman, David L. Bennett, Paul Calleja, Keren Carss, Mark J. Caulfield, Patrick F. Chinnery, Peter H. Dixon, Daniel P. Gale, Roger James, Ania Koziell, Michael A. Laffan, Adam P. Levine, Eamonn R. Maher, Hugh S. Markus, Joannella Morales, Nicholas W. Morrell, Andrew D. Mumford, Elizabeth Ormondroyd, Stuart Rankin, Augusto Rendon, Sylvia Richardson, Irene Roberts, Noemi B. A. Roy, Moin A. Saleem, Kenneth G. C. Smith, Hannah Stark, Rhea Y. Y. Tan, Andreas C. Themistocleous, Adrian J. Thrasher, Hugh Watkins, Andrew R. Webster, Martin R. Wilkins, Catherine Williamson, James

Whitworth, Sean Humphray, David R. Bentley, Genomes Project Nihr BioResource for the, Nathalie Kingston, Neil Walker, John R. Bradley, Sofie Ashford, Christopher J. Penkett, Kathleen Freson, Kathleen E. Stirrups, F. Lucy Raymond, and Willem H. Ouwehand. Whole-genome sequencing of patients with rare diseases in a national health system. *Nature*, 583(7814):96–102, 2020.

[81] Kerstin Lindblad-Toh, Manuel Garber, Or Zuk, Michael F. Lin, Brian J. Parker, Stefan Washietl, Pouya Kheradpour, Jason Ernst, Gregory Jordan, Evan Mauceli, Lucas D. Ward, Craig B. Lowe, Alisha K. Holloway, Michele Clamp, Sante Gnerre, Jessica Alföldi, Kathryn Beal, Jean Chang, Hiram Clawson, James Cuff, Federica Di Palma, Stephen Fitzgerald, Paul Flicek, Mitchell Guttman, Melissa J. Hubisz, David B. Jaffe, Irwin Jungreis, W. James Kent, Dennis Kostka, Marcia Lara, Andre L. Martins, Tim Massingham, Ida Moltke, Brian J. Raney, Matthew D. Rasmussen, Jim Robinson, Alexander Stark, Albert J. Vilella, Jiayu Wen, Xiaohui Xie, Michael C. Zody, Platform Broad Institute Sequencing, Team Whole Genome Assembly, Jen Baldwin, Toby Bloom, Chee Whye Chin, Dave Heiman, Robert Nicol, Chad Nusbaum, Sarah Young, Jane Wilkinson, Kim C. Worley, Christie L. Kovar, Donna M. Muzny, Richard A. Gibbs, Team Baylor College of Medicine Human Genome Sequencing Center Sequencing, Andrew Cree, Huyen H. Dihn, Gerald Fowler, Shalili Jhangiani, Vandita Joshi, Sandra Lee, Lora R. Lewis, Lynne V. Nazareth, Geoffrey Okwuonu, Jireh Santibanez, Wesley C. Warren, Elaine R. Mardis, George M. Weinstock, Richard K. Wilson, University Genome Institute at Washington, Kim Delehaunty, David Dooling, Catrina Fronik, Lucinda Fulton, Bob Fulton, Tina Graves, Patrick Minx, Erica Sodergren, Ewan Birney, Elliott H. Margulies, Javier Herrero, Eric D. Green, David Haussler, Adam Siepel, Nick Goldman, Katherine S. Pollard, Jakob S. Pedersen, Eric S. Lander, and Manolis Kellis. A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, 478(7370):476–482, 2011.

[82] D. Backenroth, Z. He, K. Kiryluk, V. Boeva, L. Pethukova, E. Khurana, A. Christiano, J. D. Buxbaum, and I. Ionita-Laza. FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am J Hum Genet*, 102(5):920–942, 2018.

[83] Kévin Vervier and Jacob J. Michaelson. Tisan: estimating tissue-specific effects of coding and non-coding variants. *Bioinformatics*, 34(18):3061–3068, 2018.

[84] Carles A. Boix, Benjamin T. James, Yongjin P. Park, Wouter Meuleman, and Manolis Kellis. Regulatory genomic circuitry of human disease loci by integrative epigenomics. *Nature*, 590(7845):300–307, 2021.

