Computational methods for the analysis of epigenomic activity in the human immune system



Miriam Stricker

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

11th January 2023

Acknowledgements

Personal

This thesis could not have been written without the support, encouragement, and love of the people around me.

First, I would like to thank my supervisors, Pier Palamara and Satu Nahkuri, who guided me through the DPhil process. Pier and Satu tirelessly met with me most weeks, helped me when I was stuck, and gave detailed and valuable feedback on all abstracts, posters, practice presentations, manuscripts, and more. I also want to thank my secondary supervisors, Simon Myers (and research group) and Wei-Yi Cheng, for their valuable discussions. Also remembering all my mentors who brought me here in the first place, to name Ilme Schlichting and Bruce Doak.

My colleagues and friends at the Department of Statistics made the work over the last three years a joy. I loved the chats we had over lunch and in the coffee breaks, the overall supporting atmosphere at the department.

I also had tremendous support from friends from outside the office. The people in my SABS cohort, the friends I made through rowing with Wolfson College Boat Club, especially Sarah Martin and Ros Cooper. Also, the fierce women that got us to Head of the River in Summer Eights 2019 and Head of the River in Summer Torpids 2021. And not to forget the "11 seconds"-Cambridge-beating Tethys four. Thanks also to friends from home who have supported me in various ways for a long time: Emma Efremova for reading my thesis, giving invaluable feedback and for all the study sessions we did together; Linda Pei and Max Langefeld for many fun and relaxing moments; Mareike Schröder and Rebekka Staufenberg for the best book club. Kathrin Müller and Annabelle Ziemek for long calls and visits; Melina Filzinger and Nils Fischer for great talks.

Finally, my family. This thesis would not exist if it weren't for my parents' support and encouragement to keep challenging myself. I am looking forward to many more celebrations with them.

Institutional

This work was supported by funding from the Engineering and Physical Sciences Research Council (EPSRC) and the Medical Research Council (MRC) [grant number EP/L016044/1]. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number203141/Z/16/Z. The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health.

Abstract

Epigenetics underpins the regulation of most genes, including many known to play a key role in the adaptive and innate immune system (AIIS). In this thesis, I developed a method called EpiNN that leverages epigenetic data to perform supervised gene classification with the goal of detecting AIIS-relevant regions of the genome. I trained EpiNN using a handcrafted list of genes with immune cell-specific epigenetic signatures and used it to scan the human genome for new AIIS-relevant regions. I detected 2,765 putative AIIS loci, of which 1,571 are uncharacterised by Gene Ontology (GO). I used stratified linkage disequilibrium (LD) score regression coupled with summary statistics from genome-wide association studies (GWAS) for 176 traits (average N=262k) to test whether the AIIS annotation is predictive of regional heritability for these traits. I detected significant heritability effects (average $|\tau^*| = 1.65$, $p < 0.05/176$) for 20 out of 26 immune-relevant traits. In a meta-analysis of 63 independent heritable traits, immune-relevant traits and diseases were 4.45× (SE $0.09, p = 9 \times 10^{-92}$) more enriched for heritability than other traits. The EpiNN annotation was also among the most significantly depleted in squared trans-ancestry genetic correlation in S-LDXR analysis of GWAS summary statistics from East Asian and European populations, indicating ancestry-specific effects. I analysed several of the predicted genes and found novel AIIS-specific enhancers and a novel AIIS-specific transcription start site in *DNMT1*, which was confirmed experimentally using PCR and Western Blot analyses. These results underscore the promise of leveraging supervised learning algorithms and large epigenetic data sets to detect genomic regions implicated in specific classes of heritable traits and diseases.

Contents

List of Figures viii

List of Tables ix

List of Abbreviations x

1 Motivation 1

2 Review of relevant biology and methodology 4

2.1 The immune system . . . 5

2.1.1 Innate immune system . . . 5

2.1.2 Adaptive immune system . . . 7

2.1.3 Immune-related phenotypes and diseases . . . 9

2.1.4 Genetic architecture of the immune system and common autoimmune diseases . . . 11

2.2 Epigenetics . . . 12

2.2.1 Chromatin Marks, ChromHMM, Epigenome Roadmap . . . 13

2.2.2 microRNAs and lncRNAs in the human immune system . . . 15

2.3 Finding immune system-relevant loci . . . 17

2.4 Classification Tasks . . . 19

2.4.1 Convolutional Neural Networks . . . 19

2.4.2 Logistic Regression . . . 21

2.4.3 Evaluating classification models . . . 22

2.5 LD Score Regression (LDSC) . . . 24

2.5.1 LDSC derivation . . . 24

2.5.2 Stratified LD Score Regression (S-LDSC) . . . 27

2.5.3 S-LDSC evaluation metrics . . . 29

3 Materials & Methods 30

3.1 Encoding of epigenetic data . . . 31

3.2 EpiNN model . . . 33

3.2.1 Choice of parameters . . . 34

3.2.2 EpiNN training . . . 34

3.2.3 Variations of data encoding and model architecture 36
3.2.4 Prediction of novel AIIS regions: genome-wide EpiNN annotation 37
3.3 Additional immune system-specific annotations 38
3.3.1 Immune system-specific histone mark annotations 38
3.3.2 Immune system-specific ChromHMM annotation 39
3.4 S-LDSC analysis 39
3.4.1 S-LDXR 41
3.5 Panther GO term analysis 41
3.6 Experimental analysis of *DNMT1* 42
3.6.1 CAGEseq data 42
3.6.2 Human peripheral blood samples 43
3.6.3 Immunoblotting 43
3.6.4 Polymerase chain reaction (PCR) 44

4 EpiNN analysis of immune-related traits and diseases 46
4.1 EpiNN training results 46
4.2 Genome-wide scan for novel AIIS-relevant regions 49
4.3 GO term analysis 49
4.4 S-LDSC analysis of individual traits 50
4.5 S-LDSC meta analysis 54
4.5.1 Enrichment of heritability in EpiNN quantiles 55
4.6 S-LDSC analysis including additional immune-related annotations . 56
4.6.1 S-LDSC joint analysis with underlying epigenetic annotations 57
4.7 S-LDXR analysis 61
4.8 Application of EpiNN to non-immune system genes 63
4.8.1 Housekeeping genes 63
4.8.2 Epithelial genes 63
4.8.3 Brain genes 65
4.9 Testing the AIIS EpiNN annotation for heritability enrichment in COVID19 infections 66

5 Biological Analysis of EpiNN Results 68
5.1 Protein-coding gene examples 69
5.1.1 *PIK3CD* 69
5.1.2 *IL2RB* 69
5.1.3 *TAGAP* 72
5.2 long non-coding RNAs 73
5.2.1 *Linc00402* 73
5.3 microRNAs 74

5.4 *DNMT1* 74
5.4.1 CAGEseq and primary sequence analysis 77
5.4.2 Western Blot 77
5.4.3 PCR 79
5.4.4 DNMT1 structure analysis 80
5.4.5 *DNMT1* conclusions 82
5.5 Additional EpiNN detections 82

6 Discussion **84**
6.1 Summary 84
6.2 Implications for future work 85
6.3 Limitations and future work 87

Appendices

A Appendix **89**
A.1 Applying S-LDSC to evolutionary annotations: CoalNN 89

B Appendix **93**

References **112**

List of Figures

1.1 Graphical overview of the EpiNN algorithm 3

3.1 CNN architecture graphical overview 33

4.1 Precision-recall curve of EpiNN, EpiNN-1D and EpiNN-binary . . . 47
4.2 S-LDSC analysis of EpiNN heritability enrichment and effect size . 51
4.3 S-LDSC analysis of 63 independent traits 53
4.4 S-LDSC meta analysis 54
4.5 S-LDSC meta-analysed quantiles 55
4.6 Joint S-LDSC analysis 56
4.7 Correlation between S-LDSC annotations. 57
4.8 Annotation Histograms 60
4.9 S-LDXR analysis results 62
4.10 S-LDSC on COVID19 data 67

5.1 Karyogram of training genes and predicted genes 70
5.2 Example genes predicted by EPiNN 71
5.3 In silico AIIS-relevance for the *DNMT1* gene 75
5.4 Experimental evidence of AIIS-relevance for the *DNMT1* gene and proposed mechanism 76
5.5 Structure for truncated DNMT1 predicted by Alphafold2 80

A.1 S-LDSC effect sizess for CoalNN 90

B.1 DNMT1 Western Blot replicates 95
B.2 *DNMT1* PCR replicate 95

List of Tables

2.1 Confusion matrix 22

3.1 Performance and hyperparameters of EpiNN models 36
3.2 Overview of best EpiNN models with hyperparameters 36
3.3 Image length analysis 37
3.4 Encodings used for EpiNN input data 38
3.5 Overview of S-LDSC annotations 39

4.1 EpiNN confusion matrix 48
4.2 GO Term analysis 50
4.3 Correlations between S-LDSC annotations. 58
4.4 Overlap between binarised annotations. 59
4.5 S-LDXR immune/blood traits 61

5.1 *DNMT1* CAGEseq data 78
5.2 Western Blot antibodies 79

B.1 List of EpiNN training genes 96
B.2 List of 63 complex traits 104
B.3 List of S-LDSC results for 63 complex traits 105
B.4 List of S-LDSC results for 176 complex traits 106
B.5 S-LDXR results 109
B.6 *DNMT1* CAGEseq raw data 110

List of Abbreviations

3'-UTR	3' untranslated region
aa	amino acid
AIIS	adaptive and innate immune system
AUPRC	area under the precision-recall curve
bp(s)	base pair(s)
CAGEseq	cap analysis gene expression sequencing
ChromHMM	Algorithm from [1]
CNN	Convolutional Neural Network
DNMT1	gene/protein symbol of DNA (cytosine-5)-methyltransferase 1
DPE	downstream promoter element
EpiNN	The here presented model to identify AIIS-relevant regions in the human genome
EpiNN-1D	Variation of EpiNN model
EpiNN-binary	Variation of EpiNN model
EpiNN-logistic	Variation of EpiNN model
GWAS	genome-wide association study
GO	Gene Ontology
HLA	human leukocyte antigen
hg38	Homo sapiens (human) genome assembly GRCh38
IRIS	List of immune system-relevant genes from [2]
ImmunoGene	List of immune system-relevant genes from [3]
IS, nIS	immune system, non-immune system
kb	kilo bases
LD	linkage disequilibrium
(S-)LDSC	(Stratified) Linkage-disequilibrium score (regression analysis)

(S-)LDXR . . (Stratified) trans-ethnic genetic correlation

LINC . . . . . Long Intergenic Non-Protein Coding

LOC (gene) . symbol for genes of uncertain function

MAF . . . . . minor allele frequency

MIR . . . . . microRNA

MHC . . . . . major histocompatibility complex

mRNA . . . . messenger RNA

PCR . . . . . polymerase chain reaction

RA . . . . . rheumatoid arthritis

RGB . . . . . Red-Green-Blue colour space

Roadmap . . . Epigenome Roadmap project: Database of epigenetic data [4]

SCID . . . . . severe combined immunodeficiency

SD . . . . . standard deviation

SE . . . . . standard error

SNP . . . . . single-nucleotide polymorphisms

TF . . . . . transcription factor

TSS . . . . . transcription start site

Tx . . . . . transcription

1 Motivation

Adaptive and innate immune system-relevant genes (AIIS-relevant genes) are at the core of the human immune system and dictate the course and therapeutic opportunities of many traits and diseases such as rheumatoid arthritis [5, 6], Crohn's Disease [7], and cancer [8]. The availability of increasingly large genomic data sets has enabled association analysis of complex autoimmune traits (e.g. [9, 10]), shedding light on their genetic architecture. However, despite this progress, currently known common susceptibility loci, including the Human Leukocyte Antigen (HLA) region [11], explain only between 15% [10, 12] and 36% [13] of the observed phenotypic variance in autoimmune traits. Although these efforts have, in some cases, led to the fine-mapping of underlying causal variations [14], the genetic mechanisms that trigger autoimmunity remain largely unknown [15]. Notably, of the detected candidate causative variants for autoimmune diseases, around 90% are estimated to be non-protein-coding. Of these, only a small fraction alter recognisable transcription factor sequence motifs [14]. This highlights the likely importance of regulatory variation in the genetic architecture of immune-relevant traits.

Current methods for gene classification are predominantly based on messenger RNA (mRNA) data. While expression data is cost-effective, it may be affected by temporal changes during the cell cycle or by environmental stimuli [16]. It is also not uncommon for mRNA levels to differ significantly from the actual protein levels,

with either mRNA not always being translated or mRNA of long-lived proteins already having degraded [17]. Moreover, proteins that are highly impactful but numerically few per cell, such as many transcription factors, generally have very few mRNA transcripts, therefore often falling below the detection threshold [18]. While proteome analysis overcomes some of these limitations, it is more expensive and reliant on antibody binding, which is not always as specific or effective as hoped [19]. Fortunately, epigenetic data can mediate some of these challenges and has further even been found to better differentiate between cell types than expression levels [20].

Since epigenetic mechanisms are known to play an important role in regulating the human immune system [21] and have been linked to susceptibility to autoimmune diseases [22], analysis of epigenetic data provides a valuable alternative route to identifying AIIS-relevant genes. Studying epigenetic mechanisms has provided insights into macrophage diversity and plasticity [21] and generally reflects cell identity better than expression data [20]. In this work, I hypothesise that the classification of AIIS genes based on epigenetic data might yield new AIIS regions in the genome that have not been detected by expression or proteomics data analysis. Based on this hypothesis, I aim to use epigenetic data to identify a set of AIIS genes and to determine whether epigenetically-identifiable AIIS regions differ from transcriptionally-identifiable AIIS regions.

To this end, I introduce a new approach called EpiNN (see fig. 1.1), which uses a neural network trained on epigenetic data encoded into images (for human readability) to recognise genomic regions likely to have AIIS-relevant functionality. I show that training the EpiNN neural network on publicly available epigenetic data of different human cells and tissues [23] achieves high testing accuracy in identifying held-out AIIS genes. I then apply EpiNN using a sliding window approach to detect novel AIIS-relevant regions along the human genome. I validate novel regions reported by EpiNN using GO term analysis [24] and stratified linkage disequilibrium score (S-LDSC) regression analysis [25–27]. I find that the genomic regions implicated by EpiNN are highly enriched for GO terms with immune system function and for heritability in immune system-relevant phenotypes. I compare

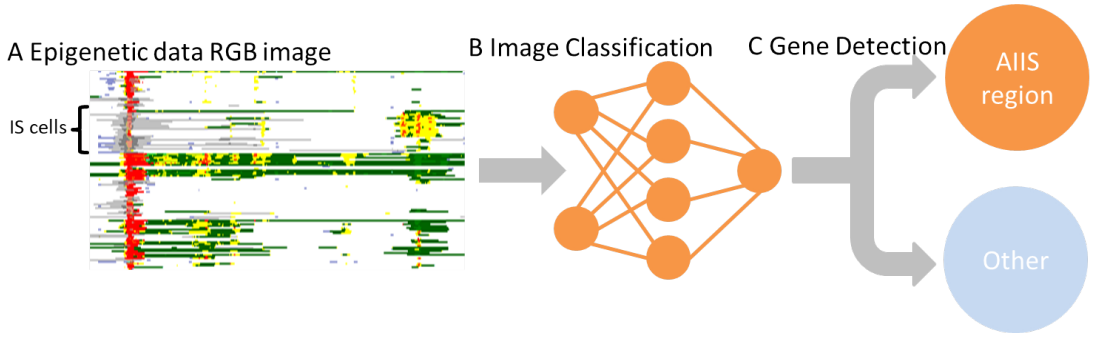


Figure 1.1: Graphical overview of the EpiNN algorithm. A: The model represents epigenetic data from the Roadmap data set [4] as an RGB (Red-Green-Blue) image. Colours encode 15 epigenetic states output by ChromHMM [1]; rows represent 111 different cell and tissue types; columns represent base pair positions along the genome. B: The model uses images of known immune system genes to train a CNN. C: The model outputs a score reflecting the likelihood of a region being AIIS-relevant. This score is used to perform binary classification of a gene/region or build genome-wide annotations of AIIS-relevance.

informativeness for immune system (IS) and non-immune system traits (nIS) traits and analyse the informativeness of the epigenetic-based EpiNN model when considering other models based on expression data and underlying epigenetic annotations. I perform functional validation for one newly discovered AIIS region, confirming its AIIS relevance as a putative novel immune system-specific transcription start site on the *DNMT1* gene. Collectively, I show that the EpiNN model complements existing methods by providing an additional projection angle of a complex cellular process and may be leveraged to gain new insights into the immune system-relevant regions of the human epigenome.

2

Review of relevant biology and methodology

Contents

2.1 The immune system 5
2.1.1 Innate immune system 5
2.1.2 Adaptive immune system 7
2.1.3 Immune-related phenotypes and diseases 9
2.1.4 Genetic architecture of the immune system and common autoimmune diseases 11
2.2 Epigenetics 12
2.2.1 Chromatin Marks, ChromHMM, Epigenome Roadmap . 13
2.2.2 microRNAs and lncRNAs in the human immune system 15
2.3 Finding immune system-relevant loci 17
2.4 Classification Tasks 19
2.4.1 Convolutional Neural Networks 19
2.4.2 Logistic Regression 21
2.4.3 Evaluating classification models 22
2.5 LD Score Regression (LDSC) 24
2.5.1 LDSC derivation 24
2.5.2 Stratified LD Score Regression (S-LDSC) 27
2.5.3 S-LDSC evaluation metrics 29

In this chapter I briefly introduce the relevant biology that guided the selection of AIIS relevant tissues and phenotypes, and epigenetic data sources.

2.1 The immune system

The human immune system protects us from pathogens and is involved in cancer clearing [28], wound healing [29], and pain signalling [30]. It is a complex dynamic network composed of organs, tissues, cells, molecules and soluble factors that interact in constantly changing processes.

Three lines of defence mechanisms protect our body against external invaders: The physical barrier, the innate immune system, and the adaptive immune system [31]. The physical barrier is comprised of epithelial cells and gland secretions such as stomach acid which serve to keep microorganisms from entering the body in the first instance. In the case of pathogens overcoming the first barrier, the innate immune system quickly mounts immediate but non-specific responses to the invading pathogens and activates the adaptive immune system [32]. The adaptive immune system recognises foreign antigens, generates a specific and appropriate immune response to each threat and creates an immunological memory of the encountered pathogen [33].

In this work, I combine the signals of multiple cell types and tissues that are part of the immune system to identify loci important for its regulation. I will refer to these genetic loci as adaptive and innate immune system relevant (AIIS relevant). The following section will, therefore, briefly review the different components of the human immune system with respect to AIIS relevant tissues.

2.1.1 Innate immune system

The innate immune system provides immediate protection against microbial invasion. It mediates the general immune responses, such as inflammation, the complement system, and non-specific cellular responses. It is either activated by the presence of foreign antigens, or cytokine and chemokine signalling by infected cells. These signals activate an inflammatory response, which increases blood flow to transport immune cells to the site of an infection.

The leukocytes of the innate immune system can be categorised into granulocytes (neutrophiles, basolphiles and eosinophiles), phagocytes (macrophages and dendritic

cells), mast cells, and natural killer cells. These cells travel in the circulatory system or are embedded in tissues.

Phagocytes circulate through the body, looking for potential threats, which they engulf and destroy. Macrophages are efficient phagocytic cells that can leave the circulatory system and invade other tissues. They can also release cytokines to recruit other cells to an area with pathogens. Mast cells are found in connective tissues and are involved in wound healing. When activated, they release cytokines, granules and mediators , such as histamine, which causes blood vessels to dilate, increasing blood flow and cell trafficking to the area of infection.

Neutrophils, of which the bone marrow of an average healthy adult makes approximately 100 billion new cells per day [34], are typically the first cells to arrive at an infection site due to their sheer number. Their cytoplasm contains granules which are toxic to invaders. Another granulocytes, eosinophils, target multi-cellular parasites by secreting highly toxic proteins and free radicals. These toxins also cause tissue damage during allergic reactions, which is why activation of eosinophils is highly regulated to prevent unnecessary tissue damage. In contrast, natural killer cells do not attack pathogens directly but destroy infected host cells to stop the spread of infection. Infected or compromised host cells can signal to natural killer cells for destruction by expressing specific receptors and by antigen presentation.

Cells of the innate immune system are supported by the complement system [35], which facilitates the search for and removal of antigens by agglomerating them. Through the process called opsonisation, complement factors tag infected cells and identify circulating pathogens expressing the same antigens. Next, macrophages and neutrophiles undergo chemotaxis to sites of infection via the action of cytokines. Lysis is then induced by proteins of the complement system by puncturing the membranes of foreign cells. Finally, agglutination of antibodies cluster and bind pathogens together for better access for immune cells.

2.1.2 Adaptive immune system

The adaptive immune system reacts more slowly and provides a specialised defence against infections. Unlike the innate immune system, it can retain "immune memory". The two types of adaptive immunity, humoral immunity and cell-mediated immunity, are mediated by different lymphocytes which recognise foreign substances with their highly diverse and variable receptors.

Humoral immunity is mediated by B lymphocytes which produce antibodies that defend against microbes by preventing them from invading tissue cells. Antibodies can also enhance the uptake of microbes living outside cells into phagocytes which kill microbes by ingestion.

Cell-mediated immunity is mediated by T lymphocytes and targets microbes that have already entered host cells. T lymphocytes have various pathways to fight against microbes: Some T lymphocytes activate phagocytes to destroy pathogens living within intracellular vesicles. Other T lymphocytes kill any host cell that harbours infectious microbes in the cytoplasm or nucleus. In both cases, the T cells recognise microbial antigens displayed on host cell surfaces.

Adaptive immune system lymphocytes

T lymphocytes develop and mature in the thymus gland. They can be distinguished from other lymphocytes by the presence of a T-cell receptor on the cell surface. All T cells originate from haematopoietic stem cells in the bone marrow. The antigen receptors of most T lymphocytes recognise only peptide fragments of protein antigens bound to specialised peptide display molecules, the major histocompatibility complex (MHC) molecules, on the surface of specialised cells, the antigen-presenting cells.

B lymphocytes express membrane-bound antibodies that serve as the receptors recognising foreign antigens. If an antibody receptor binds to a soluble or surface-bound antigen, the proliferation and differentiation of the antigen-specific B cells

is induced. B cells can contribute to autoimmune diseases by secretion of auto-antibodies, presentation of auto-antigens, secretion of inflammatory cytokines, or modulation of antigen processing and presentation [36].

When naive lymphocytes recognise microbial antigens, the antigen-specific lymphocytes proliferate and differentiate into effector cells and memory cells. A naive T lymphocyte has differentiated in the bone marrow and successfully undergone positive and negative central selection in the thymus but has not yet encountered its cognate antigen. They reside in and circulate between peripheral lymphoid organs and survive for several months and up to a few years, ready to find and respond to antigens. If not activated by an antigen, naive lymphocytes undergo apoptosis. Upon recognition of an antigen, naive lymphocytes can differentiate into effector lymphocytes that can produce molecules with the function to eliminate antigens.

Tissues of the immune system

Together with leukocytes, the primary lymphoid organs (bone marrow and thymus), which generate and mature T and B lymphocytes, are considered tissues of the immune system. In addition, there are secondary or peripheral lymphoid organs, in which adaptive immune responses to microbes are initiated (e.g. lymph nodes, spleen, and certain membrane layers in the bowel).

The **thymus** is enriched with T cells, as this is the primary site of T cell activation. The **bone marrow** also contains an immune microenvironment.

Lymph nodes are located along lymphatic channels throughout the body. Fluid (lymph) constantly leaks out of small blood vessels and is transported by lymphatic vessels to the lymph nodes and later back into the blood circulation. Lymph contains a mixture of substances absorbed from epithelia and other tissues. Antibody-presenting cells in the lymph nodes can sample the antigens of microbes in the lymph. In addition, dendritic cells sample antigens of microbes from the skin and other tissues and transport these antigens to the lymph nodes. Therefore, antigens of microbes entering through epithelia are accumulated in draining lymph nodes.

Blood entering the **spleen** is checked for blood-borne antigens, which are captured and accumulated by dendritic cells and macrophages. The spleen contains abundant numbers of phagocytes, which ingest and destroy microbes in the blood.

2.1.3 Immune-related phenotypes and diseases

Infectious diseases

Infectious diseases are among the leading causes of death globally, according to the WHO (2020 data, [37]). They are caused by a wide range of pathogens, including viruses, bacteria, parasites, and fungi.

The immune system can overcome most infections without medical intervention. However, in some cases, an infection can overwhelm the immune system and cause severe illness or death. The chance of that happening is particularly severe for immuno-suppressed patients or when dealing with novel viruses that the immune system has not previously seen, such as SARS-Cov-2. There are many strategies available to help treat and prevent infectious diseases e.g. antibiotics, anti-fungals and vaccination. But for many infectious diseases the treatment is variably efficacious, or not available.

Cancer

Cancer can severely affect the immune system. Cancers that spread into the bone marrow, such as leukaemia or lymphoma, can reduce the number of immune system cells in the blood and make patients more susceptible to infections.

On the other hand, the immune system function can be enhanced in order to target cancer cells: Cancer immunotherapy is a treatment field that uses activated immune system cells to recognise and destroy cancer cells [38, 39]. Various immunotherapy types use different mechanisms like cytokines [40], vaccines [41], monoclonal antibodies [42], viral therapies [43], adoptive cell transfer [44], and immune checkpoint inhibitors [45].

Immunodeficiency

Immunodeficiency can be inherited (primary immunodeficiencies) or, much more prevalently, acquired environmentally (secondary immunodeficiencies), such as developing AIDS through an infection with HIV.

An example of an immunodeficiency with a genetic cause is severe combined immunodeficiency (SCID) [46]. X-linked SCID is caused by mutations in the *IL2RG* gene which affects the function of both T and B cells. More than a dozen genes have been implicated in other types of SCID [47].

Allergies

Allergies arise if the immune system overreacts to foreign molecules or substances that are usually harmless. Allergies may affect only specific tissues, such as the skin, or in more harmful cases might cause systemic reactions. Allergic diseases are associated with IgE molecules which bind specifically to high-affinity IgE receptors on mast cells and mediate the release of cytokines on exposure to an allergen [48].

Autoimmune diseases

Autoimmune diseases are linked to the same loss of tolerance as allergies but are directed against self-antigens instead of external substances.

In recent years, numerous genetic mutations have been associated with autoimmune diseases. In some cases, they are caused by the mutation of a single gene, such as the transcription factor (TF) *AIRE* involved in the central tolerance mechanism (autoreactive lymphocyte clones usually being deleted before they develop into fully immunocompetent cells) [49], but most are polygenic [50].

Even with extensive GWAS studies, only fractions of the total disease heritability are currently explained for most autoimmune diseases. For instance, genome-wide loci identified by GWAS only explain 19.5% of the disease variance or 39% of disease heritability in rheumatoid arthritis (RA) [51]. Another example is hypothyroidism for which GWAS is able to explain between 2.3% and 20% in different subtypes [52, 53].

2.1.4 Genetic architecture of the immune system and common autoimmune diseases

Most autoimmune diseases are complex traits with multiple genes contributing to the phenotype. While polygenic signal is found genome-wide, the major histocompatibility complex (MHC) usually harbours several genes with large effects for these traits. The MHC is a group of genes that encode a variety of cell surface markers, antigen-presenting molecules, and other proteins involved in immune function, including the human leukocyte antigen (HLA) genes. The MHC region lies on the short arm of chromosome six at position 6p21.3. The classical MHC spans 3.6 megabases (Mb) and comprises more than 200 genes, including both various immune system genes as well as genes without any known immune function. An extended MHC, spanning about 7.6 Mb of the human genome, has been proposed. The extended MHC is suggested to encompass 421 genes of which around 20% have a putative immuno-regulatory role and might be associated with autoimmune diseases [54]. There seems to be evidence that some MHC alleles might enhance the autoimmune response [55] by influencing the expressed repertoire of T cells and consequently contributing to autoimmune phenotypes. Indeed, most autoimmune diseases have been associated with MHC alleles. Identification of disease-causing genetic variants within MHC has been difficult due to extensive linkage disequilibrium (LD) and high variability among the MHC alleles, therefore, the MHC is excluded from most GWAS.

The MHC region alone can explain between 2% and 30% of variance in autoimmune diseases. On the low end, it only explains around 2% of Systemic Lupus Erythematosus [56] but about 29.6% of phenotypic variance of Type 1 Diabetes [57].

GWAS results for several autoimmune diseases have also found plenty of disease-associated regions outside the MHC region. Some of the GWAS-SNPs are in protein-coding genes and have been identified to be non-synonymous amino acid changes, such as *PTPN22* for Type 1 Diabetes [58] and rheumatoid arthritis [59], *NOD2* and *ATG16L1* for Crohn's disease [60, 61], and *IL23R* for Crohn's disease and psoriasis [62, 63]. Interestingly, there are no amino acid changes for many genes

strongly implicated in complex genetic diseases, including *STAT4*, a transcription factor with SNPs associated with lupus and rheumatoid arthritis [64].

In line with these results, nearly 90% of the detected candidate causative variants for autoimmune diseases are estimated to be non-protein-coding, and with only a fraction of causal non-protein-coding variants altering recognisable transcription factor sequence motifs [14], which highlights the likely importance of regulatory variation in the genetic architecture of immune-relevant traits.

2.2 Epigenetics

While most cells in the human body contain the same DNA sequence, it is interpreted very differently in different cell types, leading to the vast variety of tissue types and functions. The changes in phenotype without underlying changes in genotype may be explained by epigenetic regulation. As defined by Medawar and Medawar, this field encompasses "... all the processes that go into implementing the genetic instructions contained within the fertilised egg. Genetics proposes; epigenetics disposes" [65]. At the molecular level, the epigenetic modulation of gene expression is accomplished through various structural and chemical modifications of the DNA in the nucleosome [66, 67].

Due to the tissue-specificity of epigenetic marks, they are helpful in studying differences between cell types. Unlike transcription-based data, which is also tissue-specific, epigenetic marks do not only cover RNA- or protein-coding genes but are present genome-wide, giving an indication of (tissue-specific) enhancers, transcription start sites or heterochromatin. Indeed, it has been found that epigenetic signals can differentiate between cell types better than expression levels are able to [20]. Epigenetic patterns can vary over time and have been linked to cell differentiation [68]. Epigenetic programming also ensures cell plasticity by inducing cell stage changes, thereby enabling the same cell to quickly switch between functions and gene transcription programs. For example, studying epigenetic mechanisms in macrophages has provided insights into sub-cell type diversity and plasticity [21].

This plasticity, specifically the ability to change between different transcription programs, is one hallmark of immune system cells. They have to react to invaders and activate specialised responses instantly. Thus, the study of epigenetics is particularly interesting in immune system cells with respect to their plasticity and cell differentiation. It has already been shown that epigenetics plays an essential role in regulating the human immune system [21] and has also been linked to susceptibility to autoimmune diseases [22]. These discoveries have led to novel drug discovery ideas: By 2018, seven drugs exploiting epigenetic modifications had received regulatory approval, and numerous other candidates were in clinical trials [69].

Epigenetic regulation takes effect in different ways, such as through DNA methylation, histone modifications, and non-coding RNAs (e.g. lncRNAs or microRNAs). The first, DNA methylation, typically occurs at base pair level, when DNA methyltransferases transfer a methyl group to the pyrimidine C5 position of cytosine residues [70]. The second, histone modifications, affect the approximately 200 base pairs wrapped around one histone octamer [71]. Histone modifications modulate chromatin structure and therefore affect processes including DNA transcription and repair. The third, non-coding RNAs, are a class of genes transcribed into RNA but not translated into proteins which have a regulatory function on protein-coding genes [72].

2.2.1 Chromatin Marks, ChromHMM, Epigenome Roadmap

The method I describe in this work is based on chromatin markers. Five markers (H3K4me3, H3K4me1, H3K36me3, H3K27me3, and H3K9me3) are used as input data for the model, and one (H3K27ac) is later evaluated as a consistency check.

H3K4me1 marks the mono-methylation at the fourth lysine residue of the histone H3 protein and is enriched at active and non-active enhancers [73]. H3K4me3 marks the tri-methylation at the fourth lysine residue of the histone H3 protein and is highly enriched at active promoters near transcription start sites [74]. H3K9me3 marks the tri-methylation at the ninth lysine residue of the histone H3 protein and is associated with heterochromatin and, therefore, the downregulation of gene

expression [75]. H3K27me3 marks the tri-methylation of lysine 27 on histone H3 protein and is also associated with the down-regulation of nearby genes. Unlike H3K9me3 which prevents multiple TFs from binding and silences genes across all cell types, H3K27me3 allows the binding of certain TFs, allowing activation in certain celltype-specific states [76]. H3K36me3 marks the tri-methylation at the 36th lysine residue of the histone H3 protein, is enriched within the borders gene bodies, and may be involved in defining exons [77]. H3K27ac marks the acetylation of the lysine residue at N-terminal position 27 of the histone H3 protein and is associated with active enhancers but unlike H3K4me1, not non-active enhancers, therefore only marking enhancers actually used [78].

There are a number of platforms available allowing the interrogation of epigenetic data: **ChromHMM** [1] is a computational model that uses combinations of chromatin modification patterns to learn a genome-wide annotation of the underlying epigenetic states at each genomic position. It is based on a multivariate hidden Markov model that models the observed combination of 5 chromatin marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3) and assigns one of 15 epigenetic states, such as enhancer, transcription start site or heterochromatin. Epigenetic states are analysed at 200 bp intervals that roughly approximate the number of base pairs wrapped around a histone octamer. ChromHMM data is publicly available on the NIH Roadmap Epigenomics Web Portal [79].

Epigenome Roadmap [4] is a collection of 111 reference human epigenomes with profiles for histone modifications, DNA accessibility, DNA methylation, RNA expression, and ChromHMM annotations. The 111 epigenomes represent human primary tissues and cell types representative of all major lineages in the human body. All data is publicly available at different sites, such as the WashU Epigenome Browser [80].

2.2.2 microRNAs and lncRNAs in the human immune system

lncRNAs

Immune system cells have to adapt rapidly to incoming threats with the use of carefully coordinated mechanisms involving cell differentiation or dynamic changes in gene expression. Recently, long non-coding RNAs (lncRNAs) have been linked to the coordination of immune cell lineage development and the control of dynamic gene expression changes, a hallmark of immune cell activation [81]. Altered expression of lncRNAs has also been found in several immune-related diseases such as inflammatory bowel disease, diabetes type 1 and multiple sclerosis [82].

lncRNAs are non-protein coding transcripts that function as RNA molecules and can interact with RNA, DNA or proteins to promote or restrain the expression of protein-coding genes. This multitude of interaction methods results in the need to investigate the biological function of any lncRNA on a case-by-case basis. lncRNAs are distinguished from smaller non-coding RNAs, like microRNAs, by a minimal length of at least 200 bp. They are known to coordinate diverse aspects of cell and tissue development in various cell types [83]. The current version of GENCODE Release 41 reports the number of known lncRNAs to be 19095 out of 61852 known genes, which is the same order of magnitude as the number of reported protein-coding genes (19370) [84]. However, most lncRNAs are only expressed in specific cell types and at lower levels than protein-coding genes, which makes them harder to detect with transcription-based methods.

Two examples of lncRNA activity in the immune system are the control of the development of CD4+ TH1 cells and the regulation of transcriptional programs in innate immune cells: The lncRNA lincRNA-MAF-4 coordinates the development of TH1 cells in humans through epigenetic silencing of the transcription factor c-MAF4 [85]. lincRNA-MAF-4 is located around 140 kb upstream of the gene encoding c-MAF, a transcription factor involved in differentiating naïve CD4+ T cells to TH2 [86] and TH17 cells [87]. Possible pathways leading to MAF expression silencing are either a 3D contact between lincRNA-MAF-4 and MAF or

an interaction betwee lincRNA-MAF-4 and the transcriptional repressors LSD1 and EZH2, which attach silencing H3K27me3 marks at the promotor of MAF in TH1 cells [81]. Some lncRNAs are involved in the immune response to the activation of the signalling pathway NF-κB upon cytokine receptor engagement in innate immune cells. They can induce different lncRNA-directed regulatory circuits, such as feed-forward (e.g. THRIL [88]), negative feedback (e.g. NKILA [89]), or release of suppression (e.g. lincRNA-EPS and lnc-13 [90]) loops.

microRNAs

In addition to lncRNAs, microRNAs have been identified as major regulators of gene expression [91].

MicroRNAs are a class of small (around 22 bp) non-coding RNA molecules, of which currently 1879 have been identified (GENCODE version 41 [84]). MicroRNA genes are often located either in intergenic regions and equipped with their own promoter structure or in introns and, therefore, primarily co-transcribed with their respective host gene [92]. MicroRNAs are transcribed and processed, enabling them to post-transcriptionally regulate gene expression by directly binding to the 3'UTR of mRNA transcripts [93, 94]. This interaction leads either to translational repression or degradation of the target mRNA [95].

For the immune system in particular, microRNA expression is connected to homeostasis as well as to disease states. MicroRNAs are involved in the development, lineage commitment, activation, and function of immune cells [96]. By targeting cellular signalling pathways (e.g. NF-κB), microRNAs have a regulatory impact on both innate and adaptive immune cells [91].

Interestingly, microRNAs can not only be released into the cytoplasm but also secreted into the inter-cellular space, where they can be taken up by other cells. Unfortunately, microRNAs also act as mediators of tumour immune escape through this very same mechanism: MicroRNAs secreted by cancer cells are taken up by immune cells, where they down-regulate anti-tumour immune response [97].

2.3 Finding immune system-relevant loci

Since the first sequencing of the human genome [98], researchers have tried to find genes regulating the human immune system. Twin studies have been used to estimate the heritability of phenotypes [13], but they have a limited sample size. The introduction of faster and cheaper sequencing technologies has enabled the construction of large-scale databases, which can be interrogated for correlations between genotype and phenotype.

Genome-wide association studies (GWAS) [99] use large collections of genomic data to find statistical associations between genomic variants such as single-nucleotide polymorphisms (SNPs) and phenotypes. GWAS rely on large sample sizes and on the presence of LD across genomic variants, which allows a set of observed variants to tag the effects of causal variation, which may not always be directly observed. GWAS has emerged as a powerful study design for the analysis of complex traits, which are highly polygenic and comprise a large number of small effects on phenotypes, allowing for a number of downstream applications [100] GWAS has also enabled association analyses of complex autoimmune traits (e.g. [9, 10]), shedding light on their genetic architecture. Despite this progress, currently known common susceptibility loci, including the Human Leukocyte Antigen (HLA) region [11], explain only between 15% [10, 12] and 36% [13] of the observed phenotypic variance in autoimmune traits. Similarly, identifying the genetic factors controlling variation in the immune system is still in its early discovery phase, with the overlap of loci reported in more than one study still limited [101]: GWAS for functionally relevant phenotypes of the adaptive immune system in healthy individuals could only explain 6%–20% of variance [102]. Although these efforts have, in some cases, led to the fine-mapping of underlying causal variations [14], the genetic mechanisms that trigger autoimmunity remain largely unknown [15]. In part that is because, as mentioned in chapter 1, due to GWAS results to 90% falling into non-protein-coding regions, and with only a fraction of these hits altering recognisable transcription factor sequence motifs [14]. This makes the interpretation difficult, especially without the integration of additional data sources.

Another popular method to study cell differences is to focus on the transcribed genes by sequencing and measuring mRNA levels [103]. Unlike genome sequencing, transcription data is cell type-specific and can be used to study transcriptional differences between immune and non-immune system cells. Transcription data has, for example, been used in Abbas et al [2] who identify immune system-relevant genes based on expression data, resulting in a list (Immune response in silico, "IRIS") consisting of 2272 genes. A second paper from 2013 [3] used expressed sequence tag profiling based on transcriptional data, resulting in a second list ("ImmunoGene") of 2,232 potential immune system-relevant genes.

The pros and cons of using transcription data or proteome data have been discussed in chapter 1. But the most important reason I chose not to use this data for EpiNN is that transcription and proteome data both only include transcribed genes but not regulatory features like enhancers. Since it is known from GWAS that around 90% of causal SNPs are non-protein coding [14], transcription data can only ever explain parts of phenotype variability.

Like transcriptomics and proteomics, epigenetic data is specific to cell type and is therefore helpful in finding differences between cell types. However, unlike transcriptomics and proteomics, epigenetic data is genome-wide, covering protein-coding and RNA-coding genes, but also enhancers, transcription start sites and down-regulatory elements like heterochromatin. Fig. 2g in [20] highlights how highly heterogeneous epigenetic signals are projected onto relatively uniform expression levels in immune system cells, visualising how epigenetic signals can differentiate between cell types more robustly than RNA expression levels [20].

Currently, epigenetic data is often used to fine-map GWAS results to overcome the limitations by LD [14]. Using this additional information on the regulatory elements of the locus in the cell type relevant for the disease, can point to the most promising variants to test for functional follow-up [104–107].

Another area of research is identifying (super-)enhancers that regulate multiple autoimmune diseases. For example, enhancer as well as super-enhancer marks

have been found to be enriched in cell types and loci relevant to autoimmune diseases [14, 108–110].

In this work, I leverage discoveries that have already been made by methods being applied to AIIS genes. My model uses supervised learning based on known sets of AIIS gene to find additional loci (genes, as well as regulatory regions) relevant to the immune system. In this setup, I do not need phenotypic information like GWAS, therefore I can use GWAS results as an orthogonal data set to test the prediction results of my method using stratified LD score regression.

2.4 Classification Tasks

The finding of novel AIIS-relevant regions requires solving a classification task for each genomic region, which means, deciding whether the region should be classified as AIIS-relevant or not. For sequential data, such as DNA sequences, there are multiple methods commonly used for classification, including, among others, logistic regression, hidden Markov models (HMMs) (e.g. [1]), recurrent neural networks (RNNs) and variations that improve performance for learning long-term dependencies such as long short-term memory networks (LSTMs), and convolutional neural networks (CNNs). In this work, I adopted CNNs, which have been shown to be effective in several genomic applications (e.g. [111, 112]). In CNNs, sequential data is analysed by sliding a window (called filter) over the sequential data.

2.4.1 Convolutional Neural Networks

CNNs are a type of artificial neural network that is widely used for different tasks, such as image classification, object detection or natural language processing [113]. They are particularly good at recognising complex, non-linear patterns in data [114].

CNNs are made up of interconnected nodes organised in layers with different functionalities. The first layer is typically a convolutional layer followed by a pooling layer, followed by more pairs of convolutional and pooling layers, which extract features and down-sample the input matrix. The deeper a CNN, the more pairs of

convolutional and pooling layers. A fully connected layer follows, which produces a vector of outputs, on which classification can be performed.

Convolution shifts two functions over each other and captures their amount of overlap as an output function. The convolutional layer in CNNs works by sliding a kernel (a smaller matrix than the input matrix), also called filter, over the input matrix and multiplying each matrix element in the kernel by the corresponding matrix element in the input image. The output image is then the sum of all the products. The kernel matrix elements are learned during the training phase.

Pooling is applied to reduce the size of the matrices without losing important information. Pooling works by dividing the matrix into smaller regions and then only keeping either the average or maximum value of all the matrix elements in each region, therefore reducing the matrix size.

Fully-connected layers are made up of several nodes, which are connected to every other node in the layer. The strengths of the connections are determined by weights and biases, which are learned during training. The output of a fully connected layer is a vector of values, one for each node in the layer.

Activation functions are used to introduce non-linearity into the CNN, which allows CNNs to detect complex patterns. Multiple non-linear functions can be used for this, such as Sigmoid, Softmax, tanH or ReLU, which is defined as $f(x) = \max(0, x)$.

The goal of **training CNNs** is to find the weights and kernel matrix entries of the CNN that minimise the loss function. For classification, the loss function is often defined as the cross-entropy between the predicted class labels and the ground truth class labels and minimised using backpropagation. Backpropagation uses the gradients of the loss function to update the weights of the CNN. The training process typically consists of the following steps: collection of labelled training and testing data; preprocessing the data (e.g. normalisation); initialising the weights of the model; training the model; evaluating the model performance.

Because of the large number of parameters in CNNs which can be higher than the number of training points available, CNNs have to include **regularisation** to

prevent **overfitting**. Regularisation can be achieved by adding a penalty to the loss function, such as L1 or L2 regularisation. In CNNs, overfitting is often prevented by the introduction of **dropout layers** in the architecture, where a dropout layer randomly drops a certain amount of nodes from the neural network.

Hyperparameters in CNNs

In addition to weights, several parameters in a CNN that can be tuned, including the network architecture. These hyperparameters include the number of convolutional layers (which determines the complexity of the features and functions a CNN can learn: a deeper network can learn more features but may be more difficult to train), the size of the convolution kernels (which determines the receptive field of the network: a larger receptive field can capture more context, but may be more difficult to train), learning rate (which is the rate is the step size used to update the weights of the CNN during training: a small learning rate converges slowly, while a large learning rate might overshoot the local minimum and not converge at all), regularisation parameters (such as the dropout rate of dropout layers, which determines how much regularisation is applied: allowing to tune bias and variance of the model)batch size (which is the number of training examples that are used in each iteration of training: a larger batch size uses more data in each training iteration but requires more memory) and number of epochs (which is the number of training iterations the models makes through the complete training data during the training phase: a large number might give the model the chance to overfit the training data while a small number may underfit.).

Hyperparameters are not independent of each other. They are therefore typically tuned jointly using cross-validation and the best combination of hyperparameters is chosen for the final model.

2.4.2 Logistic Regression

Logistic regression is a widely-used approach for binary classification, which is also commonly adopted in genomics (e.g. [115]). A logistic regression model is equivalent

to a neural network comprising a single neuron that is fully connected to each input feature. To apply logistic regression to matrix data, the matrices are first unravelled into one long vector, on which logistic regression can be applied.

The logistic function is defined as:

$$\sigma(z) = \frac{1}{1 + \exp(-z)}, \tag{2.1}$$

where $z = wx + b$, with weights w and bias b, which are learned to minimise a binary cross-entropy loss.

Logistic regression is usually regularised to prohibit overfitting. Common approaches include the addition of penalty terms to the loss function, such as the following ridge term:

$$L + \lambda \sum w^2, \tag{2.2}$$

where L is the loss function and λ a free parameter that can be tuned using cross-validation.

2.4.3 Evaluating classification models

Classification methods may be evaluated using a variety of metrics, including accuracy, precision, recall, and F1 score. These metrics are based on a confusion matrix (see table 2.1) that visualises correct and incorrect predictions.

		Model prediction	
	Total	Predicted Positive (PP)	Predicted Negative (PN)
Actual condition	Real Positive (P)	True Positive (TP)	False Negative (FN)
	Real Negative (N)	False Positive (FP)	True Negative (TN)

Table 2.1: Confusion matrix. Exemplary confusion matrix of a model.