[85] James Malone, Ele Holloway, Tomasz Adamusiak, Misha Kapushesky, Jie Zheng, Nikolay Kolesnikov, Anna Zhukova, Alvis Brazma, and Helen Parkinson. Modeling

sample variables with an experimental factor ontology. *Bioinformatics*, 26(8):1112–1118, 2010.

[86] Tune H. Pers, Pascal Timshel, and Joel N. Hirschhorn. SNPsnap: a Web-based tool for identification and annotation of matched SNPs. *Bioinformatics*, 31(3):418–420, 2015.

[87] J Schreiber, J Bilmes, and WS Noble. Completing the ENCODE3 compendium yields accurate imputations across a variety of assays and human biosamples. *Genome Biol*, 21:82, Mar 2020.

[88] J Friedman, T Hastie, and R Tibshirani. Regularization Paths for Generalized Linear Models via Coordinate Descent. *J Stat Softw*, 33:1–22, 2010.

[89] Gitte Vanwinckelen and Hendrik Blockeel. On estimating model accuracy with repeated cross-validation. *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, pages 39–44, 2012.

[90] Janez Demsar. Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, (7):1–30, 2006.

[91] EM Ramos, D Hoffman, HA Junkins, D Maglott, L Phan, ST Sherry, M Feolo, and LA Hindorff. Phenotype-Genotype Integrator (PheGenI): synthesizing genome-wide association study (GWAS) data with existing genomic resources. *Eur J Hum Genet*, 22:144–7, Jan 2014.

[92] Kyoko Watanabe, Sven Stringer, Oleksandr Frei, Maša Umićević Mirkov, Christiaan de Leeuw, Tinca J. C. Polderman, Sophie van der Sluis, Ole A. Andreassen, Benjamin M. Neale, and Danielle Posthuma. A global overview of pleiotropy and genetic architecture in complex traits. *Nature Genetics*, 51(9):1339–1348, 2019.

[93] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Systemic scleroderma.

[94] Marina D Kraaij and Jacob M van Laar. The role of b cells in systemic sclerosis. *Biologics: Targets and Therapy*, 2(3):389–395, 2008.

[95] Benjamin Thoreau, Benjamin Chaigne, and Luc Mouthon. Role of b-cell in the pathogenesis of systemic sclerosis. *Frontiers in Immunology*, 13:933468, 2022.

[96] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Primary sclerosing cholangitis.

[97] Lilly Kristin Kunzmann, Tanja Schoknecht, Tobias Poch, Lara Henze, Stephanie Stein, Marvin Kriz, Ilka Grewe, Max Preti, Johannes Hartl, and Nadine Pannicke. Monocytes as potential mediators of pathogen-induced t-helper 17 differentiation in patients with primary sclerosing cholangitis (psc). *Hepatology*, 72(4):1310–1326, 2020.

[98] Temitope O Keku, Joseph A Galanko, Sharon C Murray, John T Woosley, and Robert S Sandler. Rectal mucosal proliferation, dietary factors, and the risk of colorectal adenomas. *Cancer epidemiology, biomarkers prevention: a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology*, 7(11):993–999, 1998.

[99] Santosh Dulal and Temitope O Keku. Gut microbiome and colorectal adenomas. *Cancer journal (Sudbury, Mass.)*, 20(3):225, 2014.

[100] Yuanhao Yang, Hannah Musco, Steve Simpson-Yap, Zhihong Zhu, Ying Wang, Xin Lin, Jiawei Zhang, Bruce Taylor, Jacob Gratten, and Yuan Zhou. Investigating the shared genetic architecture between multiple sclerosis and inflammatory bowel diseases. *Nature Communications*, 12(1):5641, 2021.

[101] K. K. Farh, A. Marson, J. Zhu, M. Kleinewietfeld, W. J. Housley, S. Beik, N. Shoresh, H. Whitton, R. J. Ryan, A. A. Shishkin, M. Hatan, M. J. Carrasco-Alfonso, D. Mayer, C. J. Luckey, N. A. Patsopoulos, P. L. De Jager, V. K. Kuchroo, C. B. Epstein, M. J. Daly, D. A. Hafler, and B. E. Bernstein. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature*, 518(7539):337–43, 2015.

[102] C. McDowell, U. Farooq, and M. Haseeb. *Inflammatory Bowel Disease*. StatPearls Publishing Copyright © 2022, StatPearls Publishing LLC., Treasure Island (FL), 2022.