The above metrics are defined as follows:

Accuracy measures the percentage of images that are correctly classified but is only well-defined for the testing set with similar numbers of data points in each category c:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Real Positives} + \text{Real Negatives}} \tag{2.3}$$

Precision measures the specificity of a model and is defined as:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{2.4}$$

Recall measures the sensitivity of a model and is defined as:

$$\text{Recall} = \frac{\text{True Positives}}{\text{Real Positives}} \tag{2.5}$$

F1 score is the harmonic mean of precision and recall.

The area under the precision-recall curve (AUPRC) is a useful metric combining precision and recall metrics information at different classification thresholds. As the name suggests, it calculates the area under a curve that plots precision against recall at multiple thresholds. The AUPRC ranges between 1 (best) and 0.

The binary cross entropy loss may be seen as a differentiable approximation to the binary classification loss, which penalises the differences between predicted class probabilities and true class labels. The loss function for two classes is defined as:

$$H_p = -(y \log(p) + (1 - y) \log(1 - p)) \tag{2.6}$$

where log is the natural logarithm, y is a binary indicator (0 or 1) of whether the class label c is the correct classification for observation o and p is the predicted probability for observation o being of class c.

Cross-validation may be used to perform model selection and determine the best set of hyperparameters. Cross-validation is an umbrella term for resampling methods, the most used one being k-fold cross-validation. In k-fold cross-validation, the data is split into k equal-sized subsets. The model is then trained on k-1 subsets and tested on the remaining subset. This process is repeated k times, with each subset used once as the testing set. The aggregated performance of each model on the held-out test sets is then used to select the best model.

2.5 LD Score Regression (LDSC)

In this work, to test whether regions predicted to be AIIS-relevant by my model are enriched for complex trait heritability, I used a framework called Stratified Linkage Disequilibrium Score Regression (S-LDSC) [25–27, 116].

S-LDSC assesses the contribution of a genomic annotation to disease and complex trait heritability, while conditioning on a set of additional annotations. This enables quantifying the degree to which a proposed annotation, such as the one produced by EpiNN, captures signal that has not been previously described using other annotations in the model [26]. To understand S-LDSC, I first introduce LDSC, which is then extended to S-LDSC by considering models comprising multiple annotations. At a high level, LDSC works by testing whether markers with large correlation to several other genomic variants (high "LD score") tend to yield stronger GWAS associations. This is likely to be caused by polygenicity of the trait, rather than inflation of summary statistics due to, for instance, population stratification [25]. One of the appeals of LDSC is that it is easily applied, as it only requires the availability of summary association statistics from a GWAS, as well as a reference panel for the computation of LD scores.

2.5.1 LDSC derivation

LDSC derivation starts by assuming a linear polygenic model for the dependence of phenotype y on genotype x_k with effect sizes β_k for SNP k with environmental effects ϵ:

$$y = \sum_k x_k \beta_k + \epsilon, \tag{2.7}$$

where the phenotype and genotype are standardised ($\mathrm{Var}(y) = 1$ and $\mathrm{Var}(x_k) = 1$) to simplify the derivation.

In practice, only estimates of marginal effect sizes $\hat{\beta}$ are available from GWAS, which I refer to as $\hat{\beta}_{GWAS,j}$ for SNP j here. These estimates capture the effects of several other variants, through LD, as well as potential population stratification:

$$\hat{\beta}_{GWAS,j} = \sum_k r_{jk} \beta_k + s_j + e_j \tag{2.8}$$

$\hat{\beta}_{GWAS}$ has three sources of variation: genetic effects $\sum_k r_{jk}\beta_k$, a bias term s_j that includes confounders such as population stratification or cryptic relatedness, and estimation error e_j, where r_{jk} is the correlation of SNP k to SNP j (due to LD).

This model makes a number of strong assumptions:

- β_k is assumed to be random with mean zero and variance of per-SNP heritability, therefore: $\mathbb{E}(\beta_k) = 0$ and $\mathrm{Var}(\beta_k) = \frac{h^2_{SNP}}{M}$. h^2_{SNP} is here the narrow sense heritability [117] - or the proportion of phenotypic variance explained by additive effects - explained by the SNPs in the sum and M being the total number of SNPs used to calculate h^2_{SNP}. This assumes, that each SNP approximately contributes equally to the heritability of a phenotype.

- For the term s_j it is assumed that $\mathbb{E}(s_j) = 0$ and it is also defined that $a \equiv \mathrm{Var}(s_j)$.

- The error term e_j is assumed to come from a normal distribution: $e_j \sim \mathcal{N}(0, \frac{\sigma^2_e}{N})$, where N is the GWAS sample size and σ^2_e the variance of the residuals. It is assumed that the variance explained by a given SNP is small, i.e. the residual variance should be close to the full variance of y, which is standardised to be 1: $\sigma^2_e \approx \mathrm{Var}(y) = 1$. Therefore $\mathbb{E}(e_j) = 0$ and $\mathrm{Var}(e_j) \approx \frac{1}{N}$.

- It is assumed that the effect size estimates are uncorrelated with the stratification, i.e. a SNP that has a large effect on some phenotype is not more or less likely to have a large amount of stratification in it: $\mathbb{E}(\beta_j s_k) = 0$ for any j, k.

- It is assumed that the effect sizes are uncorrelated to the estimation errors: $\mathbb{E}(\beta_j e_k) = 0$ for any j, k.

- It is assumed that all of the effect sizes are uncorrelated to each other. Knowing the effect of one SNP is not informative of another SNP: $\mathbb{E}(\beta_j \beta_k) = 0$ for any $j \neq k$.

- It is assumed that stratification and error are uncorrelated: $\mathbb{E}(s_j e_k) = 0$ for any j, k.

LDSC regression regresses the mean χ^2 statistic of variants in a quantile against the mean LD score of variants in a quantile bin. The χ^2 statistic test statistic for a SNP j in the GWAS is computed as the square of $\hat{\beta}_{GWAS,j}$ over their sampling variance:

$$\begin{aligned}\chi_j^2 &\equiv \frac{\hat{\beta}_{GWAS,j}^2}{\mathrm{Var}(e_j)} \\ &\approx \frac{\hat{\beta}_{GWAS,j}^2}{\frac{1}{N}} \\ &= N\hat{\beta}_{GWAS,j}^2\end{aligned} \tag{2.9}$$

Then the expected value of χ_j^2 is:

$$\begin{aligned}\mathbb{E}(\chi_j^2) &= \mathbb{E}\left[N\left(\sum_k r_{jk}\beta_k + s_j + e_j\right)^2\right] \\ &= N\Bigg(\mathbb{E}\left[\left(\sum_k r_{jk}\beta_k\right)^2 + 2s_j\sum_k r_{jk}\beta_k + \right. \\ &\quad \left. 2e_j\sum_k r_{jk}\beta_k + 2s_je_j + s_j^2 + e_j^2\right]\Bigg) \\ &= N\left(\sum_k r_{jk}^2\,\mathbb{E}(\beta_k^2) + \mathbb{E}(s_j^2) + \mathbb{E}(e_j^2)\right) \\ &= N\left(\mathrm{Var}(\beta_j)\sum_k r_{jk}^2 + \mathbb{E}(s_j^2) + \mathbb{E}(e_j^2)\right) \\ &= N\left(\frac{h_{SNP}^2}{M}\sum_k r_{jk}^2 + a + \frac{1}{N}\right) \\ &= \sum_k r_{jk}^2\frac{N}{M}h_{SNP}^2 + Na + 1 \\ &= h_{SNP}^2 l_j + Na + 1\end{aligned} \tag{2.10}$$

where, by the linearity of the expectation operator, the expectation for each term could be separately computed after expanding the square, and applying definitions and assumptions listed above, namely $\hat{\beta}_{GWAS,j}$, s_j, and e_j being uncorrelated. By defining $l_j = \sum_k r_{jk}^2\frac{N}{M}$, called the LD score, a linear form can be achieved. The LD score is the sum over all SNPs of the squared correlation coefficient between of SNP j and all other SNPs. In practice, $\mathbb{E}(\chi^2)$ and N will be given in GWAS summary statistics and l_j can be obtained from a reference data set (see below). When regressing the χ^2-statistic on the LD score, the slope will be the heritability

h^2 and the intercept $Na + 1$. Knowing N, the variance due to stratification can be estimated from this intercept. Therefore, LD score regression can unravel the different contributors to genetic effect sizes in GWAS to underlying factors of heritability and population stratification.

LD scores l_j can be estimated in reference samples with the same LD structure, e.g. for this project, European ancestry individuals in the 1000 Genome Project [118] (Phase3). This allows performing LD score regression without access to raw genetics data. One consequence of this is that the estimated h^2_{SNP} is the variance explained by the SNPs used to estimate the LD scores, which might differ from the set of SNPs in the GWAS summary statistics, i.e. more SNPs might be used to calculate l_j than in the regression because the whole panel of SNPs in the reference panel might not be available from the GWAS summary statistics. It is usually assumed that the heritability and intercept results from the SNPs in the regression extrapolate to the remaining SNPs that are used to calculate LD scores but were not in the regression, i.e. that l_j is not biased despite being calculated on different SNPs than $\mathbb{E}(\chi^2_j)$. It is therefore safer to restrict allele frequency (MAF $> 5\%$) for the calculation of LD scores and exclude strong LD regions. This is because the 1000 Genome Project contains multiple rare variants, while GWAS summary statistics usually do not contain many rare variants and it cannot be assumed that the effect sizes from common variant GWAS extrapolate well to rare variants in the reference panels. This extrapolation also forbids the use of targeted GWAS chips that are enriched for regions with higher effect for a certain phenotype (e.g. metabolism-chip, immuno-chip), since there can be no generalisation from enriched regions to the whole genome. Therefore, only representative GWAS can be used.

2.5.2 Stratified LD Score Regression (S-LDSC)

S-LDSC was initially developed to estimate heritability enrichment using binary annotations by Finucane et al [26] and extended to allow working with continuous annotations by Gazal et al [27]. It has been widely applied to dissect the contribution

of several functional and evolutionary annotations to the heritability of complex traits e.g. [119] and to identify disease-relevant tissue and cell types [120].

LDSC assumes $\mathrm{Var}(\beta_j) = \frac{h^2}{M}$, i.e. the heritability contribution from each SNP is about the same genome-wide. This is a strong assumption that is broken in many cases, e.g. coding vs. non-coding variants, or SNPs in brain genes vs. muscle genes for a psychiatric phenotype. S-LDSC extends LDSC to allow different regions of the genome to have different contributions to trait heritability (i.e. regions with different $\mathrm{Var}(\beta_j)$).

To this end, the previous assumption can be generalised to allow the variance to differ between annotated categories C with effect sizes τ:

$$\mathrm{Var}(\beta_j) = \sum_{c:j\in C_c} a_{cj}\tau_c \tag{2.11}$$

where C_c indicates the SNPs in the c-th category. τ_c is the per-SNP contribution of one unit of annotation a_c to heritability and a_{cj} the value of annotation c for SNP j. a_{cj} may be binary (0/1) or continuous.

This assumption changes the expression of the expected χ^2-statistic:

$$\begin{aligned}
\chi_j^2 &\equiv \frac{\hat{\beta}^2_{GWAS,j}}{\mathrm{Var}(e_j)} \approx N(\sum_k \beta_k r_{jk} + s_j + e_j)^2 \\
\mathbb{E}(\chi_j^2) &= N(\sum_k \mathbb{E}(\beta_k^2) r_{jk}^2) + \mathbb{E}(s_j^2) + \mathbb{E}(e_j^2) \\
&= N(\sum_c a_{cj}\tau_c \sum_{k\in C_c} r_{jk}^2) + Na + 1 \\
&= N\sum_c \tau_c l_{j,c} + Na + 1
\end{aligned} \tag{2.12}$$

where

$$l_{j,c} \equiv \sum_{k\in C_c} a_{cj} r_{jk}^2, \tag{2.13}$$

is the newly defined stratified LD score of SNP j with respect to annotation c and r_{jk} is the genotypic correlation between SNPs j and k. τ_c is now estimated via multiple regression with $l_{j,c}$ computed from reference data for a choice of annotations.

When applying S-LDSC, the following practical guidelines should be kept in mind: Large sample sizes are required for partitioning heritability [26] for S-LDSC to have reasonable power. This is true in my analyses since many GWAS summary

statistics I use are based on the UK Biobank, allowing for large sample size. Second, the LD reference panel has to match the population in the GWAS [121]. In these analyses, this is both European ancestry using the 1000 Genome reference panel. Third, S-LDSC might underestimate heritability if the traits in the analyses are not highly polygenic [122], which again is averted by the careful choice of GWAS summary statistics used.

2.5.3 S-LDSC evaluation metrics

I assess the informativeness of an annotation c using two metrics. The first metric is enrichment E, defined as follows (for binary and probabilistic (0-1) annotations):

$$E_c = \frac{\frac{h^2_{SNP}(c)}{h^2_{SNP}}}{\frac{\sum_j a_{cj}}{M}}, \tag{2.14}$$

where $h^2_{SNP}(c)$ is the heritability explained by the SNPs in annotation c, weighted by the annotation values. This compares the heritability per SNP within the annotation to the genome-wide average heritability per SNP.

The second metric is the standardised effect size $\tau^\star$ defined as follows (for binary, probabilistic, and continuous-valued annotations):

$$\tau^\star_c = \frac{\tau_c \mathrm{sd}_c}{\frac{h^2_{SNP}}{M}}, \tag{2.15}$$

where sd_c is the standard deviation of annotation c, h^2_{SNP} is the total SNP heritability, and M is the total number of SNPs in which heritability is estimated. $\tau^\star$ represents the proportionate change in per-SNP heritability associated with a 1 standard deviation increase in the annotation value. The main difference between enrichment and $\tau^\star_c$ is that $\tau^\star_c$ quantifies effects that are unique to the focal annotation c (after conditioning on all other annotations by testing whether $\tau^\star$ differs from zero in multiple regression), whereas enrichment does not differentiate between effects that are unique to the focal annotation and those that are shared by other annotations.

3
Materials & Methods

Contents

3.1 Encoding of epigenetic data 31
3.2 EpiNN model 33
3.2.1 Choice of parameters 34
3.2.2 EpiNN training 34
3.2.3 Variations of data encoding and model architecture . . . 36
3.2.4 Prediction of novel AIIS regions: genome-wide EpiNN annotation 37
3.3 Additional immune system-specific annotations 38
3.3.1 Immune system-specific histone mark annotations . . . 38
3.3.2 Immune system-specific ChromHMM annotation 39
3.4 S-LDSC analysis 39
3.4.1 S-LDXR 41
3.5 Panther GO term analysis 41
3.6 Experimental analysis of *DNMT1* 42
3.6.1 CAGEseq data 42
3.6.2 Human peripheral blood samples 43
3.6.3 Immunoblotting 43
3.6.4 Polymerase chain reaction (PCR) 44

The EpiNN model comprises of a convolutional neural network (CNN) trained on epigenetic data to identify genomic regions belonging to a specific class of genomic regions, such as AIIS genes. A schematic overview is given in fig. 1.1. This section provides details on the encoding of epigenetic data, the EpiNN model, the training and deployment procedures, and the experimental setup for downstream analyses.

3.1 Encoding of epigenetic data

The method represents epigenetic data for a given genomic region in the EpiNN model using an image-like $n \times m \times 3$ format. Each image contains $n = 111$ rows representing 111 different cell and tissue types [4]. Each row encodes ChromHMM [1] data from the epigenome roadmap [4], and each column m represents the epigenetic state of 200 base pairs (bps), which is close to the number of base pairs wrapped around a histone. Images were set to have a fixed size spanning 100 kilobases (kb) by optimisation (see next paragraph). Each matrix entry was encoded as a 3-dimensional value, where similar states are assigned similar values. To facilitate human readability for the labelling of training examples and in downstream analysis, I encoded each state using a Red-Green-Blue (RGB) colour scheme, adopting the scheme used in Kundaje et al. [4], with a few modifications, e.g. to reorder the shades of green for the transcription states (see table 3.4). Instead of the weak transcription (TxWk) being displayed by the darkest shade of green of the three transcription states, I set it to be the lightest shade of green to represent the weakness of the transcription signal. I used Tx dark green, TxFlnk green, and TxWk lime green; instead of Tx green, TxFlnk lime green, TxWk dark green.

To standardise the reading direction of the genes in the images, I horizontally mirrored all the epigenomes on the minus strand of the genome. Since the epigenetic signals at the start of a gene (e.g. transcription start sites) differ significantly from those at the end, I standardised the reading direction to remove this added complexity from the data. In the preprocessed images, the start of a gene was always on the left side of the image and the end of the gene on the right. When sets of genes were analysed (e.g. to train the EpiNN model), each image contained a single gene. When a gene did not fit into a single image, I found that only including the start of each gene, which contains the most diverse and relevant epigenetic signal, resulted in a good performance. For the training of the model, the label of an image was set to 1 when there was an AIIS gene in the image, and to 0 if not.

Because the CNN learns epigenetic patterns localised within regions of the training images, ordering rows (cell/tissue types) impacts the model's performance.

Due to the reduced volume of available training data, the model did not learn an optimal ordering for the rows (cell/tissue types) in the input images and instead sorted image rows according to tissue similarity. I decided instead to adopt the ordering used in Kundaje et al. [4] but changed the position of the two thymus data sets to be placed underneath the T-cell data sets, as the thymus is where the T cells reside during their maturation. To facilitate comparison with existing literature, however, I chose to display all images in this thesis using the colour scheme and row order of Kundaje et al [4].

Image size in bps

I had to standardise the input images dimensions, so that I could use the model to sweep the human genome for novel AIIS regions. Because genes generally differ in length, I considered the following options:

- Re-scaling images. This could lead to a distortion of the gene type-determining features with a negative impact on image classification and would be especially aggravating with a large spread in gene length. On the positive side, each image would reliably contain exactly one gene per image.

- Fixing image size to the longest gene while retaining the resolution of 200 bps per pixel. This would be computationally inefficient in the CNN. Additionally, filling up the shorter genes with surrounding bps could lead to having multiple genes of different types in one image. The spread of gene length could mean up to the order of 10 genes per image. With them being of different gene classes, the classification of one specific of them would be very hard.

- Fixing the image size to the longest gene while decreasing the resolution of 200 bps per pixel. This would mitigate the computational costs but would not help with the multiple genes per image.

- Fixing the image size and retaining the resolution to a value lower than the longest gene by cutting off long genes or filling up the small genes with neighbouring bps. Cutting off the long genes could lead to losing important

information, mainly since regulating epigenetic features often are located at the start and end of a gene. Multiple genes in one image could still be a possibility but would be less likely and fewer in number.

I decided on the last option to fix the image size. As the epigenetic signal is most diverse at the start of genes, I placed the start of each gene in the middle of the image, with the gene continuing to the right, to make sure to capture the epigenetic signal before and around each gene start site.

3.2 EpiNN model

The main EpiNN model consists of a CNN comprising three stacked modules of a convolutional layer, ReLU activation, and a max pooling layer. The convolutions operate on 3×3 windows, and the max pooling layers operate on 2×2 windows. The first convolution employs 16 filters, the following one employs 32 filters, and the last employs 64 filters. After these 3 modules, I added two fully-connected layers with a dropout layer in between to reduce overfitting. The network output uses a sigmoid activation producing a score between 0 and 1, reflecting the predicted confidence that the input image contains an AIIS gene (see fig. 3.1).

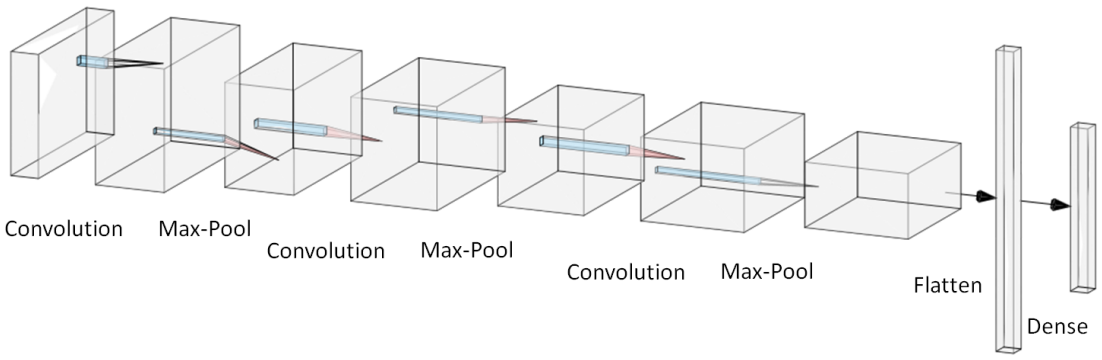


Figure 3.1: CNN architecture graphical overview. The CNN architecture then consisted of 3 stacked modules of a convolutions layer, ReLU activation and a layer of max pooling.

3.2.1 Choice of parameters

Since the training data for the EpiNN model is very limited (see below), I could only optimise some hyperparameters and chose the others to the best of my ability. The most important hyperparameter when applying a CNN to genetic and epigenetic data is the size of the convolution filter, which is a small matrix that is used to extract features from the sequence. The size of the convolution filter should be chosen such that it is large enough to capture the relevant features of the epigenetic sequence, but not so large that it captures too much noise. In this case, I chose the filter size to be 3×3, which is big enough to capture the defining features I am expecting: The edges between immune system cell types and non-immune system cell types, in the case of an AIIS-specific genomic region showing different epigenetic features in the block of immune system cells and the non-immune system cells, and in the case of non-AIIS-specific genomic regions, no difference. I chose to double to number of filters in every convolutional layer, as this allows the later layers to pick up more complex patterns from the earlier layers. The architecture itself was chosen as a common image classification architecture, for which I had not enough training data to optimise the number of convolutional layers. The other hyperparameters were optimised using grid search from a grid of wide, but commonly used, values. Other than the choice of filter size, none of the hyperparameters represent biological assumptions.

3.2.2 EpiNN training

I used the preprocessed images to train EpiNN to identify AIIS-relevant regions. The training data for the EpiNN model consisted of a hand-crafted list of AIIS genes, selected using the following criteria: Loci relevant for immune cell identity, such as most leukocyte signature matrix [123] or cancer immunity cycle genes [124], which exhibit a distinctive pattern in the Epigenome Roadmap 15 state model visualisation. Specifically, loci where most primary immune cells display highly activated promoter, elongation and/or enhancer signals in stark contrast to other cell types, where either quiescent or repressed epigenetic states typically dominate. ChromHMM [1]

profiles of genes from the MetaCore Database [125] were inspected and 888 genomic loci manually labelled where such a discrepancy between the epigenetic markers of immune cells and those of other cell types could be observed. This hand-crafted gene set was used as the training data of the EpiNN model, together with an equal number of randomly sampled genes to represent non-AIIS regions, which were manually inspected to verify that they do not harbour immune system-specific signatures.

I performed a grid search to optimise hyper-parameters in each model. These included image size in bps (10 kbp, 25 kbp, 50 kbp, 100 kbp, 200 kbp, 400 kbp), image size in pixel (100 pixels, 500 pixels, 1000 pixels), dropout rate (0.1, 0.3), and number of epochs, i.e. the number of times the learning algorithm will work through the entire training dataset (10, 50). Batch size (64) and learning rate (0.001, Adam optimiser [126]) were held constant. These constant hyperparameters didn't seem to have much influence on the result in preliminary experiments. I divided the data into 10 random subsets of approximately equal size. I reserved one such set for testing, then optimised hyperparameters and performed model selection on the remaining 9 subsets through cross-validation, using binary cross-entropy loss to evaluate the performance of each model configuration. As an additional metric, I further measured the area under the precision-recall curve (AUPRC). Final hyperparameters are shown in table 3.1. An overview of the next best AIIS EpiNN models is given in table 3.2, showing that the model does not overfit the training data.

As an additional parameter, I analysed the influence of the image length in bps on the model performance (see table 3.3 for details). I found that both too-small ($<$ 25 kbp) and too-large ($>$ 200 kbp) image lengths led to worse, but not significantly, model performance, as shown in table 3.3. I interpret this the following way: For smaller window sizes, the models lack sufficient contextual information, e.g. enhancers or other elements around the gene but not necessarily part of the gene. For very large images, multiple genes may be included in a single input. Although this may potentially allow learning interactions between neighbouring genes, it would complicate the classification task, particularly given that limited training data is available.

AIIS genes

Model Cross validation	AUPRC	Binary Cross Entropy Loss	Window length [bps]	Image length [pixel]	dropout rate	n of epochs
EpiNN	0.951	0.234	100k	1000	0.1	10
EpiNN-1D	0.944	0.243	100k	100	0.1	50
EpiNN-binary	0.935	0.251	100k	100	0.3	50
EpiNN-logistic	0.92	0.266	100k	500	-	-

Housekeeping genes

Model Cross validation	AUPRC	Binary Cross Entropy Loss	Window length [bps]	Image length [pixel]	dropout rate	n of epochs
EpiNN	0.962	0.181	50k	500	0.1	50
EpiNN-1D	0.953	0.193	50k	100	0.3	10
EpiNN-binary	0.947	0.197	100k	500	0.1	10
EpiNN-logistic	0.931	0.202	100k	500	-	-

Table 3.1: Performance and hyperparameters of EpiNN models. Sets of optimal hyperparameters and performance metrics resulting from a grid search for each model configuration.

Window length [bps]	Image length [pixel]	dropout rate	n of epochs	Training Binary Cross Entropy Loss	Training AUPRC	Test AUPRC	Test Binary Cross Entropy Loss
100k	1000	0.1	10	0.238	0.968	0.951	0.234
100k	1000	0.1	50	0.24	0.968	0.942	0.251
100k	500	0.3	10	0.24	0.967	0.949	0.243
100k	1000	0.3	10	0.245	0.957	0.949	0.241
100k	500	0.1	10	0.242	0.948	0.938	0.247

Table 3.2: Overview of best EpiNN models with hyperparameters. Overview of the top 5 hyperparameter optimisations for the AIIS EpiNN model, sorted by Training AUPRC. The model performance is close for the different hyperparameters. Training and testing metrics are similar, indicating that the model does not overfit the training data.

3.2.3 Variations of data encoding and model architecture

I tested variations of the input encoding and model architecture, including one-dimensional and binary encodings of the epigenetic states (EpiNN-1D and EpiNN-binary models, see table 3.4), as well as a logistic regression model (EpiNN-logistic, equivalent to a single-node fully connected network). The EpiNN-logisitc model was trained on the same data as EpiNN, but the matrices were unravelled into a long vector so that the logisitc function could be applied. The EpiNN-logistic model

window length in kbp	Accuracy achieved	Std dev
10	0.928	0.085
25	0.939	0.071
50	0.95	0.050
100	0.95	0.071
200	0.95	0.058
400	0.922	0.064

Table 3.3: Image length analysis. I analysed the influence of the image length in bps on the model performance. One can see that the cross-validated, held-out accuracy of the optimised model for different window lengths decreases both for smaller window sizes and for larger window sizes.

was regularised using L2 norm using a penalty of 1.0 (see chapter 2). Models with lower dimensional encodings further leverage prior knowledge about relationships between chromatin states to compress the input space, which may lead to better performance in small training data sets. For the EpiNN-1D model, states were ranked and encoded using a single value ranging between 0 and 1. Transcription-enhancing states were assigned larger values, and transcription-inhibiting states were assigned lower values. For the EpiNN-binary model, transcription-enhancing states were assigned value 1, and transcription-inhibiting states were assigned value 0. See table 3.4 for details on the colour schemes.

3.2.4 Prediction of novel AIIS regions: genome-wide EpiNN annotation

I applied the trained and optimised EpiNN model to predict novel AIIS-relevant regions along the genome using a sliding window approach. To this end, I divided the genome in overlapping windows of 100 kbp, using intervals of 5 kbp as a step size. Due to the resulting overlap between windows, this approach yields 20 EpiNN predictions for each 5 kbp window. To combine these into a single score, I computed a rolling average with Gaussian weights (SD=0.5) so that predictions, where the 5 kbp window is in the centre of the image, have the highest weight in the final score for the region. Finally, I produced a genome-wide annotation with these EpiNN predictions by taking the maximum predicted value between the forward

Mnemonic	Description	Colour Name	EpiNN	1D	binary
TssA	Active TSS	Red	255,0,0	1	1
TssAFlnk	Flanking Active TSS &	Orange Red	255,69,0	0.9	1
TxFlnk	Transcr. at gene 5' and 3'	Green	0,128,0	0.7	1
Tx	Strong transcription	DarkGreen	0,100,0	0.8	1
TxWk	Weak transcription	LimeGreen	50,205,50	0.4	1
EnhG	Genic enhancers	GreenYellow	194,225,5	0.5	1
Enh	Enhancers	Yellow	255,255,0	0.6	1
ZNF/Rpts	ZNF genes repeats	Medium Aquamarine	102,205,170	0.3	1
Het	Heterochromatin	PaleTurquoise	138,145,208	0.05	0
TssBiv	Bivalent/Poised TSS	IndianRed	205,92,92	0.15	0
BivFlnk	Flanking Bivalent TSS/Enh	DarkSalmon	233,150,122	0.1	0
EnhBiv	Bivalent Enhancer	DarkKhaki	189,183,107	0.2	0
ReprPC	Repressed PolyComb	Silver	128,128,128	0.05	0
ReprPCWk	Weak Repressed PolyComb	Gainsboro	192,192,192	0.05	0
Quies	Quiescent/Low	White	255,255,255	0	0

Table 3.4: Encodings used for EpiNN input data. I transformed the output of the ChromHMM model, which uses a one-hot encoding across 15 possible chromatin states, to a 3-dimensional, 1-dimensional, or binary encoding. To facilitate human readability in labelling and follow-up analysis, the 3-dimensional representation encodes each state using a Red-Green-Blue (RGB) colour scheme where similar states are assigned similar values. I used the colour scheme adopted in [4], applying a few minor modifications, e.g. reordering the shades of green used for the transcription states to reflect signal strength.

and backward strand directions. To reduce noise in several downstream analyses, I clipped all annotation values below 0.75, setting them to 0.

3.3 Additional immune system-specific annotations

I created several additional immune system-specific annotations to evaluate the unique contribution of the EpiNN model in addition to that of data sets that are directly or indirectly provided in the input.

3.3.1 Immune system-specific histone mark annotations

I downloaded all available immune system cell tracks for each H3K27me3, H3K36me3, H3K9me3, H3K4me1, H3K4me3, and H3K27ac from the Epigenome Roadmap data base [4]. I averaged the signal over all cell types to obtain a genome-wide annotation for each individual chromatin mark, resulting in 6 annotations: H3K27me3$_{IS}$, H3K36me3$_{IS}$, H3K9me3$_{IS}$, H3K4me1$_{IS}$, H3K4me3$_{IS}$, and H3K27ac$_{IS}$.

3.3.2 Immune system-specific ChromHMM annotation

I averaged the ChromHMM data [1] over all immune system cells and non-immune system cells separately. Transcription-inducing states were marked as "1" and transcription-inhibiting states as "0", as discussed in EpiNN-binary. I subtracted the non-immune system average from the immune system average, resulting in an immune-specific annotation ChromHMM$_{IS}$.

An overview of all annotations is given in table 3.5

Annotation	percentage of genome covered	binary/continuous
H3K36me3$_{IS}$	18.2%	continuous
H3K4me1$_{IS}$	17.0%	continuous
H3K9me3$_{IS}$	19.2%	continuous
H3K27me3$_{IS}$	21.2%	continuous
H3K4me3$_{IS}$	22.2%	continuous
H3K27ac$_{IS}$	19.4%	continuous
ChromHMM$_{IS}$	0.4%	continuous
EpiNN	3.5%	continuous, thresholded
IRIS	2.8%	binary
ImmunoGene	1.6%	binary

Table 3.5: Overview of S-LDSC annotations. Additional details on annotations used for the S-LDSC analyses: Each annotation is marked as continuous (taking continuous values) or binary (taking values in $\{0, 1\}$). The EpiNN annotation used for S-LDSC analyses is additionally thresholded by setting all values < 0.75 to 0. The "percentage of genome covered" column denotes the fraction of the genome for which the annotation has a non-0 value, i.e. regions where continuous annotations are defined and regions where binary annotations take value 1.

3.4 S-LDSC analysis

I tested whether regions predicted to be AIIS-relevant by EpiNN are enriched for complex trait heritability using Stratified Linkage Disequilibrium Score Regression (S-LDSC) [25–27, 116]. In the analyses, I compare my focal annotations against the baseline model, consisting of 97 curated annotations (Baseline-LD model version 2.2), including functional (e.g. coding regions, enhancer regions, or CTCF sites) and evolutionary annotations (e.g. conserved regions, or recombination rates) [26].

I computed the statistical significance (p-values) of the enrichment and $\tau^\star$ of each annotation via block-jackknife over 200 blocks [26]

To increase statistical power, I performed a random-effects meta-analysis [127] of heritability using summary association statistics from 63 studies for 57 unique heritable traits. Of these, 57 studies involved traits with negligible genetic correlation to all other traits ($r^2 < 0.1$) and 6 more involved duplicate traits (genetic correlation $r^2 > 0.9$ with another trait) but disjoint samples [128]. For simplicity, I refer to this set of 63 studies as "63 independent traits" throughout. A subset of 5 studies involved immune system-relevant traits. I performed meta-analyses including either all studies or the subset of immune-relevant and non immune-relevant traits. The traits analysed are reported in table B.2.

In addition to these 63 studies used for meta-analysis, I performed further heritability analyses for the EpiNN annotation using a wider list of 176 summary association statistics from several diseases and complex traits, reported in table B.4. These are not necessarily independent, but contain a broader set of 26 immune system relevant traits and diseases. To test whether the EpiNN annotation had a significant heritability effect size τ^* when conditioned on other annotations, such as IRIS and ImmunoGene, I performed joint S-LDSC analyses where all tested annotations were jointly included in addition to the Baseline-LD model. To account for the effects of highly-correlated annotations in the joint S-LDSC analysis, I performed forward stepwise elimination [27, 129]. In this procedure, I iteratively added one annotation at a time to an existing set of annotations, starting from the Baseline-LD model, and only retained annotations that remained Bonferroni-significant at each iteration. Because the set of annotations retained by this heuristic procedure may vary depending on the order in which annotations are considered, I repeated this analysis using all possible orders but pruned specific combinations that were unlikely to be jointly significant. Specifically, whenever annotation A became non-significant after adding annotation B, I only included annotation B in any combination of annotations that would have otherwise have contained both A and B.

3.4.1 S-LDXR

S-LDXR was run in collaboration with Steven Gazal. S-LDXR [130] is a method to estimate the enrichment of stratified squared multi-ancestry genetic correlation across functional categories of SNPs using GWAS summary statistics and population-matched linkage disequilibrium (LD) reference panels. S-LDXR first estimates the genome-wide squared multi-ancestry genetic correlation r_g^2, which is defined as the squared correlation between per-allele effect sizes across two ancestries; I note that S-LDXR focused on per-allele effect sizes rather than per-SNP h^2_{SNP} (as S-LDSC) to account for differences in allele frequencies across populations. Then, S-LDXR estimates the enrichment/depletion of squared multi-ancestry genetic correlation within an annotation c, which is defined as $r_g^2(c)/r_g^2$, where $r_g^2(c)$ is the squared multi-ancestry genetic correlation of variants within annotation c. We applied the S-LDXR method using recommended settings, reference files (i.e. 481 East Asian and 489 European samples in the 1000 Genomes Project), and a background set of functional annotations (i.e. the baseline-LD-X model a set of 62 functional SNP-annotations known to impact per-allele effect sizes). We analysed 15 immune traits from the total of 31 traits analysed in Shi et al. [130]; GWAS were obtained for both ASI and EUR ancestries using Biobank Japan, UK Biobank, and other sources (see Suppl. Table 2 in [130]).

3.5 Panther GO term analysis

I tested for functional enrichment in regions corresponding to high EpiNN annotation values using gene ontology (GO) term analysis. I first used HUGO identifiers to match annotated genomic regions to known genes. For every gene, I then computed the average EpiNN annotation value. To reduce noise. I performed a GO term analysis of the biological process ("GO biological process complete"), the cellular components ("GO cellular component complete"), and the molecular function ("GO molecular function complete") assignments using the PANTHER [24] tool for all genes with an average EpiNN score higher than 0.5. The analysis was run using the

following settings: analysis type: PANTHER overrepresentation test (released 20200407); annotation GO ontology database (DOI: 10.5281/zenodo.3727280) released on 2020-03-23; reference list: Homo Sapiens (all genes in database); test type: FISHER; correction: Bonferroni. I report results satisfying Bonferroni-corrected $p < 0.05$. To verify that the results of these analyses are not driven by the set of genes used in training, I also repeated these analyses after excluding the set of training genes.

3.6 Experimental analysis of *DNMT1*

The experimental analysis of *DNMT1* was performed in collaboration with Calliope Dendrou's research group, namely Calliope Dendrou and Weijiao Zhang.

3.6.1 CAGEseq data

Cap analysis of gene expression sequencing (CAGEseq) [131, 132] measures DNA expression and maps transcription start sites. We apply it here to test whether a putative transcript start site is the actual start site of a transcribed mRNA. in CAGEseq, RNA is isolated and reverse-transcribed using generic primers that are not specific to certain RNA sequences. In a step called cap capture, only RNA with complete 5'prime ends are selected, which are then transcribed into cDNA, sequenced and mapped to the genome. This way, the transcription start site of each sequence can be identified.

CAGEseq data was collected from the FANTOM5 [133] database for the following cell types: CD14-/CD16+ Monocytes, Dendritic Cells monocyte immature derived, CD19+ B Cells pluriselect, CD34+ Progenitors donor 1, CD34+ Progenitors donor 2, CD34 stem cells adult bone marrow derived, CD4+ CD25+ CD45RA+ naive regulatory T cells donor 1, CD4+ CD25+ CD45RA+ naive regulatory T cells donor 2, CD4+ CD25- CD45RA- memory conventional T cells donor 1, CD4+ CD25- CD45RA- memory conventional T cells donor 2, CD8+ T Cells pluriselect donor 090309 donation 1, CD8+ T Cells pluriselect donor 090309 donation 2, CD8+ T Cells pluriselect donor 090309 donation 3, CD8+ T Cells pluriselect donor

090325 donation 1, CD8+ T Cells pluriselect donor 090325 donation 2, CD8+ T Cells pluriselect donor 090612 donation 1, CD8+ T Cells pluriselect donor 090612 donation 2, CD8+ T Cells pluriselect donor 090612 donation 3, Common myeloid progenitor, Eosinophils, and Basophils.

3.6.2 Human peripheral blood samples

To perform experimental validation for the *DNMT1* gene, which was predicted to be AIIS-relevant by EpiNN, we performed PCR and Western Blot analysis on immune system and non-immune system cells. Peripheral blood samples were obtained by venesection of volunteers from the Oxford BioBank, a random, population-based cohort of healthy adults. Samples were obtained with informed consent under CUREC R45381/RE001 approval. Peripheral blood mononuclear cells were separated by density gradient centrifugation using Lymphoprep (Axis-Shield, and specific cell subsets were isolated by positive selection using MicroBeads (Miltenyi). Cell viability was estimated to be $> 95\%$.

3.6.3 Immunoblotting

Cells were lysed with lysis buffer containing 20 mM HEPES pH 7.5, 100 mM sodium chloride, 0.05% Triton X-100 and 1 mM dithiothreitol (all Sigma-Aldrich, UK). Lysis buffer was supplemented with protease inhibitors (Roche Mini Complete from Roche Diagnostics, UK) immediately before use. Protein concentration was determined using a Nanodrop One spectrophotometer (Thermo Fisher, UK) with the BCA kit (Thermo Fisher, USA). Lysates were mixed with Laemmli sample buffer (BioRad, UK) and denatured at 95°C for 10 min. The denatured lysates were then loaded onto a 12% polyacrylamide gel made using the TGX FastCast Acrylamide Kit (BioRad, UK), with 20 µg of protein loaded per lane. Following blotting onto nitrocellulose membrane, anti-actin (clone C4, Millipore) and anti-DNMT1 (clone JF09-89, Novus Biologicals) were used for protein detection, with staining with IRDye 680LT- and 800CW-conjugated secondary antibodies and visualisation

using an Odyssey Infrared Imaging System (LI-COR Biosciences). Analyses were performed with n=3 individuals in technical duplicate.

3.6.4 Polymerase chain reaction (PCR)

Cells were lysed with buffer RLT and RNA isolation, and genomic DNA removal and cDNA synthesis were performed with Qiagen kits, according to the manufacturer's protocols. PCR was performed using the NEBNext® High-Fidelity 2X PCR Master Mix (NEB) with primers against the *DNMT1* transcript with the alternative start site (5'-GCATAAGCTCACCACAAAGGA-3' and 5'-GGCCTCGAGGAAGTAGAAGC-3') and positive control primers (against *TYK2*, 5'-GTACTGCTACGATCCGACCA-3' and 5'-GCAGCCCTTGTACTTGATGA-3'). The PCR program used was 98°C for 30 s, and 30 cycles of 98°C for 10 s, 66°C for 30 s, and 72°C for 20 s, followed by a further 2 min at 72°C. PCR products were resolved by gel electrophoresis and visualised with a UV transilluminator. Analyses were performed with n=3 individuals in technical duplicate.

Other PCR primers tried unsuccessfully:

For completeness, these are the other primers we designed and tested but which did not work.

> *DNMT1* intron 28 starting from the CAGE tags, chr19:10142221-10142901

cacactcgagaggctgctaccggcaggcggacacagcagggcagggccaggcgtgcagagcaagcag

ccaagctgacctaatggttgggggtctctaccctgtggttccctaattgggggcgaatctaccttg

cagttccctcatttggggaggggttatctttacctgcggttccctaagtagggggtctttacccta

tgactccctaactggggaggtctttaccctgcggttccataactgggggtctttaccctgcggttc

ccaactgggtcgggggtcattacctggcggttccctaattgtgggggaatctaccttgcagttccct

aatgtgaggaggggatatctttaccctgtgattcccttactgggaggggtctttaccctgcgattcc

ctaacttgggggatctttaccctgtggttccctaactagggggggtctttaccctgcagttccctaa

atgggggggagtctaccttgcagttccccagtgtggggaggagatatctttaccctgcagttcccta

actgagggggtctttaccctgcggttctcgaactagggggtctttaccctgcagttccctgagcacc

ccaggaacatggtactgtttaattcagcataagctcaccacaaaggagagctcctaacgaggcctct

cccgctcgcag

> *DNMT1* exon 29, 193 nt, chr19:10142028-10142220 (Forward primer (69.2 °C))
GCCTGAGAACACCCACAAGTCCACTCCAGCGAGCTACCACGCAGAC
ATCAACCTGCTCTACTGGAGCGACGAGGAGGCCGTGGTGGACTTCA
AGGCTGTGCAGGGCCGCTGCACCGTGGAGTATGGGGAGGACCTGCC
CGAGTGCGTCCAGGTGTACTCC**ATG**GGCGGCCCCAACCGCTTCTACT
TCCTCGAG

> *DNMT1* exon 30, 85nt, chr19:10141105-10141189 (Reverse primer #1 (67.0 °C))
GCCTATAATGCAAAGAGCAAAAGCTTTGAAGATCCTCCCAACCATG
CCCGTAGCCCTGGAAACAAAGGGAAGGGCAAGGGAAAAG

> *DNMT1* exon 31, 129nt, chr19:10140781-10140909 (Reverse primer #2 (66.3 °C))
GGAAGGGCAAGCCCAAGTCCCAAGCCTGTGAGCCGAGCGAGCCAG
AGATAGAGATCAAGCTGCCCAAGCTGCGGACCCTGGATGTGTTTTC
TGGCTGCGGGGGGTTGTCGGAGGGATTCCACCAAGCAG

> *DNMT1* exon 32, 283nt, chr19:10140046-10140328 (Reverse primer #3 (67.6 °C))
GCATCTCTGACACGCTGTGGGCCATCGAGATGTGGGACCCTGCGGC
CCAGGCGTTCCGGCTGAACAACCCCGGCTCCACAGTGTTCACAGAG
GACTGCAACATCCTGCTGAAGCTGGTCATGGCTGGGGAGACCACCA
ACTCCCGCGGCCAGCGGCTGCCCCAGAAGGGAGACGTGGAGATGCT
GTGCGGCGGGCCGCCCTGCCAGGGCTTCAGCGGCATGAACCGCTTC
AATTCGCGCACCTACTCCAAGTTCAAAAACTCTCTGGTGGTTTCCT
TCCTCAG

> *DNMT1* exon 33, 142nt, chr19:10139676-10139817 (Reverse primer #4 (69.0 °C))
CTACTGCGACTACTACCGGCCCCGGTTCTTCCTCCTGGAGAATGTC
AGGAACTTTGTCTCCTTCAAGCGCTCCATGGTCCTGAAGCTCACCC
TCCGCTGCCTGGTCCGCATGGGCTATCAGTGCACCTTCGGCGTGCT
GCAG

4

EpiNN analysis of immune-related traits and diseases

Contents

4.1 EpiNN training results 46
4.2 Genome-wide scan for novel AIIS-relevant regions . . 49
4.3 GO term analysis 49
4.4 S-LDSC analysis of individual traits 50
4.5 S-LDSC meta analysis 54
4.5.1 Enrichment of heritability in EpiNN quantiles 55
4.6 S-LDSC analysis including additional immune-related annotations 56
4.6.1 S-LDSC joint analysis with underlying epigenetic annotations 57
4.7 S-LDXR analysis 61
4.8 Application of EpiNN to non-immune system genes . 63
4.8.1 Housekeeping genes 63
4.8.2 Epithelial genes 63
4.8.3 Brain genes 65
4.9 Testing the AIIS EpiNN annotation for heritability enrichment in COVID19 infections 66

4.1 EpiNN training results

For each EpiNN model (EpiNN, EpiNN-1dim, and EpiNN-binary), I used cross-validation to tune its hyperparameters and to select the model architecture that per-

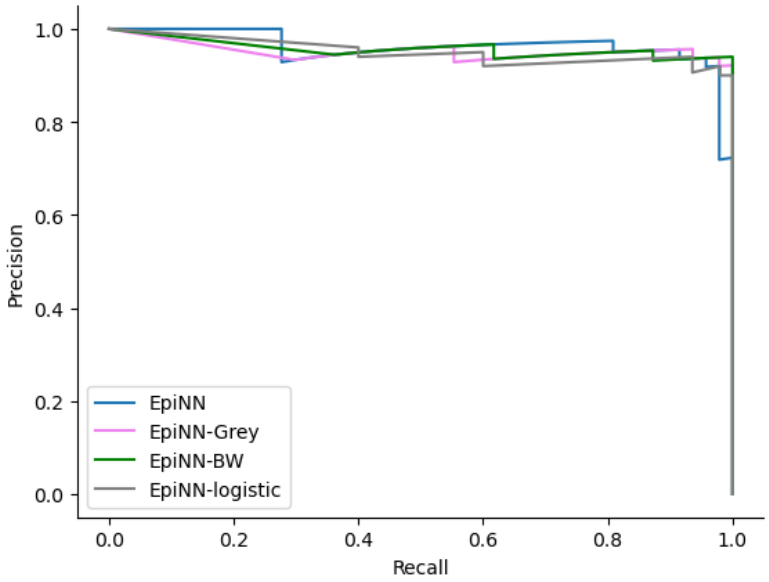


Figure 4.1: Precision-recall curve of EpiNN, EpiNN-1D and EpiNN-binary. Precision-recall curve is similar for the four models.

formed best in the prediction of held-out AIIS-relevant loci as described in chapter 3. The hyperparameters of the optimised models are detailed in table 3.1.