[103] C. Lord, M. Elsabbagh, G. Baird, and J. Veenstra-Vanderweele. Autism spectrum disorder. *Lancet*, 392(10146):508–520, 2018.

[104] G. Olivo, S. Gaudio, and H. B. Schiöth. Brain and Cognitive Development in Adolescents with Anorexia Nervosa: A Systematic Review of fMRI Studies. *Nutrients*, 11(8), 2019.

[105] E. R. Sigmon, M. Kelleman, A. Susi, C. M. Nylund, and M. E. Oster. Congenital Heart Disease and Autism: A Case-Control Study. *Pediatrics*, 144(5), 2019.

[106] Z. C. Zhou, D. B. McAdam, and D. R. Donnelly. Endophenotypes: A conceptual link between anorexia nervosa and autism spectrum disorder. *Research in Developmental Disabilities*, 82:153–165, 2018.

[107] Margherita Boltri and Walter Sapuppo. Anorexia nervosa and autism spectrum disorder: A systematic review. *Psychiatry Research*, 306:114271, 2021.

[108] D. S. Tylee, J. Sun, J. L. Hess, M. A. Tahir, E. Sharma, R. Malik, B. B. Worrall, A. J. Levine, J. J. Martinson, S. Nejentsev, D. Speed, A. Fischer, E. Mick, B. R. Walker, A. Crawford, S. F. A. Grant, C. Polychronakos, J. P. Bradfield, P. M. A. Sleiman, H. Hakonarson, E. Ellinghaus, J. T. Elder, L. C. Tsoi, R. C. Trembath, J. N. Barker, A. Franke, A. Dehghan, S. V. Faraone, and S. J. Glatt. Genetic correlations among psychiatric and immune-related phenotypes based on genome-wide association data. *Am J Med Genet B Neuropsychiatr Genet*, 177(7):641–657, 2018.

[109] C. Y. Li, T. M. Yang, R. W. Ou, Q. Q. Wei, and H. F. Shang. Genome-wide genetic links between amyotrophic lateral sclerosis and autoimmune diseases. *BMC Med*, 19(1):27, 2021.

[110] X. Yu, J. Vargas, P. H. R. Green, and G. Bhagat. Innate Lymphoid Cells and Celiac Disease: Current Perspective. *Cell Mol Gastroenterol Hepatol*, 11(3):803–814, 2021.

[111] B. Jabri and L. M. Sollid. T Cells in Celiac Disease. *J Immunol*, 198(8):3005–3014, 2017.

[112] Jingsi Ming, Mingwei Dai, Mingxuan Cai, Xiang Wan, Jin Liu, and Can Yang. Lsmm: a statistical approach to integrating functional annotations with genome-wide association studies. *Bioinformatics*, 34(16):2788–2796, 2018.

[113] A. Julià, F. J. López-Longo, J. J. Pérez Venegas, S. Bonàs-Guarch, À Olivé, J. L. Andreu, MÁ Aguirre-Zamorano, P. Vela, J. M. Nolla, J. L. M. de la Fuente, A. Zea, J. M. Pego-Reigosa, M. Freire, E. Díez, E. Rodríguez-Almaraz, P. Carreira, R. Blanco, V. M. Taboada, M. López-Lasanta, M. L. Corbeto, J. M. Mercader, D. Torrents, D. Absher, S. Marsal, and A. Fernández-Nebro. Genome-wide association study meta-analysis identifies five new loci for systemic lupus erythematosus. *Arthritis Res Ther*, 20(1):100, 2018.

[114] H. Lu, J. Zhang, Z. Jiang, M. Zhang, T. Wang, H. Zhao, and P. Zeng. Detection of genetic overlap between rheumatoid arthritis and systemic lupus erythematosus using gwas summary statistics. *Front Genet*, 12:656545, 2021.

[115] J. Bentham, D. L. Morris, D. S. C. Graham, C. L. Pinder, P. Tombleson, T. W. Behrens, J. Martín, B. P. Fairfax, J. C. Knight, L. Chen, J. Replogle, A. C. Syvänen, L. Rönnblom, R. R. Graham, J. E. Wither, J. D. Rioux, M. E. Alarcón-Riquelme, and T. J. Vyse. Genetic association analyses implicate aberrant regulation of innate and adaptive immunity genes in the pathogenesis of systemic lupus erythematosus. *Nat Genet*, 47(12):1457–1464, 2015.