I validated the performance of the optimised EpiNN model by analysing AIIS-relevant regions and, for further validation, housekeeping genes, for which a larger training and testing data set is available. For the AIIS analysis, I trained EpiNN using a hand-crafted training list of 888 known AIIS genes and a set of non-AIIS genes with matching numbers of non-AIIS and AIIS genes per chromosome, with the goal of recognising images containing AIIS genes from images not containing AIIS genes. I applied the trained model to held-out AIIS/non-AIIS gene images to test the model area under the precision-recall curve (AUPRC, see Fig. fig. 4.1), obtaining a testing AUPRC of 0.951. For the analysis of housekeeping genes, I trained on 3754 genes, again obtaining a high AUPRC of 0.962. The corresponding confusion matrices are shown in table 4.1.

I also validated the use of a CNN over simpler architectures within the EpiNN model. To this end, I trained a logistic regression model I call EpiNN-logistic (equivalent to a fully connected single node neural network, see section 3.2.3) with the same data used as input for the CNN architecture, and obtained lower

		EpiNN prediction	
	Total 88	Positive (PP) 44	Negative (PN) 44
Actual condition	Positive (P) 44 Negative (N) 44	True Positive (TP) 42 False Positive (FP) 2	False Negative (FN) 2 True Negative (TN) 42
		EpiNN-1D prediction	
	Total 88	Positive (PP) 43	Negative (PN) 45
Actual condition	Positive (P) 44 Negative (N) 44	True Positive (TP) 41 False Positive (FP) 2	False Negative (FN) 3 True Negative (TN) 42
		EpiNN-binary prediction	
	Total 88	Positive (PP) 43	Negative (PN) 45
Actual condition	Positive (P) 44 Negative (N) 44	True Positive (TP) 41 False Positive (FP) 2	False Negative (FN) 3 True Negative (TN) 42
		EpiNN-logistic prediction	
	Total 88	Positive (PP) 42	Negative (PN) 46
Actual condition	Positive (P) 44 Negative (N) 44	True Positive (TP) 40 False Positive (FP) 2	False Negative (FN) 4 True Negative (TN) 42

Table 4.1: EpiNN confusion matrix. Confusion matrix of EpiNN models for held-out test set at threshold 0.5.

AUPRC of 0.919 for the AIIS genes task and 0.931 for the housekeeping genes task. These results suggest that the CNN architecture may be better suited to capture nonlinear features in the input.

To understand the influence of the dimensionality reduction in the epigenetic data, I tested EpiNN (3-dimensional colour space) against models based on even lower-dimensional data: EpiNN-1D (greyscale, 1-dimensional colour space) and EpiNN-

binary (black-white, binary colour space). EpiNN showed higher testing AUPRC (0.951) than the EpiNN-1D (AUPRC 0.944) and EpiNN-binary (AUPRC 0.935) for predicting AIIS-relevant loci. For housekeeping genes, similarly, EpiNN (AUPRC 0.962) outperforms EpiNN-1D (AUPRC 0.953) and EpiNN-binary (AUPRC 0.947), indicating, that the additional information of ordering the epigenetic states in a 3D space is useful to the model, compared to a 1D scale or binary classification.

4.2 Genome-wide scan for novel AIIS-relevant regions

I applied the trained EpiNN model to perform a genome-wide scan for novel AIIS-relevant genomic regions. To this end, I used a sliding window approach: a window of 100 kbp was shifted in steps of 5 kbp over the whole genome, obtaining a predicted AIIS score for each considered region. The genome was surveyed both in forward and backward direction to achieve invariance to plus/minus strand orientation. A genome annotation with step sizes of 5 kbp was calculated by applying a Gaussian filter to average over all images containing the region of interest (see chapter 3). This results in a genome-wide annotation of scores for predicted AIIS relevance that I hereafter refer to as the EpiNN annotation. The EpiNN annotation was further mapped onto a list of all known genetic regions with a HUGO name, to identify positively predicted genes. 2873 novel potential AIIS genes were identified that could be matched to a HUGO name (of ∼33,000 current HUGO names [134]).

4.3 GO term analysis

I performed a gene ontology (GO) term analysis of the biological process ("GO biological process complete") assignment using the PANTHER Over-Representation Test [24]. Of the candidate AIIS-relevant genes with EpiNN score greater than 0.5, 1413 could not be mapped and 158 did not have a GO term, leaving 1194 genes. The GO terms with the lowest p-value are shown in Tab. table 4.2. All GO terms were linked to immune system function, including e.g. defence response

(GO:0006952) with a 2.04× enrichment (Bonferroni-corrected $p = 6.4 \times 10^{-15}$) and immune response (GO:0006955), 2× enriched ($p = 7.2 \times 10^{-21}$).

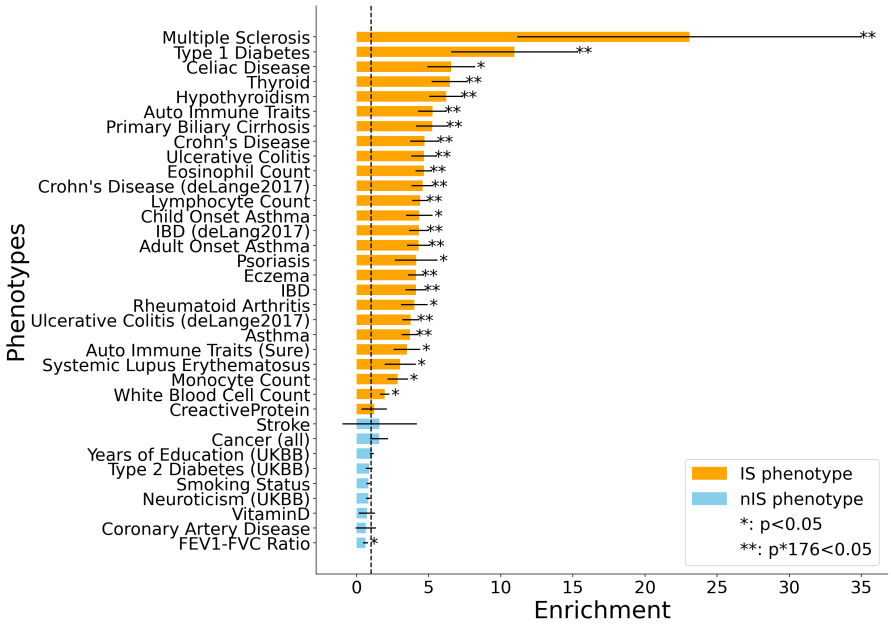
Noticeably, most of the unmapped genes were either LOC (uncharacterised genes, 860 of 1,413), MIR (microRNA, 158 of 1,413), or LINC genes (Long Intergenic Non-Protein Coding, 120 of 1,413). LINC genes have been found to play a role in autoimmune diseases [135]; Linc00402, for instance, was recently validated to have an immune function [136], and was independently implicated by EpiNN.

GO biological process complete	total	found	expected	Enrichment	Bonferroni-corrected p-value
Unclassified (UNCLASSIFIED)	2745	158	180.2	0.88	$< 10^{-22}$
immune system process (GO:0002376)	2849	339	187.03	1.81	4.06×10^{-22}
immune response (GO:0006955)	1972	259	129.46	2	7.20×10^{-21}
regulation of immune system process (GO:0002682)	1701	218	111.67	1.95	1.16×10^{-15}
defense response (GO:0006952)	1416	190	92.96	2.04	6.37×10^{-15}
leukocyte activation (GO:0045321)	935	138	61.38	2.25	8.44×10^{-13}
regulation of immune response (GO:0050776)	1129	156	74.12	2.1	1.23×10^{-12}
cell activation (GO:0001775)	1084	151	71.16	2.12	2.46×10^{-12}
signal transduction (GO:0007165)	5138	474	337.29	1.41	1.29×10^{-11}
response to biotic stimulus (GO:0009607)	1455	179	95.52	1.87	1.43×10^{-10}
response to other organism (GO:0051707)	1418	175	93.09	1.88	2.48×10^{-10}

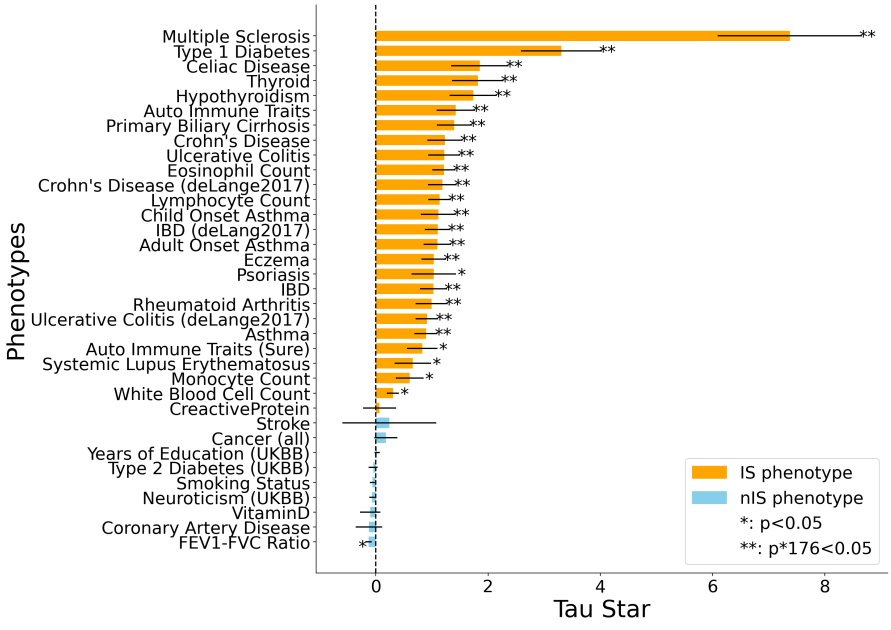
Table 4.2: GO Term analysis. I report the 10 most significantly enriched/depleted GO processes for the set of immune system-relevant genes predicted by EpiNN. The analysis was run using the following settings: Analysis Type: PANTHER Overrepresentation Test (Released 20200407), Annotation Version and Release Date: GO Ontology database DOI: 10.5281/zenodo.3727280 Released 2020-03-23, Reference List: Homo sapiens (all genes in database), Test Type: FISHER, Correction: Bonferroni.

4.4 S-LDSC analysis of individual traits

I next analysed EpiNN's genome-wide predictions to test whether, as a whole, these putative AIIS-relevant regions are enriched for phenotypic effects in immune-related traits and diseases. To this end, I tested whether the AIIS annotation is



(a)



(b)

Figure 4.2: S-LDSC analysis of EpiNN heritability enrichment and effect size. Results are shown for all immune system traits and a selection of 9 (out of 176) non-immune system traits (see table B.4 for numerical values and additional results). Traits are labelled as immune system-related (IS) or non-immune system-related (nIS). For all analyses, S-LDSC was run including EpiNN together with 96 Baseline-LD annotations. **a.** 18 out of 26 immune system traits were significantly enriched ($p < 0.05/176$). **b.** 20 out of 26 had significant ($p < 0.05/176$) effect sizes τ^*.

predictive of regional heritability based on the EpiNN genome annotation, leveraging stratified LD score regression (S-LDSC) [25–27, 116] coupled with genome-wide association summary statistics for 176 traits (average $N = 262k$, see chapter 3). I measured the informativeness of the EpiNN annotations using two standard metrics: heritability enrichment and the standardised effect on heritability, denoted as τ^* [26, 27] (see chapter 3). A key difference between the enrichment and τ^* metrics is that τ^* conditions on other annotations that are present in the model, thus quantifying the unique contribution of a tested annotation. I annotated 26 of the 176 traits as being immune-related (see table B.4) and had my classification checked by Calliope Dendrou.

For 20 of these, including e.g. Crohn's disease, multiple sclerosis, and hypothyroidism, I detected a significant (Z-test $p < 10^{-100}$) heritability enrichment (average over 26 traits: 3.97, SE 0.05) and effect size ($|\tau^*|$ average over 26 traits: 0.987, SE 0.015) for the EpiNN annotation (see fig. 4.2). These significant effect sizes imply that the EpiNN annotation captures signal that is not detected using the 97 other functional and evolutionary annotations present in the baseline-LD model. The 26 immune system-relevant phenotypes include 21 auto-immune diseases such as type-1 diabetes ($\tau^* = 3.3$, SE 0.7, Z-test $p = 2.2 \times 10^{-6}$), multiple sclerosis ($\tau^* = 7.4$, SE 1.3, Z-test $p = 4.7 \times 10^{-9}$), and ulcerative colitis ($\tau^* = 1.2$, SE 0.29, Z-test $p = 1.0 \times 10^{-5}$) as well as immune-related biomarkers, e.g.: eosinophils count ($\tau^* = 1.2$, SE 0.21, Z-test $p = 2.1 \times 10^{-9}$) and lymphocyte count ($\tau^* = 1.1$, SE 0.2, Z-test $p = 8.8 \times 10^{-9}$).

Analysis of additional 150 traits that were not directly labelled as immune-relevant revealed a small set of additional traits with heritability enrichments/depletions for the EpiNN annotations (see table B.4). In particular these included respiratory diseases ($\tau^* = 0.85$, SE 0.2, Z-test $p = 1.3 \times 10^{-5}$) and dermatological diseases ($\tau^* = 1.1$, SE 0.3, Z-test $p = 7.6 \times 10^{-5}$). These phenotypes are not strictly IS phenotypes, as they contain a multitude of diseases, but some of these diseases are IS traits (i.e. allergy eczema for dermatological diseases or asthma for respiratory diseases). The detected enrichment in these two phenotypes is

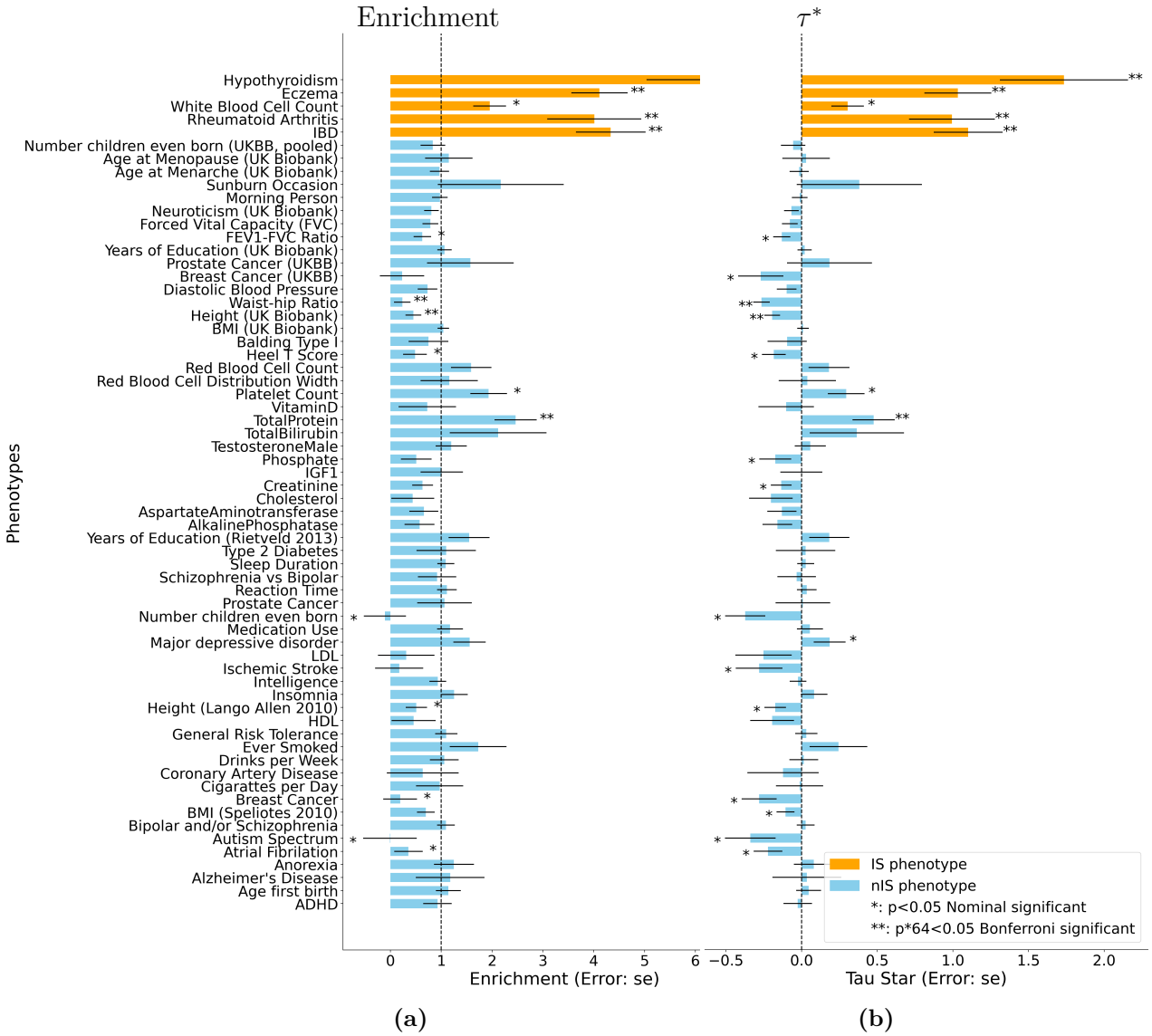
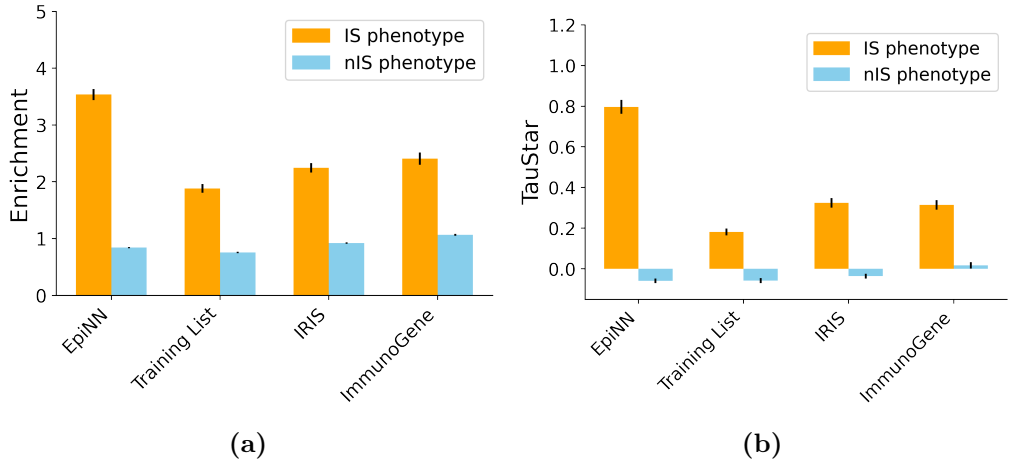


Figure 4.3: S-LDSC analysis of 63 independent traits. a. Heritability enrichment and **b.** effect size τ^* for immune system (IS) and other (nIS) traits and diseases.

therefore expected and in line with the other immune phenotype enrichments. The other enriched phenotypes are waist-hip ratio ($\tau^* = -0.26$, SE 0.05, Z-test $p = 6.8 \times 10^{-7}$), standing height ($\tau^* = -0.19$, SE 0.05, Z-test $p = 8.9 \times 10^{-5}$), and urate blood levels ($\tau^* = -0.25$, SE 0.07, Z-test $p = 1.0 \times 10^{-4}$). Interestingly, these traits have a negative effect size estimate.



(a) (b)

Figure 4.4: S-LDSC meta analysis Meta-analysis for 5 IS and 58 nIS traits and diseases. See table 3.5 for a description of individual annotations. **a.** Enrichment meta-analysis for EpiNN training and testing annotations, and for the IRIS [2] and ImmunoGene [3] expression-based annotations. **b.** Effect size τ^* meta-analysis. Each annotation is separately conditioned on 96 other Baseline-LD annotations.

4.5 S-LDSC meta analysis

To confirm the specificity of the detected enrichments for immune-relevant phenotypes, I performed a meta-analysis of 63 independent traits, which included 4 autoimmune diseases: inflammatory bowel disease (IBD), hypothyroidism, eczema, and rheumatoid arthritis (see chapter 3, fig. 4.3, fig. 4.4, table B.2, and table B.3). These autoimmune diseases also exhibited large effect sizes ($|\tau^*| > 0.5$; Z-test $p < 0.05/63$), again conditioned on 97 other functional and evolutionary annotations (hypothyroidism $\tau^* = 1.73$, SE 0.42, Z-test $p = 2.1 \times 10^{-5}$; IBD $\tau^* = 1.1$, SE 0.23, Z-test $p = 6.8 \times 10^{-7}$; eczema $\tau^* = 1.03$, SE 0.22, Z-test $p = 1.5 \times 10^{-6}$; rheumatoid arthritis $\tau^* = 0.99$, SE 0.28, Z-test $p = 2.3 \times 10^{-4}$; white blood cell count $\tau^* = 0.30$, SE 0.11, Z-test $p = 2.4 \times 10^{-3}$). In addition, EpiNN was specifically enriched for heritability in this class of traits, with a 4.45× (SE 0.09, Z-test $p < 1.0 \times 10^{-100}$) higher enrichment observed in autoimmune diseases (3.53×, SE 0.10, Z-test $p < 1.0 \times 10^{-100}$) compared to non-immune system traits (0.80×, SE 0.03, Z-test $p = 1,3 \times 10^{-11}$).

4.5.1 Enrichment of heritability in EpiNN quantiles

I computed the proportion of heritability explained by each quantile (N=50) of the EpiNN annotation, which provides a more intuitive interpretation of the strength of the annotation's effect (see fig. 4.5). I observed that SNPs in the highest quantile of the annotation had an enrichment of 3.31 (s.e. = 0.44) for the heritability of meta-analysed immune-system traits, 4.2× more than SNPs in the smallest quantile (0.77, s.e. = 0.05).