[116] R. E. Peterson, K. Kuchenbaecker, R. K. Walters, C. Y. Chen, A. B. Popejoy, S. Periyasamy, M. Lam, C. Iyegbe, R. J. Strawbridge, L. Brick, C. E. Carey, A. R. Martin, J. L. Meyers, J. Su, J. Chen, A. C. Edwards, A. Kalungi, N. Koen, L. Majara, E. Schwarz, J. W. Smoller, E. A. Stahl, P. F. Sullivan, E. Vassos, B. Mowry, M. L. Prieto, A. Cuellar-Barboza, T. B. Bigdeli, H. J. Edenberg, H. Huang, and L. E. Duncan. Genome-wide association studies in ancestrally diverse populations: Opportunities, methods, pitfalls, and recommendations. *Cell*, 179(3):589–603, 2019.

[117] Kye Won Park, Ho-Sung Ryu, Eunsoon Shin, YoonGi Park, Sang Ryong Jeon, Seong Yoon Kim, Jae Seung Kim, Seong-Beom Koh, and Sun Ju Chung. Ethnicity- and sex-specific genome wide association study on parkinson's disease. *npj Parkinson's Disease*, 9(1):141, 2023.

[118] Gitte Vanwinckelen and Hendrik Blockeel. On estimating model accuracy with repeated cross-validation. *BeneLearn 2012: Proceedings of the 21st Belgian-Dutch Conference on Machine Learning*, pages 39–44, 2012.

[119] James Z. Wang, Zhidian Du, Rapeeporn Payattakool, Philip S. Yu, and Chin-Fu Chen. A new method to measure the semantic similarity of go terms. *Bioinformatics*, 23(10):1274–1281, 2007.

[120] Guangchuang Yu. Using meshes for mesh term enrichment and semantic analyses. *Bioinformatics*, 34(21):3766–3767, 2018.

[121] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Celiac disease.

[122] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Hypothyroidism.

[123] C. L. Ch'ng, M. K. Jones, and J. G. Kingham. Celiac disease and autoimmune thyroid disease. *Clin Med Res*, 5(3):184–92, 2007.

[124] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Multiple myeloma.

[125] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Multiple sclerosis.

[126] S. H. Tsung. Monoclonal gammopathy associated with multiple sclerosis. *Ann Clin Lab Sci*, 8(6):472–5, 1978.

[127] E. Melamed and M. W. Lee. Multiple sclerosis and cancer: The ying-yang effect of disease modifying therapies. *Front Immunol*, 10:2954, 2019.

[128] C. M. Kitahara, D. Krmendiné Farkas, J. O. L. Jørgensen, D. Cronin-Fenton, and H. T. Sørensen. Benign thyroid diseases and risk of thyroid cancer: a nationwide cohort study. *Journal of Clinical Endocrinology and Metabolism*, 103:2216–2224, 2018.

[129] M. Cristofanilli, Y. Yamamura, S. W. Kau, T. Bevers, S. Strom, M. Patangan, L. Hsu, S. Krishnamurthy, R. L. Theriault, and G. N. Hortobagyi. Thyroid hormone and breast carcinoma: primary hypothyroidism is associated with a reduced incidence of primary breast carcinoma. *Cancer*, 103:1122–1128, 2005.

[130] S. Balasubramaniam, E. Ron, G. Gridley, A. B. Schneider, and A. V. Brenner. Association between benign thyroid and endocrine disorders and subsequent risk of thyroid cancer among 4.5 million us male veterans. *Journal of Clinical Endocrinology and Metabolism*, 97:2661–2669, 2012.

[131] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Gout.

[132] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Juvenile arthritis.

[133] MedlinePlus [Internet]. Bethesda (MD): National Library of Medicine (US). Eating disorders.

[134] Jonathan Klonowski, Qianqian Liang, Zeynep Coban-Akdemir, Cecilia Lo, and Dennis Kostka. aenmd: Annotating escape from nonsense-mediated decay for transcripts with protein-truncating variants. *bioRxiv*, 2023.