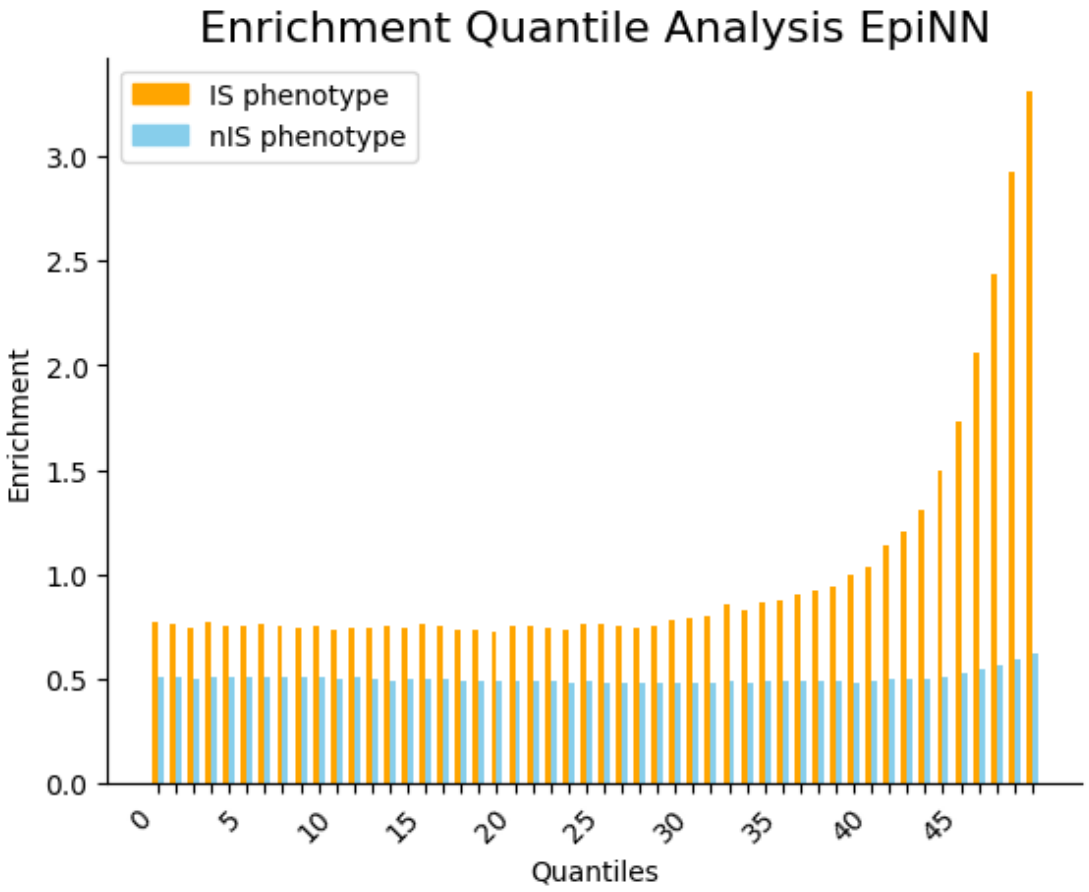


Figure 4.5: S-LDSC meta-analysed quantile. Enrichment analysis for the EpiNN annotation, divided into 50 quantiles. Meta-analysis for 5 IS and 58 nIS traits and diseases.

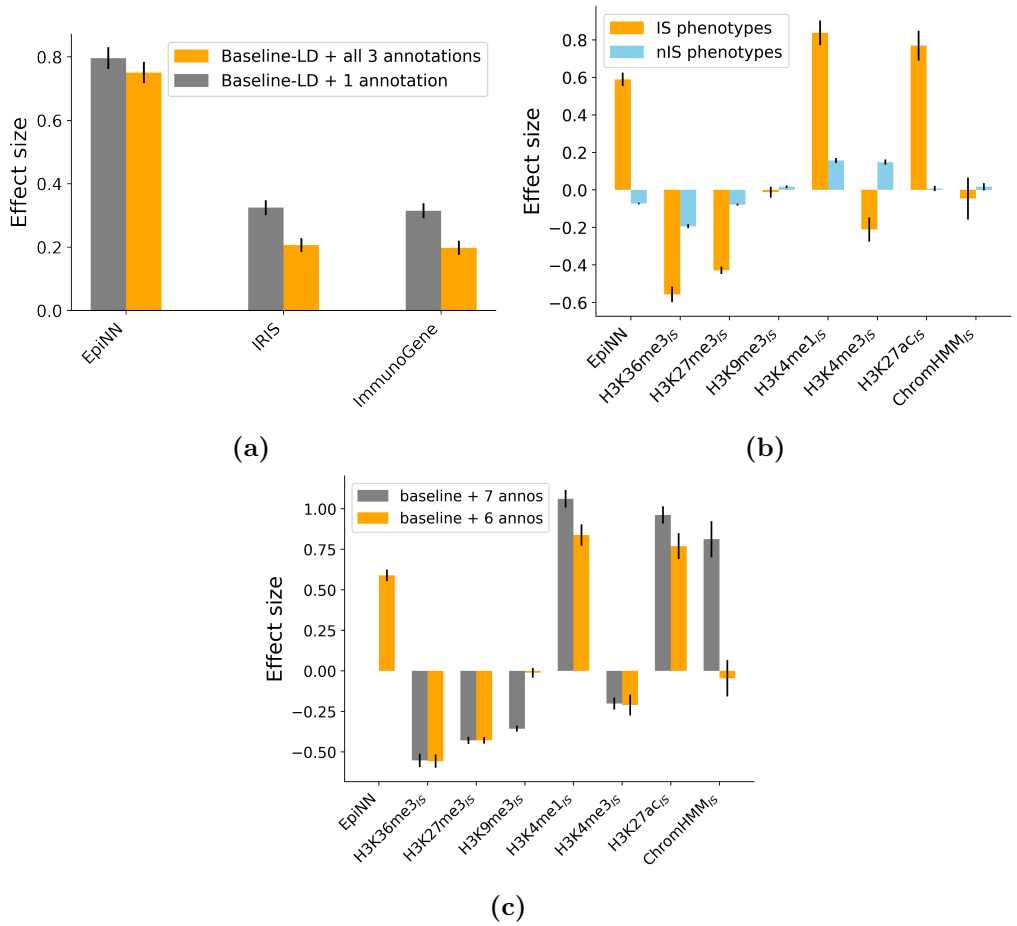


Figure 4.6: Joint S-LDSC analysis a. Effect size τ^* meta-analysis. All three annotations were jointly added into the model and also conditioned on 96 other Baseline-LD annotations. Grey bars represent effect sizes obtained when only conditioning on 96 other Baseline-LD annotations. **b.** Effect size τ^* for a S-LDSC analysis where EpiNN were jointly analysed with 96 other Baseline-LD annotations as well as annotations for IS-specific annotations related to EpiNN's input. **c.** Effect size τ^* for a S-LDSC analysis only for IS-specific phenotypes. Grey: joint analysis of 6 annotations with the Baseline-LD annotations. Orange: Joint analysis of 7 annotations with the Baseline-LD annotations. With the addition of EpiNN, the formerly strong ChromHMM$_{IS}$ signal is fully absorbed.

4.6 S-LDSC analysis including additional immune-related annotations

I used the same S-LDSC setup to test two other existing AIIS gene lists and compare them to EpiNN: Abbas et al [2] identified immune system relevant genes based on expression data, resulting in a list of 2,272 genes, which I refer to as "IRIS"; Giallourakis et al [3] used expressed sequence tag profiling based on transcriptional data, resulting in 2,232 potential immune system relevant genes,

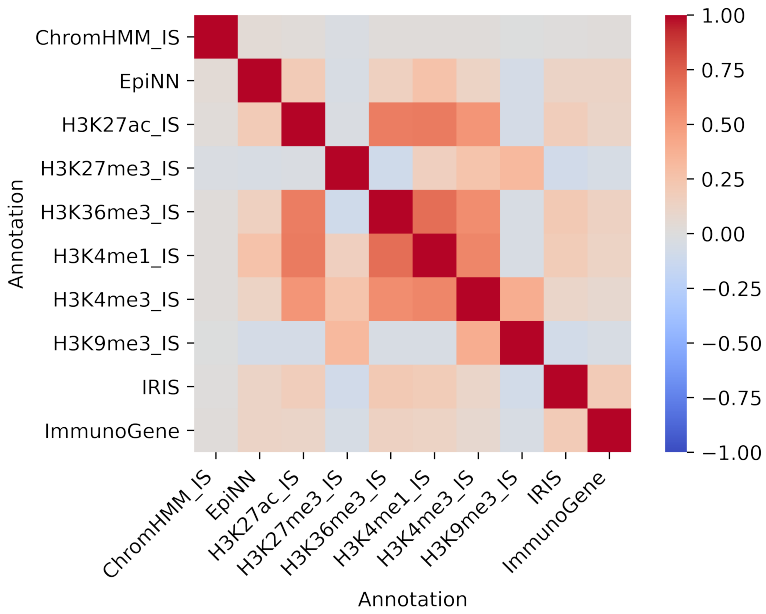


Figure 4.7: Correlation between S-LDSC annotations. Pearson correlation between EpiNN and several other annotations used in S-LDSC analyses. Numerical values are reported in table 4.3.

called "ImmunoGenes". I observed that, while significantly greater than 0, the marginal effect size τ^* of these annotations, meta-analysed across immune system traits was smaller than those observed for the EpiNN models (see fig. 4.6).

To test whether the EpiNN annotation provides orthogonal information to that contained in these expression-based annotations, I performed a joint S-LDSC analysis including EpiNN, IRIS, and ImmunoGenes in the same model. In this joint analysis, the heritability effect size of each annotation, conditional on the other two annotations, was slightly reduced but remained significant (see fig. 4.6a). The effect of the EpiNN annotation, in particular, remained large and significant ($\tau^* = 0.75$, SE 0.03), indicating that EpiNN provides novel information when conditioned on these expression-based models.

4.6.1 S-LDSC joint analysis with underlying epigenetic annotations

I additionally performed joint S-LDSC analyses of immune system-specific variants of the epigenetic annotations that are leveraged by EpiNN, to verify that EpiNN captures novel signal compared to these annotations. I created (see section 3.3)

	EpiNN	ChromHMM$_{IS}$	H3K27ac$_{IS}$	H3K27me3$_{IS}$	H3K36me3$_{IS}$
EpiNN	1	0.040	0.195	-0.041	0.155
ChromHMM$_{IS}$	0.040	1	0.027	-0.026	0.021
H3K27ac$_{IS}$	0.195	0.027	1	-0.028	0.632
H3K27me3$_{IS}$	-0.041	-0.026	-0.028	1	-0.098
H3K36me3$_{IS}$	0.155	0.021	0.632	-0.098	1
H3K4me1$_{IS}$	0.260	0.021	0.639	0.159	0.688
H3K4me3$_{IS}$	0.131	0.018	0.518	0.256	0.556
H3K9me3$_{IS}$	-0.063	-0.001	-0.070	0.326	-0.047
ImmunoGene	0.120	0.016	0.116	-0.048	0.144
IRIS	0.119	0.015	0.176	-0.089	0.215

	H3K4me1$_{IS}$	H3K4me3$_{IS}$	H3K9me3$_{IS}$	ImmunoGene	IRIS
EpiNN	0.260	0.131	-0.063	0.120	0.119
ChromHMM$_{IS}$	0.021	0.018	-0.001	0.016	0.015
H3K27ac$_{IS}$	0.639	0.518	-0.070	0.116	0.176
H3K27me3$_{IS}$	0.159	0.256	0.326	-0.048	-0.089
H3K36me3$_{IS}$	0.688	0.556	-0.047	0.144	0.215
H3K4me1$_{IS}$	1	0.594	-0.041	0.129	0.192
H3K4me3$_{IS}$	0.594	1	0.397	0.077	0.105
H3K9me3$_{IS}$	-0.041	0.397	1	-0.045	-0.082
ImmunoGene	0.129	0.077	-0.045	1	0.202
IRIS	0.192	0.105	-0.082	0.202	1

Table 4.3: Correlations between S-LDSC annotations. I provide numerical values for the correlations displayed in fig. 4.7

immune system-specific annotations of the 5 histone marks that are used in ChromHMM (H3K4me1$_{IS}$, H3K4me3$_{IS}$, H3K9me3$_{IS}$, H3K27me3$_{IS}$, H3K36me3$_{IS}$), an immune system-specific ChromHMM annotation (ChromHMM$_{IS}$), and an immune system-specific annotation of histone marker H3K27ac (H3K27ac$_{IS}$), which has been shown to harbour immune-related causal variation [14].

Several of these annotations were highly correlated with each other (e.g. $r = 0.69$ for H3K4me1$_{IS}$ and H3K36me3$_{IS}$, see table 4.3 and fig. 4.7), while the strongest correlation for EpiNN was observed for the H3K4me1$_{IS}$ annotation ($r = 0.26$). Interestingly, H3K4me1$_{IS}$ seems to mark a broad range of enhancers (percentage of SNPs covered 17.0%, number of SNPs over 0.75 certainty: 127173), while H3K27ac$_{IS}$ also covers a broad range of SNPs (19.4%) but has a certainty over 0.75 for only 22 of those (see table 3.5 and table 4.4). EpiNN captures 3713 of the H3K4me1$_{IS}$ SNPs and all 22 of H3K27ac$_{IS}$. As seen in table 4.4, it also covers a subset of SNPs from the other histone markers, meaning that EpiNN marks not

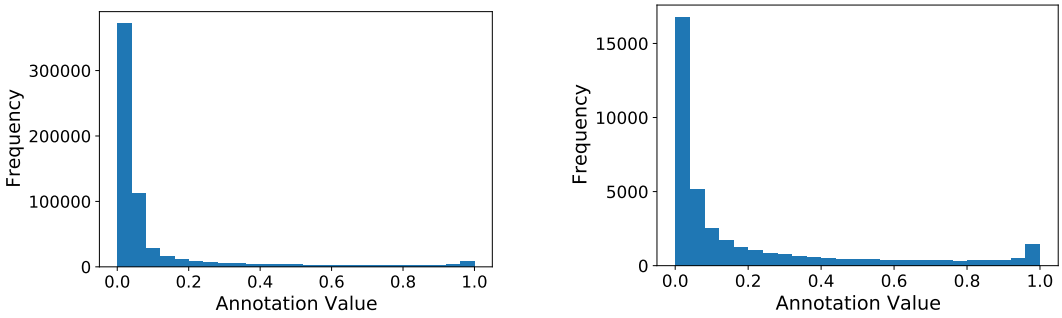
		EpiNN		
		in annotation	not in annotation	total # of SNPs
H3K36me3$_{IS}$ binarised	in annotation	94,199	934,148	1,028,347
	not in annotation	296,123	8,760,980	
H3K4me1$_{IS}$ binarised	in annotation	47,015	80,158	127,173
	not in annotation	343,307	9,567,786	
H3K9me3$_{IS}$ binarised	in annotation	1,660	113,184	114,844
	not in annotation	388,662	9,489,405	
H3K27me3$_{IS}$ binarised	in annotation	26,257	1,009,975	1,036,232
	not in annotation	364,065	8,617,211	
H3K4me3$_{IS}$ binarised	in annotation	3,713	12,942	16,655
	not in annotation	386,609	9,591,700	
H3K27ac$_{IS}$ binarised	in annotation	22	0	22
	not in annotation	390,300	9,600,951	
ChromHMM$_{IS}$ binarised	in annotation	353	150	503
	not in annotation	389,969	9,601,132	
IRIS	in annotation	49,090	241,685	290,775
	not in annotation	341,232	9,408,334	
ImmunoGene	in annotation	35,277	132,740	168,017
	not in annotation	355,045	9,503,466	
	total # of SNPs	390,322		

Table 4.4: Overlap between binarised annotations. Annotations were binarised using a threshold of 0.75 (as done for EpiNN). "in annotation" refers to the number of SNPs present in the binary annotation; "not in annotation": refers to the number of SNPs not present in the binary annotation.

only one functional element as these histone annotations do, but catches the most important elements throughout the genome. Histograms with the distribution of the annotation values can be found in fig. 4.8.

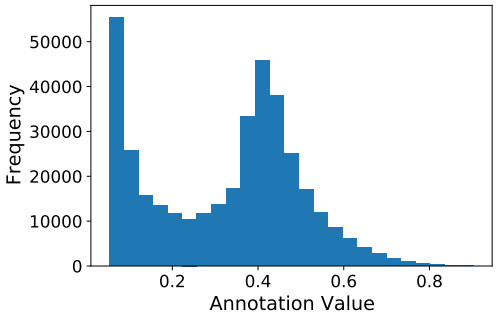
I also built binarised versions of these annotations by thresholding at a value of 0.75 and computed the amount of overlap with the EpiNN annotation (see table 4.4). The EpiNN annotation was found to overlap with all of these binarised annotations, suggesting that EpiNN does not closely reflect a single underlying annotation, but rather learns to combine signal from these diverse functional elements.

In a joint S-LDSC analysis including these 7 annotations (see fig. 4.6b and table 3.5), EpiNN retained a significant heritability effect ($\tau^* = 0.58$, SE 0.036, Z-test $p = 4.2 \times 10^{-60}$), indicating that EpiNN's combined signal is complementary to that provided by the other immune-specific functional annotations I considered. H3K4me1$_{IS}$ and H3K27ac$_{IS}$ also had significant and large effects. This was expected, as these annotations mark enhancers, which have been shown to harbour a large fraction of fine-mapped variants in autoimmune diseases [14]. The immune-specific

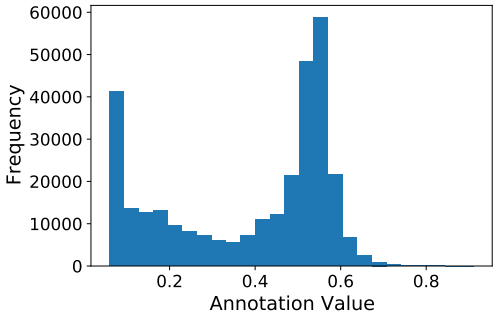


(a) EpiNN annotation

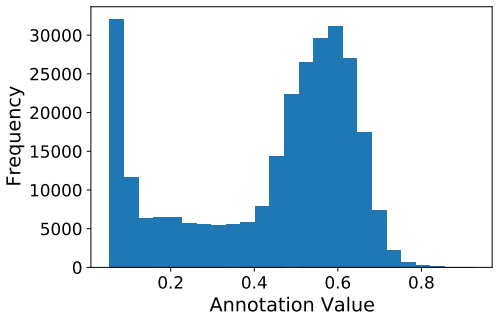
(b) EpiNN gene predictions



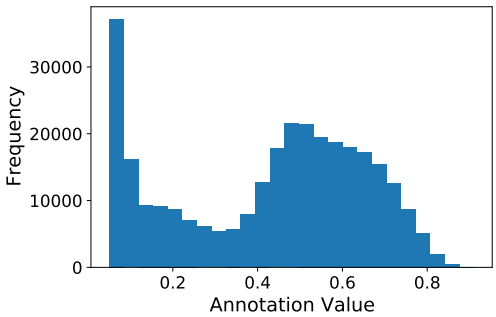
(c) H3K4me1$_{IS}$ annotation



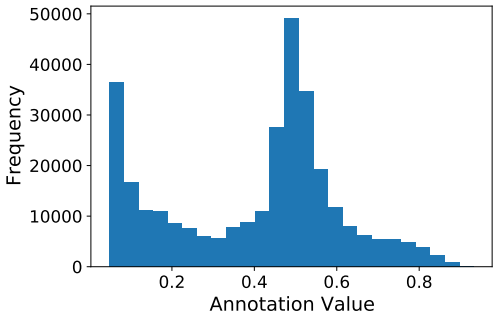
(d) H3K4me3$_{IS}$ annotation



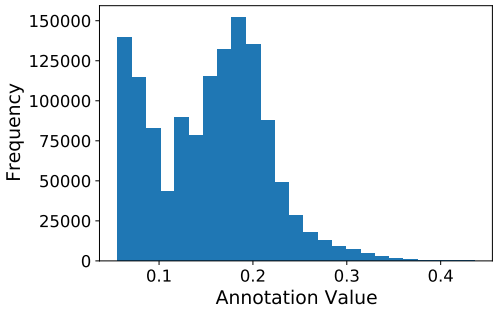
(e) H3K9me3$_{IS}$ annotation



(f) H3K27me3$_{IS}$ annotation



(g) H3K36me3$_{IS}$ annotation



(h) H3K27ac$_{IS}$ annotation

Figure 4.8: Annotation Histograms: Histograms of the distribution of annotation values for different annotations. For EpiNN, there is a clear separation between positive (1) and negative (0) predictions, while for the histone marks, intermediate values persist.

ChromHMM$_{IS}$, which also should capture multiple functional elements, did not yield a significant effect, indicating that EpiNN and the other annotations present in the model better represent the signal provided by this annotation. Finally, to verify that these heritability effects were not significantly affected by the presence of highly correlated annotations in the joint model, I repeated this analysis, this time following a forward-stepwise elimination procedure [27, 129] (see chapter 3). In this analysis, the EpiNN, H3K3me1$_{IS}$, H3K4m3$_{IS}$, H3K27me3$_{IS}$, H3K36me3$_{IS}$, and H3K27ac$_{IS}$ annotations were again retained as significant, with heritability effects similar to those observed in fig. 4.6b.

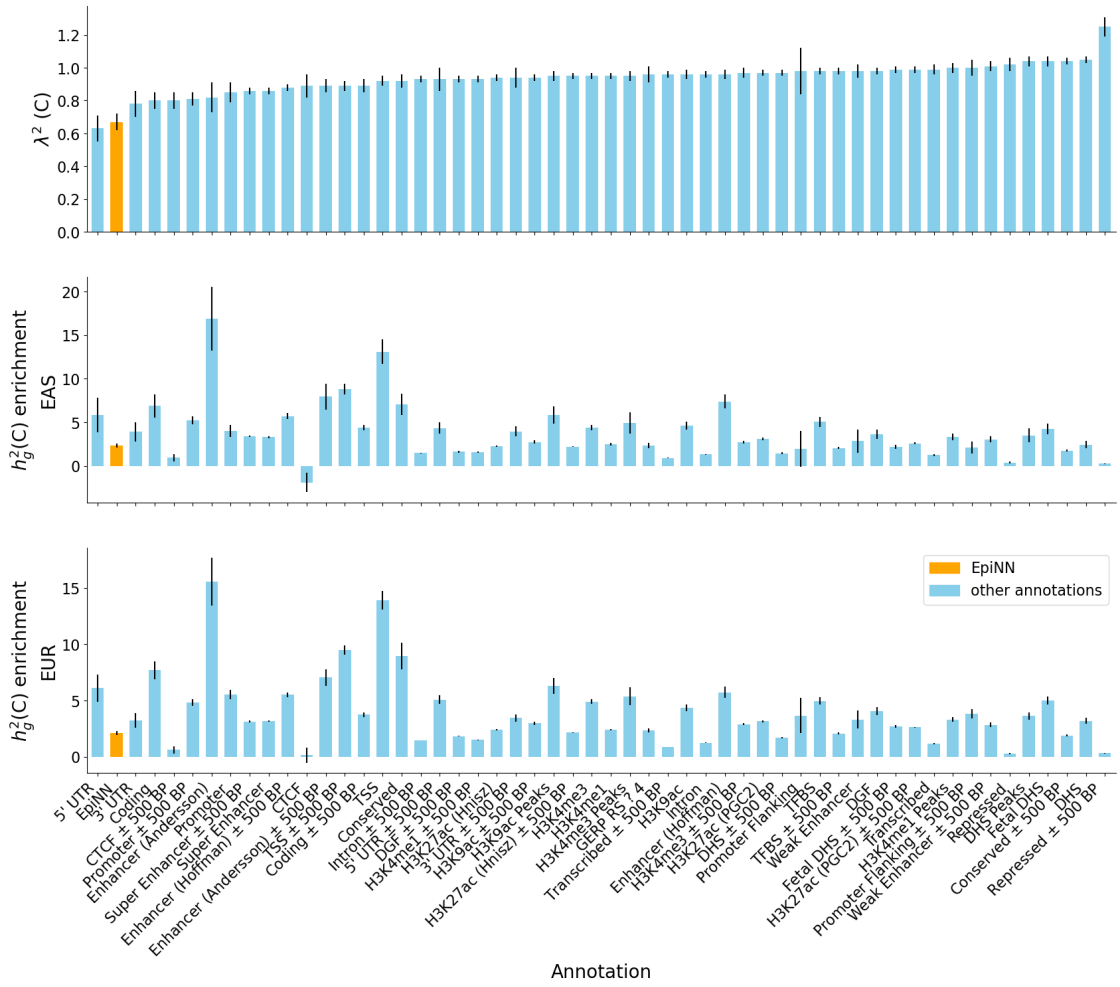
4.7 S-LDXR analysis

This analysis was performed in collaboration with Steven Gazal.