[135] Jonathan Klonowski, Qianqian Liang, Zeynep Coban-Akdemir, Cecilia Lo, and Dennis Kostka. aenmd: annotating escape from nonsense-mediated decay for transcripts with protein-truncating variants. *Bioinformatics*, 39(9), 2023.

[136] Søren Lykke-Andersen and Torben Heick Jensen. Nonsense-mediated mrna decay: an intricate machinery that shapes transcriptomes. *Nature Reviews Molecular Cell Biology*, 16(11):665–677, 2015.

[137] Lisa Backwell and Joseph A. Marsh. Diverse molecular mechanisms underlying pathogenic protein mutations: Beyond the loss-of-function paradigm. *Annual Review of Genomics and Human Genetics*, 23(1):475–498, 2022.

[138] Madhuri Bhuvanagiri, Anna M Schlitter, Matthias W Hentze, and Andreas E Kulozik. Nmd: Rna biology meets human genetic medicine. *Biochemical Journal*, 430(3):365–377, 2010.

[139] Jiu Cheng, Phillip Belgrader, Xianbo Zhou, and Lynne E. Maquat. Introns are cis effectors of the nonsense-codon-mediated reduction in nuclear mrna abundance. *Molecular and Cellular Biology*, 14(9):6317–6325, 1994.

[140] Lukas Gerasimavicius, Benjamin J Livesey, and Joseph A Marsh. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature communications*, 13(1):3895, 2022.

[141] Ângela Inácio, Ana Luísa Silva, Ana Morgado, Francisco J. C. Pereira, João Lavinha, and Luísa Romão. Comment on 'nonsense-mediated mrna decay modulates clinical outcome of genetic disease'. *European Journal of Human Genetics*, 15(5):533–534, 2007.

[142] Ken Inoue, Mehrdad Khajavi, Tomoko Ohyama, Shin-ichi Hirabayashi, John Wilson, James D. Reggin, Pedro Mancias, Ian J. Butler, Miles F. Wilkinson, Michael Wegner, and James R. Lupski. Molecular mechanism for distinct neurological phenotypes conveyed by allelic truncating mutations. *Nature Genetics*, 36(4):361–369, 2004.

[143] Ken Inoue, Tomoko Ohyama, Yosuke Sakuragi, Ryoko Yamamoto, Naoko A. Inoue, Yu Li-Hua, Yu-ichi Goto, Michael Wegner, and James R. Lupski. Translation of

sox10 3 untranslated region causes a complex severe neurocristopathy by generation of a deleterious functional domain. *Human Molecular Genetics*, 16(24):3037–3046, 2007.

[144] Noriko Miyake, Hidehisa Takahashi, Kazuyuki Nakamura, Bertrand Isidor, Yoko Hiraki, Eriko Koshimizu, Masaaki Shiina, Kazunori Sasaki, Hidefumi Suzuki, and Ryota Abe. Gain-of-function mn1 truncation variants cause a recognizable syndrome with craniofacial and brain abnormalities. *The American Journal of Human Genetics*, 106(1):13–25, 2020.

[145] Yavuz Bayram, Janson J White, Nursel Elcioglu, Megan T Cho, Neda Zadeh, Asuman Gedikbasi, Sukru Palanduz, Sukru Ozturk, Kivanc Cefle, and Ozgur Kasapcopur. Rest final-exon-truncating mutations cause hereditary gingival fibromatosis. *The American Journal of Human Genetics*, 101(1):149–156, 2017.

[146] Pamela A. Frischmeyer and Harry C. Dietz. Nonsense-mediated mrna decay in health and disease. *Human Molecular Genetics*, 8(10):1893–1900, 1999.

[147] Georgina W Hall and Sweelay Thein. Nonsense codon mutations in the terminal exon of the beta-globin gene are not associated with a reduction in beta-mrna accumulation: a mechanism for the phenotype of dominant beta-thalassemia. 1994.

[148] Angela Inácio, Ana Luísa Silva, Joana Pinto, Xinjun Ji, Ana Morgado, Fátima Almeida, Paula Faustino, João Lavinha, Stephen A Liebhaber, and Luísa Romão. Nonsense mutations in close proximity to the initiation codon fail to trigger full nonsense-mediated mrna decay. *Journal of Biological Chemistry*, 279(31):32170–32180, 2004.