	Trait	N_{ASI}	N_{EUR}
BASO	Basophil Count	62,076 [137]	131,860 [138]
EO	Eosinophil Count	62,076 [137]	337,539 [139]
HBA1C	Hemoglobin A1c	42,790 [137]	337,539 [139]
HGB	Hemoglobin	108,769 [137]	132,596 [138]
HTC	Hematocrit	108,757 [137]	132,699 [138]
LYMPH	Lymphocyte Count	62,076 [137]	337,539 [139]
MCH	Mean Corpuscular Hemoglobin	108,054 [137]	337,539 [139]
MCHC	MCH Concentration	108,728 [137]	132,586 [138]
MCV	Mean Corpuscular Volume	108,256 [137]	132,353 [138]
MONO	Monocyte Count	62,076 [137]	337,539 [139]
NEUT	Neutrophil Count	62,076 [137]	131,564 [138]
PLT	Platelet Count	108,208 [137]	337,539 [139]
RA	Rheumatoid Arthritis	22,343 [5]	37,598 [5]
RBC	Red Blood Cell Count	108,794 [137]	337,539 [139]
WBC	White Blood Cell Count	107,964 [137]	337,539 [139]

Table 4.5: S-LDXR immune/blood traits. We report the sample size of the 15 immune traits for ASI and EUR ancestries that were meta-analysed for the S-LDXR analysis.

Regulatory effects and variant effect sizes might differ across human populations, especially for variants within genomic regions that are under stronger gene-environment interactions, such as immune-related annotations [130]. We therefore



(a)

Figure 4.9: S-LDXR analysis results. S-LDXR results for 54 functional annotations across 15 immune diseases and complex traits. **a.** Enrichment/depletion of squared trans-ethnic genetic correlation for annotation C (λ^2(C)). **b.-c.** Population-specific estimates of heritability enrichment for each annotation.

tested whether immune-related regions implicated by EpiNN were depleted for trans-ancestry genetic correlation. To this end, we applied S-LDXR [130] on 15 diseases and complex traits (see table 4.5) with GWAS data from East Asian (mean $N = 83K$) and European (mean $N = 235K$) populations (see chapter 3, fig. 4.9 and table B.5). We found that variants within the EpiNN annotation were similarly enriched in both populations (East Asian population enrichment: 2.32×, SE 0.22; European population enrichment: 2.11×, SE 0.16), suggesting that the EpiNN annotation may be used to effectively predict AIIS-relevant regions in non-European ancestries. At the same time, the EpiNN annotation was the second most

significantly depleted for squared trans-ancestry genetic correlation among 54 tested annotations (0.67×, SE 0.05, $p = 3.9 \times 10^{-10}$). This depletion indicates the presence of underlying population-specific causal effect sizes, which are likely due to stronger gene-environment interactions at immune-related loci with large EpiNN scores.

4.8 Application of EpiNN to non-immune system genes

While immune system-relevant genes are of particular interest for drug development, EpiNN, can be applied to other gene categories as well. As previously mentioned, I applied it to housekeeping genes, but also to epithelial-specific genes, and brain-specific genes.

4.8.1 Housekeeping genes

Housekeeping genes, or constitutively expressed genes, are genes that are relevant in almost all cell types and often code for processes that are necessary to keep a cell alive, independent of that cell's function. To validate the function of the EpiNN model, I applied the same procedure described for AIIS genes to housekeeping genes. The training set for this is task was a well-curated published list of 3754 housekeeping genes [140]. Re-ordering of cells was not necessary for this use case, as it was expected that for housekeeping genes, all or nearly all cell types would behave similarly. The training metrics are slightly better than the training metrics of the AIIS gene prediction (see table 3.1 for details). That is most likely due to two major factors: The first being the bigger size of the training set and the second the less complex task: Finding a signal that is common between all cell types rather than a signal that differs between AIIS and non-AIIS cell types.

4.8.2 Epithelial genes

I further applied EpiNN to epithelial-specific genes. The training list came from an expression-based study [141] and contained only 268 genes, whose coordinates were generated using ensemble biomart, and which I preprocessed by removing

unconventional chromosomes (for which epigenetic data was not available). This left 241 training genes.

There are 8 different types of epithelial cells in the ChromHMM data set (2x foreskin fibroblast, 2x foreskin melanocyte, 2x foreskin keratinocyte, breast vHMEC mammary epithelial, breast myoepithelial) which is a much smaller number than immune cells. Re-ordering of cell types was not necessary of this use case either, as these cell types were already aggregated in a single block in the original ordering.

When inspecting the training images, it became clear that very few images showed skin cell specificity in the ChromHMM data and the signal was often only found in a subcategory of the different skin cell types (often keratinocytes rather than fibroblasts). This meant that the training images had relatively poor quality, mostly due to the low number of skin cell types.

I trained the model despite the low data quality, resulting in the following metrics with a threshold of 0.5:

- PRAUC: 0.702083
- Accuracy: 0.718182
- Precision: 0.652622
- Recall: 0.702083

These are poor training results compared to the AIIS EpiNN model, indicating that the model did not learn to discern well between epithelial-specific and other genes. Nevertheless, I still applied the trained model to the whole human genome using sliding windows which resulted in a genome-wide annotation for regions marking epithelial-specific regions. I applied the same S-LDSC analyses as before but none of the 63 complex traits had significantly enriched heritability, which is in line with my low confidence in the model based on the poor training performance.

4.8.3 Brain genes

I took a similar approach for brain-specific genes which might be interesting for neurological phenotypes. In this case, the training data came from the Protein Atlas [142], searching for enriched expression in brain tissues (query: *tissue_category_rna:brain; Tissue enriched, Group enriched, Tissue enhanced AND sort_by:tissue specific score*). The coordinates of these genes were again found using ensemble biomart and preprocessed by removing genes on unconventional chromosomes, and genes too close to the chromosome ends, leaving 2,498 training genes.

Similar to the epithelial genes, the specificity of the brain genes was not very high, when I visually inspected the ChromHMM tracks. Re-ordering of cell types to increase the signal was not possible in this case either, as the relevant cell types were already aggregated in a single block in the original ordering. Still, because of the order of magnitude difference in available training data, it had the potential to be a better learning task.

Unfortunately, training EpiNN on brain-specific genes also resulted in unpromising performance metrics compared to the AIIS EpiNN model:

- PRAUC: 0.742583
- Accuracy: 0.746923
- Precision: 0.689932
- Recall: 0.742583

Similarly to above, I still applied the model to the genome and used S-LDSC to analyse the results, again obtaining a null result, which was to be expected from the low training metrics.

4.9 Testing the AIIS EpiNN annotation for heritability enrichment in COVID19 infections

With the AIIS EpiNN annotation labelling immune-specific regions in the human genome, and COVID19 infections being prevalent throughout the time of this thesis, I tested whether the EpiNN annotation was predictive for COVID19 infections, based on data from the COVID-19 Host Genetics Initiative [143]. I calculated S-LDSC enrichment and effect size for data releases 5 (January 18, 2021) and 6 (June 15, 2021). Different types of COVID19 infection traits were available: very severe respiratory confirmed COVID19 vs population hospitalised COVID19 vs not hospitalised COVID19, hospitalised COVID19 vs population, and COVID19 vs population. Data was gathered either by 23andme or 23andme and the UKBB. I discarded the data sets for filtered and European-only data due to the small sample size.

Unfortunately, the EpiNN annotation was not informative for the heritability of any of the degrees of COVID19 severity in any of the data releases (see fig. 4.10). This indicates that the loci found by EpiNN are neither predictive for general COVID19 infection nor for the severity of a COVID19 infection.

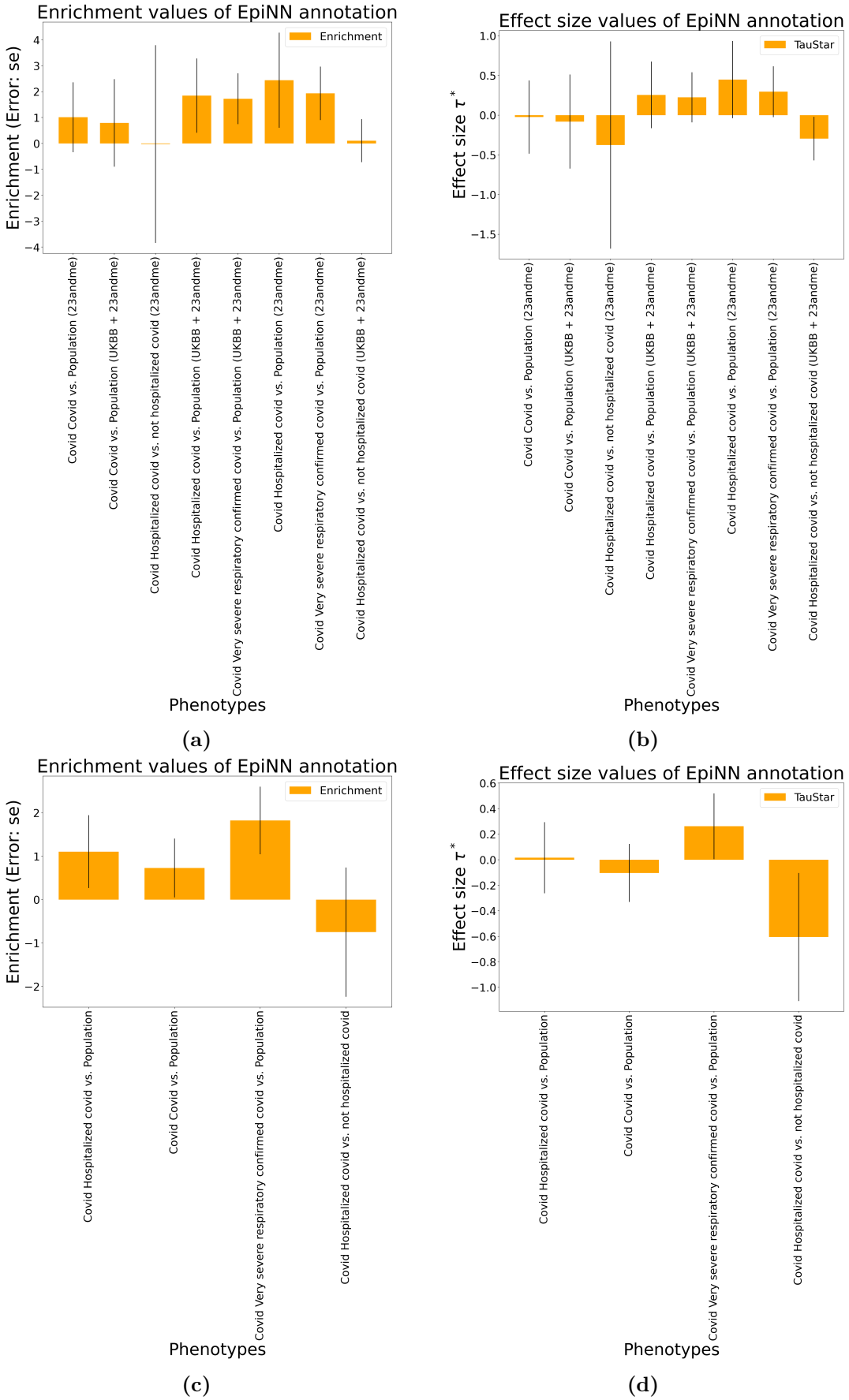


Figure 4.10: S-LDSC on COVID19 data: (a) and (b): Release 5 (c) and (d): Release 6. Neither release shows heritability enrichment for the EpiNN annotation for any degrees of COVID19 severity.

5

Biological Analysis of EpiNN Results

Contents

5.1 Protein-coding gene examples 69
5.1.1 *PIK3CD* 69
5.1.2 *IL2RB* 69
5.1.3 *TAGAP* 72
5.2 long non-coding RNAs 73
5.2.1 *Linc00402* 73
5.3 microRNAs 74
5.4 *DNMT1* 74
5.4.1 CAGEseq and primary sequence analysis 77
5.4.2 Western Blot 77
5.4.3 PCR 79
5.4.4 DNMT1 structure analysis 80
5.4.5 *DNMT1* conclusions 82
5.5 Additional EpiNN detections 82

This chapter presents biological findings of the predicted AIIS genes by highlighting predictions of known immune system-relevant genes as positive control (*PIK3CD*) as well as novel findings in *IL2RB*, *TAGAP*, and *DNMT1*.

To better understand the genes predicted by EpiNN, I first looked into how they are distributed over different gene categories based on their names. Out of 3471 total genes, there are:

- 784 LOC (uncharacterised genes)

- 122 LINC (Long Intergenic Non-Protein Coding)
- 86 MIR (mircoRNA)
- 45 SLC genes
- 44 CD genes
- 30 FAM genes

They are distributed evenly over the whole genome, including but in no way limited to the HLA region, as can be seen in the karyogram in fig. 5.1, indicating that genes that are relevant for the immune system are spread throughout the genome.

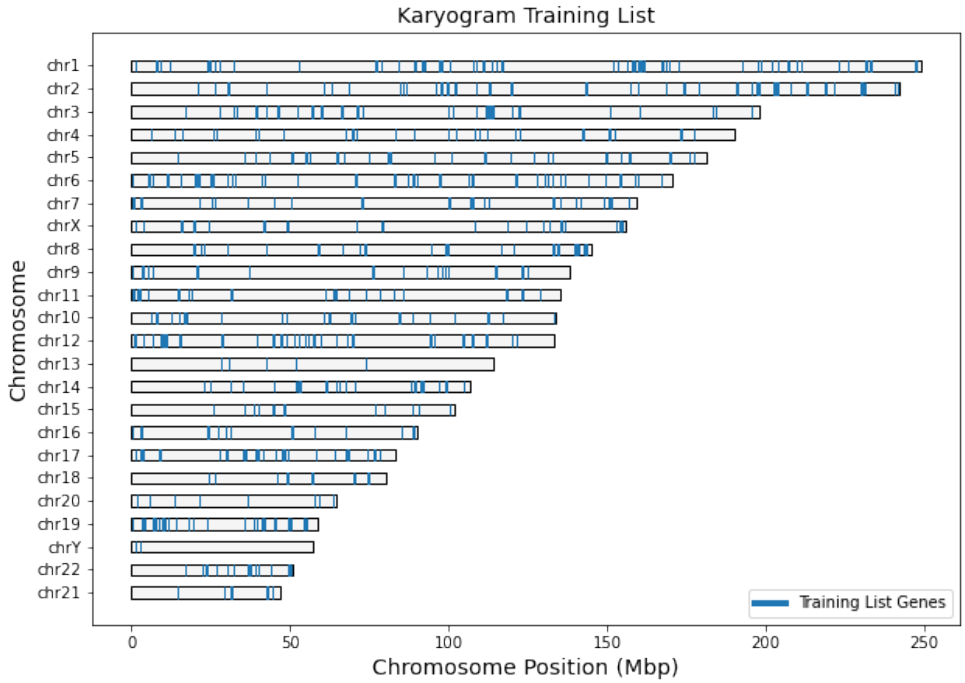
5.1 Protein-coding gene examples

5.1.1 *PIK3CD*

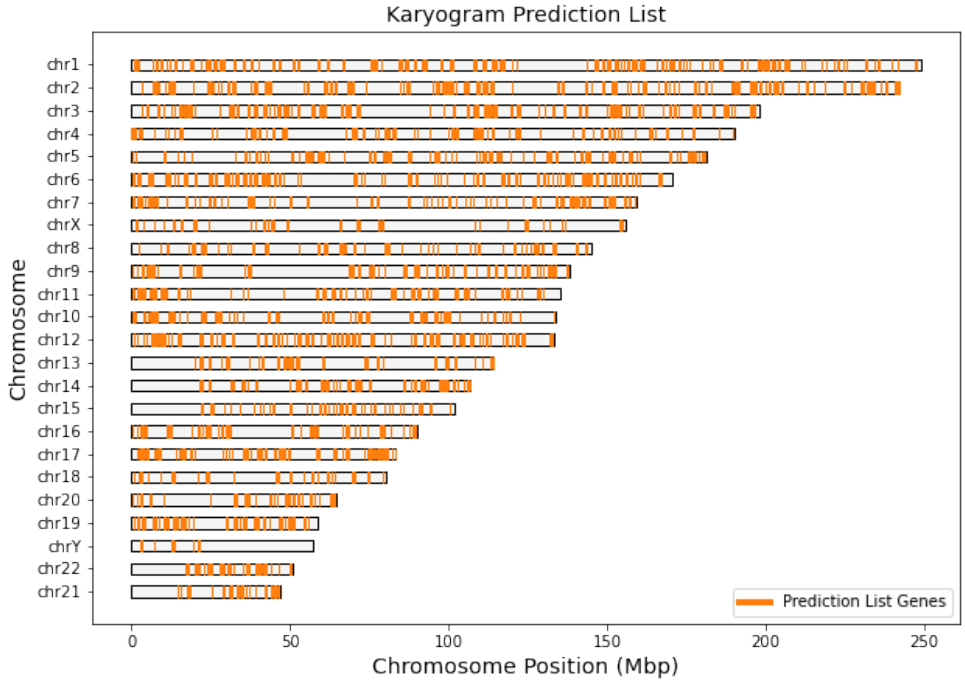
PIK3CD (see. fig. 5.2a) is a known AIIS gene [144], which was not part of the training list and therefore a good validation for the prediction list. *PIK3CD* is known to be involved in the immune response. It codes for the p110 delta subunit of Phosphoinositide 3-kinase, which phosphorylate inositol lipids and are found primarily in leukocytes. Diseases associated with *PIK3CD* include Immunodeficiency and Roifman-Chitayat Syndrome. Among its related pathways are Gastric cancer and Human cytomegalovirus infection [145].

5.1.2 *IL2RB*

IL2RB is an immune system-relevant gene, encoding a subunit of the interleukin 2 receptor, which is involved in T cell-mediated immune response. The novel finding is an upstream enhancer element (“1”) (see fig. 5.2b) targeted by e.g. transcription factors *AP-1*, *FOXM1*, and *TBX21*. Analysing Hi-C data from CD4+ T-cells shows changes between non-activated and activated cells concerning these enhancers: In non-activated CD4+ cells only enhancers “1” and “2” are connected. In activated CD4+ cells there are the following connections: “1”-“2”, “1”-“3”, “1”-“G”, “2”-“3”.

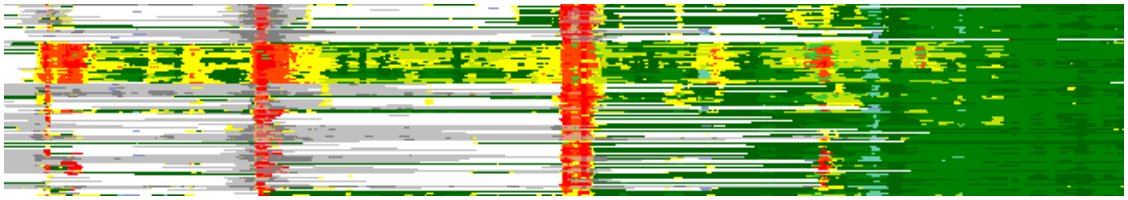


(a) Training data.

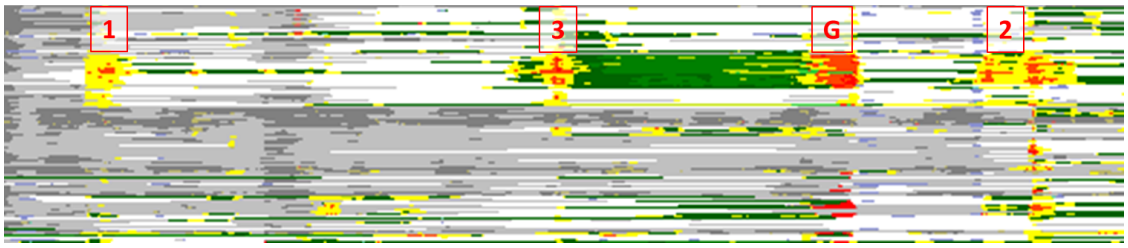


(b) Predictions.

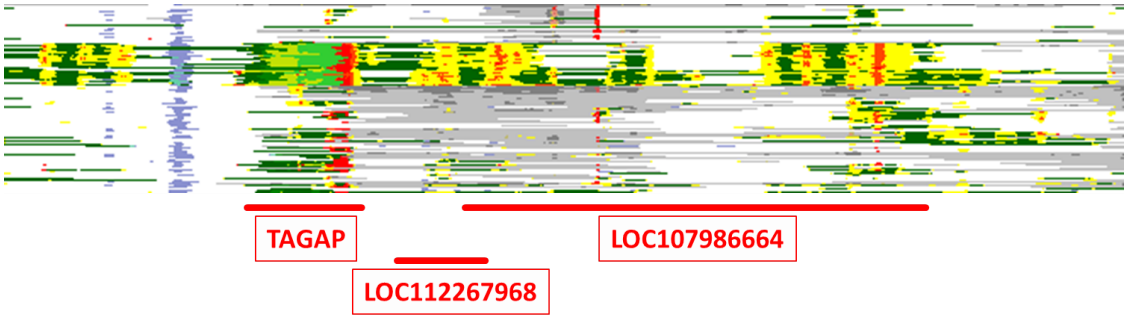
Figure 5.1: Karyogram of training genes and predicted genes. Chromosomes are ordered by size. Predicted genes are distributed over all chromosomes.



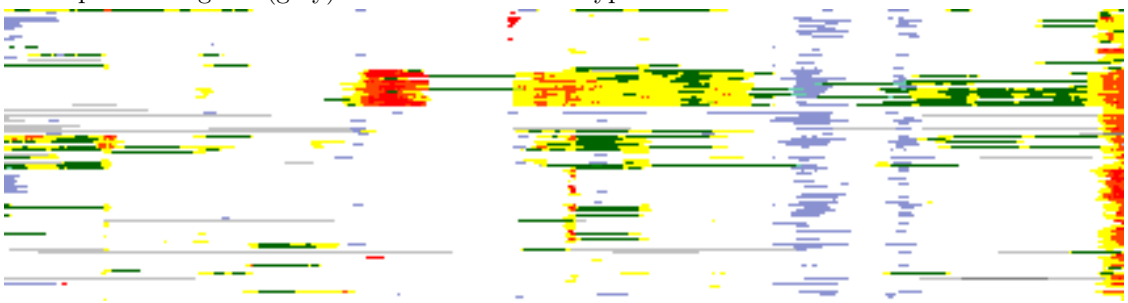
(a) Epigenetic signature of *PIK3CD* gene, which is a validation example of a positively identified AIIS gene.



(b) Epigenetic profile of *IL2RB* with transcription start site "G" as well as enhancers "1", "2", and "3".



(c) Epigenetic profile of *TAGAP* and neighbouring ncRNA loci *LOC112267968* and *LOC107986664*. The epigenetic pattern shows clustering of transcription elongation (green) in immune system cells and repression signals (grey) in most other tissue types.



(d) Epigenetic signature of *Linc00402*, which was very recently experimentally validated to be immune system-relevant[136].

Figure 5.2: Example genes predicted by EPiNN. Displayed are the original ChromHHM tracks.

Since this is an ensemble measurement, not all these connections must be present in each cell at the same time [146]. In addition, "1" and "G" bind transcription factor *TBX21*. "2" and "3" bind transcription factor *GATA3* [147], both of which are involved in Naive CD4+ T-cell activation. Based on this I propose that the novel enhancer "1" is involved in CD4+ activation by tethering a cohesin loop in a tissue-specific manner: In naive CD4+ T cells, the super-enhancer "2" always binds distal enhancer "3". *IL2RB* is expressed but on a rather low level. Upon activation, the following changes occur:

(A) *IL2RB* promoter "G" connects with distal enhancer "1", while *IL2RB* expression increases by a magnitude

(B) *IL2RB* transcription termination site "3" connects with super-enhancer "2"

(C) *IL2RB* transcription termination site "3" also connects with distal enhancer "1", while a promoter at "1" is activated, leading to antisense transcription towards and until "3." (The resulting transcript could be AL022314.1.) Moreover, Epigenome Roadmap shows the silencing of this antisense transcript in naive CD4+ cells via Polycomb, i.e. H3K27 methylation, which can be removed by UTX.