[149] Sandra Jansen, Sinje Geuer, Rolph Pfundt, Rachel Brough, Priyanka Ghongane, Johanna C Herkert, Elysa J Marco, Marjolein H Willemsen, Tjitske Kleefstra, and Mark Hannibal. De novo truncating mutations in the last and penultimate exons of ppm1d cause an intellectual disability syndrome. *The American Journal of Human Genetics*, 100(4):650–658, 2017.

[150] Tim P Kerr, Caroline A Sewry, Stephanie A Robb, and Roland G Roberts. Long mutant dystrophins and variable phenotypes: evasion of nonsense-mediated decay? *Human genetics*, 109:402–407, 2001.

[151] Hong Joo Kim, Payam Mohassel, Sandra Donkervoort, Lin Guo, Kevin O'donovan, Maura Coughlin, Xaviere Lornage, Nicola Foulds, Simon R Hammans, and A Reghan

Foley. Heterozygous frameshift variants in hnrnpa2b1 cause early-onset oculopharyngeal muscular dystrophy. *Nature communications*, 13(1):2306, 2022.

[152] Matthew Mort, Dobril Ivanov, David N Cooper, and Nadia A Chuzhanova. A meta-analysis of nonsense mutations causing human genetic disease. *Human mutation*, 29(8):1037–1047, 2008.

[153] M Cecilia Poli, Frédéric Ebstein, Sarah K Nicholas, Marietta M de Guzman, Lisa R Forbes, Ivan K Chinn, Emily M Mace, Tiphanie P Vogel, Alexandre F Carisey, and Felipe Benavides. Heterozygous truncating variants in pomp escape nonsense-mediated decay and cause a unique immune dysregulatory syndrome. *The American Journal of Human Genetics*, 102(6):1126–1142, 2018.

[154] Janson White, Juliana F Mazzeu, Alexander Hoischen, Shalini N Jhangiani, Tomasz Gambin, Michele Calijorne Alcino, Samantha Penney, Jorge M Saraiva, Hanne Hove, and Flemming Skovby. Dvl1 frameshift mutations clustering in the penultimate exon cause autosomal-dominant robinow syndrome. *The American Journal of Human Genetics*, 96(4):612–622, 2015.

[155] Janson J White, Juliana F Mazzeu, Alexander Hoischen, Yavuz Bayram, Marjorie Withers, Alper Gezdirici, Virginia Kimonis, Marloes Steehouwer, Shalini N Jhangiani, and Donna M Muzny. Dvl3 alleles resulting in a 1 frameshift of the last exon mediate autosomal-dominant robinow syndrome. *The American Journal of Human Genetics*, 98(3):553–561, 2016.

[156] Kohei Hamanaka, Eri Imagawa, Eriko Koshimizu, Satoko Miyatake, Jun Tohyama, Takanori Yamagata, Akihiko Miyauchi, Nina Ekhilevitch, Fumio Nakamura, and Takeshi Kawashima. De novo truncating variants in the last exon of sema6b cause progressive myoclonic epilepsy. *The American Journal of Human Genetics*, 106(4):549–558, 2020.

[157] Eszter Nagy and Lynne E Maquat. A rule for termination-codon position within intron-containing genes: when nonsense affects rna abundance. *Trends in biochemical sciences*, 23(6):198–199, 1998.

[158] Hervé Le Hir, Elisa Izaurralde, Lynne E Maquat, and Melissa J Moore. The spliceosome deposits multiple proteins 20–24 nucleotides upstream of mrna exon–exon junctions. *The EMBO journal*, 19(24):6860–6869, 2000.

[159] Yasuhito Ishigaki, Xiaojie Li, Guillaume Serin, and Lynne E Maquat. Evidence for a pioneer round of mrna translation: mrnas subject to nonsense-mediated decay in mammalian cells are bound by cbp80 and cbp20. *Cell*, 106(5):607–617, 2001.

[160] Hervé Le Hir, David Gatfield, Elisa Izaurralde, and Melissa J Moore. The exon–exon junction complex provides a binding platform for factors involved in mrna export and nonsense-mediated mrna decay. *The EMBO journal*, 20(17):4987–4997, 2001.

[161] V Narry Kim, Naoyuki Kataoka, and Gideon Dreyfuss. Role of the nonsense-mediated decay factor hupf3 in the splicing-dependent exon-exon junction complex. *Science*, 293(5536):1832–1836, 2001.

[162] Jens Lykke-Andersen, Mei-Di Shu, and Joan A Steitz. Communication of the position of exon-exon junctions to the mrna surveillance machinery by the protein rnps1. *Science*, 293(5536):1836–1839, 2001.

[163] Niels H Gehring, Gabriele Neu-Yilik, Thomas Schell, Matthias W Hentze, and Andreas E Kulozik. Y14 and hupf3b form an nmd-activating complex. *Molecular cell*, 11(4):939–949, 2003.

[164] Tim A Hoek, Deepak Khuperkar, Rik GH Lindeboom, Stijn Sonneveld, Bram MP Verhagen, Sanne Boersma, Michiel Vermeulen, and Marvin E Tanenbaum. Single-molecule imaging uncovers rules governing nonsense-mediated mrna decay. *Molecular cell*, 75(2):324–339. e11, 2019.

[165] Gonçalo Nogueira, Rafael Fernandes, Juan F García-Moreno, and Luísa Romão. Nonsense-mediated rna decay and its bipolar function in cancer. *Molecular Cancer*, 20(1):1–19, 2021.

[166] Ana Luísa Silva, Francisco JC Pereira, Ana Morgado, Jian Kong, Rute Martins, Paula Faustino, Stephen A Liebhaber, and Luísa Romão. The canonical upf1-dependent nonsense-mediated mrna decay is inhibited in transcripts carrying a short open reading frame independent of sequence context. *Rna*, 12(12):2160–2170, 2006.

[167] Melissa J Landrum, Jennifer M Lee, Mark Benson, Garth R Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, Jennifer Hart, Douglas Hoffman, and Wonhee Jang. Clinvar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

[168] Konrad J Karczewski, Laurent C Francioli, Grace Tiao, Beryl B Cummings, Jessica Alföldi, Qingbo Wang, Ryan L Collins, Kristen M Laricchia, Andrea Ganna, and Daniel P Birnbaum. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, 581(7809):434–443, 2020.

[169] Elliot Sollis, Abayomi Mosaku, Ala Abid, Annalisa Buniello, Maria Cerezo, Laurent Gil, Tudor Groza, Osman Güneş, Peggy Hall, and James Hayhurst. The nhgri-ebi gwas catalog: knowledgebase and deposition resource. *Nucleic acids research*, 51(D1):D977–D985, 2023.

[170] Adrian Tan, Gonçalo R Abecasis, and Hyun Min Kang. Unified representation of genetic variants. *Bioinformatics*, 31(13):2202–2204, 2015.

[171] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2022.

[172] Valerie Obenchain, Michael Lawrence, Vincent Carey, Stephanie Gogarten, Paul Shannon, and Martin Morgan. Variantannotation: a bioconductor package for exploration and annotation of genetic variants. *Bioinformatics*, 30(14):2076–2078, 2014.

[173] Brian J Knaus and Niklaus J Grünwald. vcfr: a package to manipulate and visualize variant call format data in r. *Molecular ecology resources*, 17(1):44–53, 2017.

[174] Sergio Lifschitz, Edward H Haeusler, Marcos Catanho, Antonio B de Miranda, Elvismary Molina de Armas, Alexandre Heine, Sergio GMP Moreira, and Cristian Tristão. Bio-strings: A relational database data-type for dealing with large biosequences. *BioTech*, 11(3):31, 2022.

[175] Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T Morgan, and Vincent J Carey. Software for computing and annotating genomic ranges. *PLoS computational biology*, 9(8):e1003118, 2013.

[176] Nicole A. Teran, Daniel C. Nachun, Tiffany Eulalio, Nicole M. Ferraro, Craig Smail, Manuel A. Rivas, and Stephen B. Montgomery. Nonsense-mediated decay is highly stable across individuals and tissues. *The American Journal of Human Genetics*, 108(8):1401–1408, 2021.

[177] Julia Gustavsen, Sina Rüeger, Scott Chamberlain, Kevin Ushey, and Hao Zhu. *rsnps: Get 'SNP' ('Single-Nucleotide' 'Polymorphism') Data on the Web*, 2022. R package version 0.5.0.0.