Rearrangements (A) and (C) above will potentially shift pre-bound transcription factors (TFs) from enhancers to promoters and hence initiate/increase transcription levels. This hypothesis remains experimentally untested at this point.

5.1.3 *TAGAP*

TAGAP is a known immune system-relevant gene. It encodes the T-cell activation Rho GTPase-activating protein. It is also implicated in the predisposition to immune-related adverse events upon CPI treatment [148].

The novel finding are two ncRNA loci *LOC112267968* and *LOC107986664* (see fig. 5.2c), two long non-coding genes near *TAGAP* which show an immune cell-specific epigenetic pattern, as well as hotspots for GWAS hits for autoimmune diseases and white blood cell levels [149]. For *LOC112267968* these are rs2451258

(phenotypes: ankylosing spondylitis, psoriasis, ulcerative colitis, Crohn's disease, sclerosing cholangitis), rs2451279 (neutrophil count, basophil count, myeloid white cell count) and rs2249937 (psoriasis vulgaris) [145], which are all auto-immune diseases or immune system-related. For *LOC107986664*, this is rs182429, which has been associated to *TAGAP* and rheumatoid arthritis in Pakistani populations [150] but has actually been fine-mapped to *LOC107986664* [151], and rs1738074, which using GWAS has been to celiac disease [152] and multiple sclerosis risk [153]. These GWAS hits are usually attributed to *TAGAP* even though they fall outside the gene. But the finding of the immune-specificity of *LOC112267968* and *LOC107986664* in the same cells as *TAGAP* opens up a possible mechanism, where the GWAS hits modulate *LOC112267968* and *LOC107986664* which modulate *TAGAP*, rather than *TAGAP* directly. This hypothesis is so far untested.

5.2 long non-coding RNAs

EpiNN predicts in total 122 long non-coding RNAs to be immune system-relevant. This is interesting as they don't code for protein directly but might regulate the expression of other genes, therefore having a regulatory effect. To this end, lncRNAs don't need to be expressed in large quantities and therefore often fall under the detection threshold in expression experiments. The EpiNN results indicate that epigenetic signals might be an alternative basis for identifying and categorising lncRNAs into functional categories. An example of an identified long non-coding RNA by EpiNN is *Linc00402*.

5.2.1 *Linc00402*

Linc00402 was predicted by EpiNN to be immune system-relevant. In the epigenetic data, it shows a very clear signal for immune system cell types (see fig. 5.2d). *Linc00402* has only recently been linked to T cell alloimmunity [136] using RNA-seq of human T cells after hematopoietic stem cell transplantation, which identifies the noncoding *Linc00402* as a regulator of T cell alloimmunity. This is a good validation of the predicted genes and underlines the potential importance of the

many non-coding genes found in the predictions for the human immune system. Similar to *Linc00402*, many of these still have to be investigated experimentally.

5.3 microRNAs

Similar to long non-coding RNAs, microRNAs are difficult to capture using expression-based methods, mostly due to many microRNAs losing their poly-A tail very early on in their lifespan and therefore not being detected using transcription-based approaches, which rely on the poly-A tail for detection. But similarly to long non-coding RNAs, they can in small quantities have large regulatory effects, therefore potentially having a big influence on phenotypes. EpiNN predicts 86 microRNAs as immune system-relevant, indicating again, that epigenetic data might be useful in identifying regulatory loci and genes.

5.4 *DNMT1*

Another novel finding is a transcription start site in *DNMT1* (human DNA methyltransferase 1) showing distinct epigenetic regulation patterns specific to immune cells. Experimental evidence (PCR and Western Blot) for its existence in immune cells has been collected together with Calliope Dendrou's research group. I did the initial study design and decided on experiments to run and performed the initial research into antibodies, while Calliope Dendrou's research group refined the study design and details and performed the PCR and Wastern Blot experiments.

DNMT1 encodes a constitutively expressed epigenetic agent that transfers methyl groups to cytosine nucleotides of genomic DNA [145]. This region exhibits a distinct pattern of epigenetic regulation, specific to immune cells. Epigenome Roadmap data shows homogeneous epigenetic states across the gene body of *DNMT1* in other cell types, as well as a previously unreported immune-cell-specific, promoter-like signature in the intron 28 (see fig. 5.3). Assuming the locus in intron 28 would lead to an alternative transcription start, the resulting DNMT1 isoform could lose its autoinhibitory functions, since the isoform would lack the RFTS, CXXC and

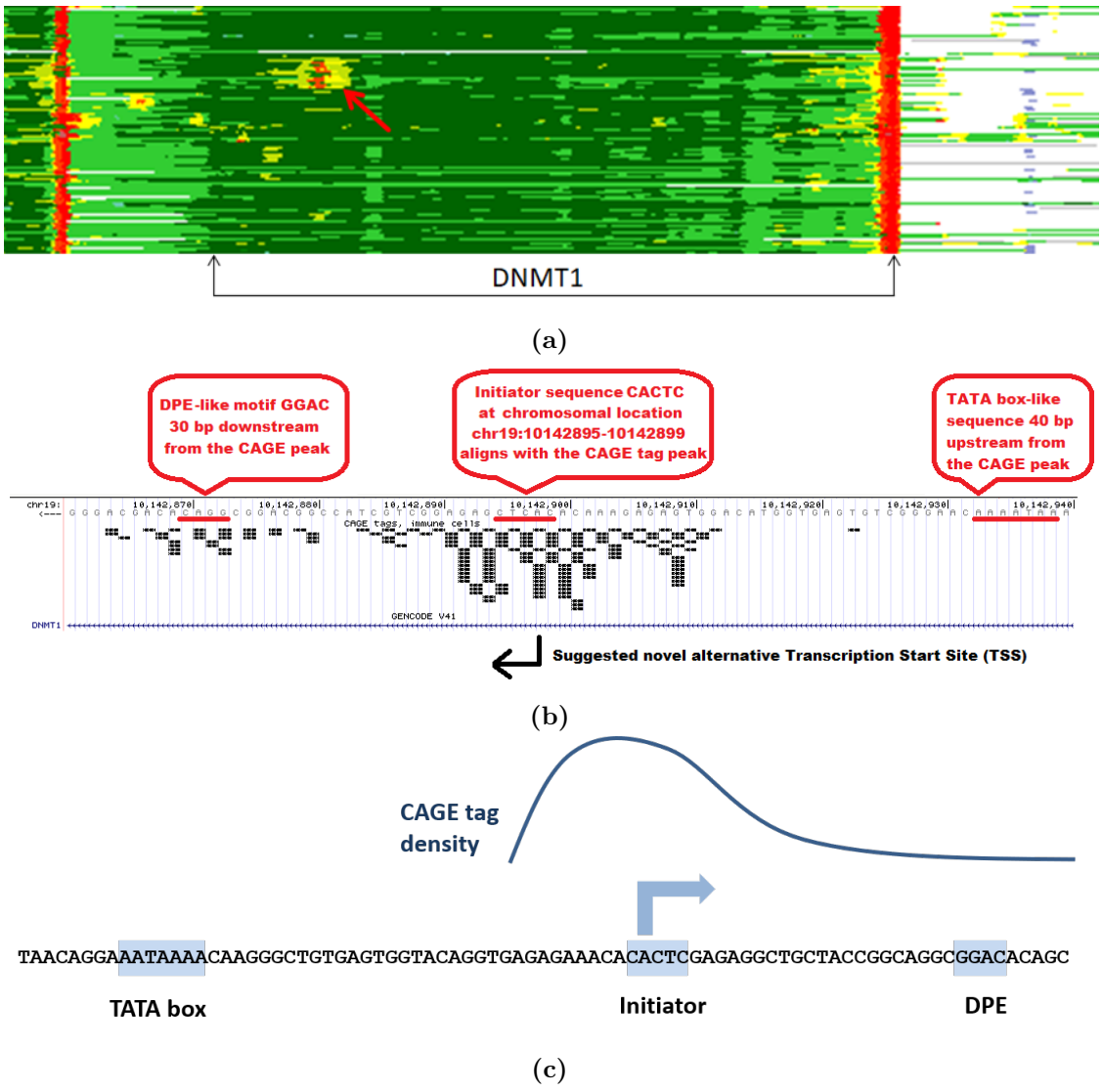
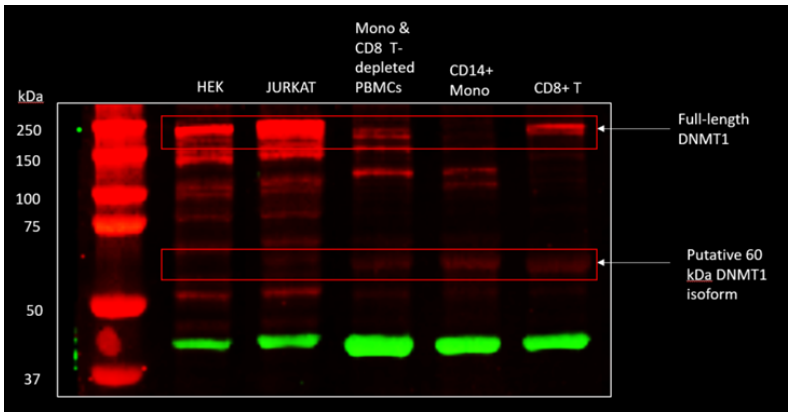
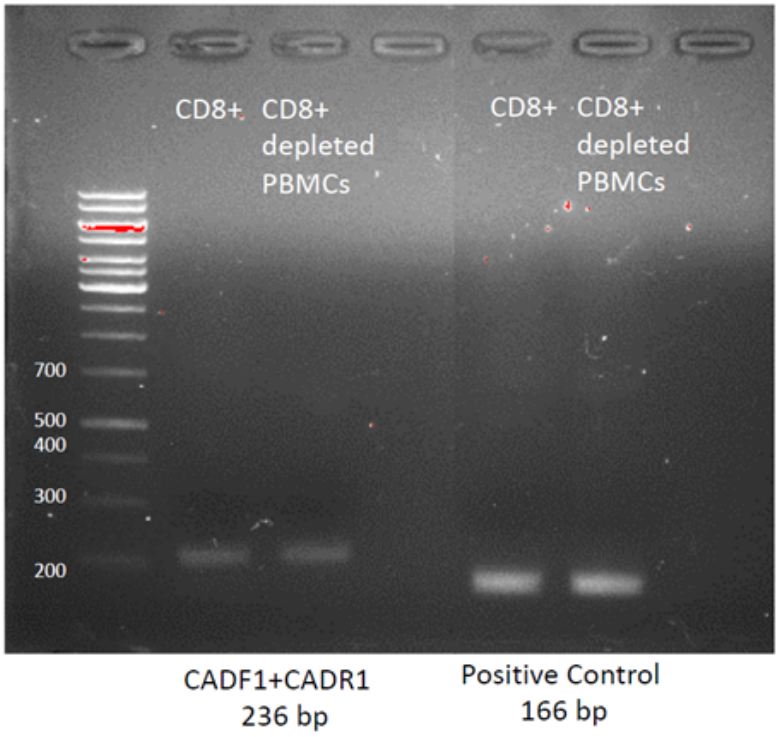


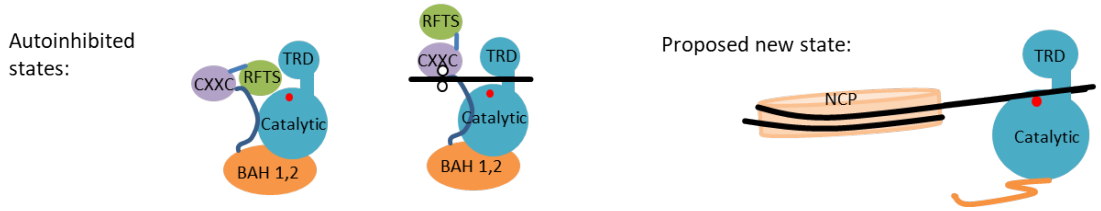
Figure 5.3: In silico AIIS-relevance for the *DNMT1* gene. a. Epigenetic signal of alternative transcription start site predicted by EpiNN. **b.** The *DNMT1* intron shows a peak of CAGE tags in CD4+ and CD8+ T cells, as well as three incomplete but telltale promoter sequence motifs: Initiator, Downstream Promoter Element (DPE), and TATA box [154]. Displayed as tracks on the UCSC Genome Browser with GRCh38 coordinates; gene in the minus strand. **c.** Schematic diagram of the observed CAGE tags in immune cells (raw data in table B.6) to clarify CAGE tag signal at initiator site.



(a)



(b)



(c)

Figure 5.4: Evidence of AIIS-relevance for the *DNMT1* gene and proposed mechanism. a. Western Blot from primary human CD8+ cells shows both the full-length 190 kDa DNMT1 protein, as well as a truncated 60 kDa protein, with size matching that of the predicted new isoform. **b.** PCR of CD8+ and CD8+-depleted PBMCs show bands of expected size for truncated *DNMT1* transcript. **c.** Proposed mechanism: removal of both auto-inhibitory domains (RFTS and CXXC) may lead to unchecked DNA methylation by the truncated protein isoform.

BAH1,2 domains. Such a shorter isoform of DNMT1 could perform unchecked methylation, with ramifications for the negative epigenetic regulation of methylated promoters in immune cells (see fig. 5.4c).

The putative alternative transcription start might result in a new protein starting at position MET 1077 of the full-length DNMT1. The proposed amino acid chain is shown in Appendix B.

5.4.1 CAGEseq and primary sequence analysis

Bioinformatic analyses of other data sets support the finding: FANTOM5 CAGE-seq data [133, 155] shows a moderate signal at the new TSS in CD8+ T cells, CD4+/CD25-/CD45RA- memory T cells, and a weaker one in CD4+/CD25+/CD45RA+ naive regulatory T cells (see fig. 5.3b and table 5.1). After deducting the exact location of the candidate TSS from the peak in the CAGE data, one can observe that the primary sequence shows a TATA box at position -41..-35, an initiator consensus sequence at -1..+4, and a downstream promoter element (DPE) site at +26..+35 with respect to the hypothesised transcription start site.

5.4.2 Western Blot

We performed a Western Blot to identify a putative 60kDa DNMT1 isoform in human immune system cells (see fig. 5.4a).

The western blotting method was developed in 1979 [156] and has since been used as a stable technique to separate and identify proteins. There are multiple steps involved, which, on a high level are: sample preparation, gel electrophoresis, transfer, total protein staining, blocking, incubation and detection. In the sample preparation step, proteins are isolated from a cell mixture and mixed with loading buffer. In the second step, this protein mixture is separated by the proteins' molecular mass. This is achieved by using sodium dodecyl sulfate (SDS) buffers which denature proteins and covers them in negatively charged SDS. By loading a polyacrylamide gel and applying an electrical field, the proteins will migrate to the positively charged anode and separate. In the transfer step, the separated proteins are transferred onto a

FANTOM5 cell type	CAGEseq evidence of short *DNMT1* isoform
CD14-/CD16+ Monocytes	-
Dendritic Cells monocyte immature derived	-
CD19+ B Cells pluriselect	-
CD34+ Progenitors donor 1	-
CD34+ Progenitors donor 2	-
CD34 stem cells adult bone marrow derived	-
CD4+ CD25+ CD45RA+ naive regulatory T cells donor 1	++
CD4+ CD25+ CD45RA+ naive regulatory T cells donor 2	++
CD4+ CD25- CD45RA- memory conventional T cells donor 1	++
CD4+ CD25- CD45RA- memory conventional T cells donor 2	+++
CD8+ T Cells pluriselect donor 090309 donation 1	+++
CD8+ T Cells pluriselect donor 090309 donation 2	++
CD8+ T Cells pluriselect donor 090309 donation 3	++
CD8+ T Cells pluriselect donor 090325 donation 1	+++
CD8+ T Cells pluriselect donor 090325 donation 2	++
CD8+ T Cells pluriselect donor 090612 donation 1	++
CD8+ T Cells pluriselect donor 09062 donation 2	+++
CD8+ T Cells pluriselect donor 090612 donation 3	++
Common myeloid progenitor	-
Eosinophils	-
Basophils	+

Table 5.1: *DNMT1* CAGEseq data. *DNMT1* CAGEseq data for immune system cell based on FANTOM5 [133, 155]

membrane so that they are accessible for antibody staining of the target protein in the step after next. But first, unspecific binding of antibodies to the membrane has to be prevented by blocking the membrane with generic proteins. During the subsequent incubation step, specific primary antibodies are applied to the membrane targeting the protein of interest. A secondary antibody is then applied to bind to the primary antibody to detect the binding events. This creates the typical bands in of target protein in the western blot images, which are visualised in the last step. Using a ladder of proteins with a known size in the same gel, the approximate size of the target proteins in the bands can be estimated.

The critical protein-specific step in western blot is the choice of primary antibody that targets the protein of interest. In this specific case, there are many anti-DNMT1 antibodies commercially available but not all would also target the short DNMT1 isoform that we were trying to detect. It was therefore critical to identify which part of the DNMT1 protein the antibody's immunogen, and whether this sequence was

still present in the putative short DNMT1 isoform. Antibodies that we considered and their respective antigens in the short DNMT1 isoform are listed in table 5.2.

Name	Company	Immunogen	Species Reactivity
D4567	Sigma	1581-1595 of rat DNMT1	human, rat, mouse
JF09-89	Novus	human Dnmt1 aa 1510-1550	
JF09-89	ThermoFisher	Human Dnmt1 aa +L577:L592 1506-1549	Rat, Human, Mouse
EPR3522	Abcam	Synthetic peptide within Human Dnmt1 aa 1600-1700 (C terminal)	Human
2C1	Abcam	ecombinant fragment within Human Dnmt aa 1 to the C-terminus	Human
ABE1956	Sigma	11 amino acids from the C-terminal half with acetylated K1111, 1113, 1115, and 1117	human
C-term S1602	Abcepta	1588-1616 aa	Human

Table 5.2: Western Blot antibodies. Antibodies considered for the western blot. The clone JF09-89 from Novus gave the best results.

We compared HEK and JURKAT cell lines, as well as primary CD14+ monocytes, primary CD8+ T-cells, and peripheral blood mononuclear cells (PBMCs) depleted of primary monocytes and CD8+ T-cells from one blood donor. The Western Blot shows both the constitutive, full-length 190 kDa DNMT1 protein, as well as a band indicative of a shorter 60kDa protein, with a length matching that of the new isoform implied by EpiNN. This isoform is only visible in the tested immune system cells and not in the HEK cells, as predicted by the other presented data. While the Western Blot shows multiple additional bands around 140kDa, 120kDa, 110 kDa, 80 kDa, and 50 kDa, such bands could arise e.g. from noise incurred by post-translational modifications of DNMT1 that regulate protein stability [157]. Additional results of biological repeats are given in fig. B.1.

5.4.3 PCR

We also performed PCR to search for the putative shorter transcription isoform in immune system cells. Preliminary PCR data shows a signal at the predicted length of the shorter *DNMT1* starting at the putative TSS (see fig. 5.4b). While it is indicating that the transcript might exist, further tests including negative controls in non-immune system cell types as well as larger numbers of donors must be executed and are indeed currently being pursued. Additional results of biological repeats are given in fig. B.2.

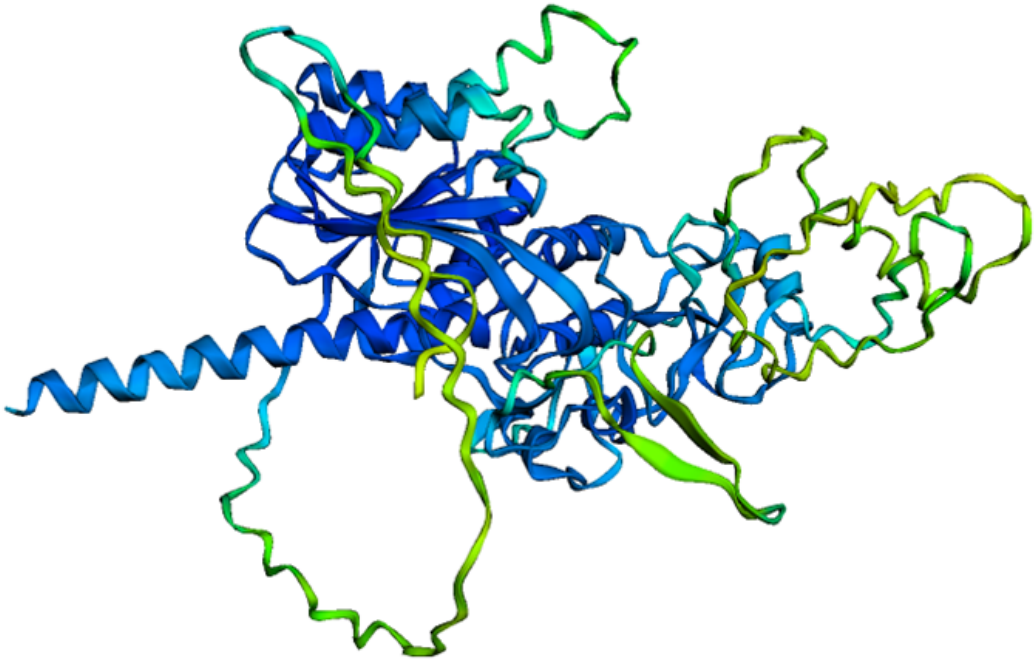


Figure 5.5: Structure for truncated DNMT1 predicted by Alphafold2. The active site around amino acid at position 1,226, the S-adenosyl-L-methionine binding site around amino acid at position 1,578, and the S-adenosyl-L-methionine binding site around amino acid position 1,580 are structured similarly to the long DNMT1 isoform, implying potential functionality. The S-adenosyl-L-methionine binding site around amino acid position 1,146 and the S-adenosyl-L-methionine binding site around amino acid position 1,191 seem unstructured and are therefore most likely not functional. Performing structural alignment of this structure and the full DNMT1 structure of the long isoform (PDB ID 4WXX, chain A) results in RMSD = 3.02, TM-score = 0.36, sequence identity = 86%, equivalent residues = 445, reference coverage = 38%, and target coverage = 82%, indicating that that the putative short DNMT1 is folded similarly to the known long isoform.

5.4.4 DNMT1 structure analysis

Comparing the truncated DNMT1 structure (see fig. 5.5 to the long isoform DNMT1 structure in the protein database (PDB ID 4WXX, chain A), using the results in high structural alignment: RMSD 3.02, TM-score 0.36, sequence identity 86%, equivalent residues 445, reference coverage 38%, and target coverage 82%.

Detailed structure comparison to full *DNMT1* protein, shows that the following sites have the same structure as in the normal protein:

- around amino acid (aa) 1226, Active Site: structured similarly to the long DNMT1 isoform
- aa 1578, Binding site (S-adenosyl-L-methionine; via carbonyl oxygen: structured similarly to the long DNMT1 isoform
- aa 1580, Binding Site (S-adenosyl-L-methionine; via amide nitrogen): structured similarly to the long DNMT1 isoform

The following sites seem more unstructured in the newly predicted protein than in the long DNMT1 isoform:

- aa 1146, Binding site (S-adenosyl-L-methionine; via carboxyl oxygen): unstructured
- aa 1191, Binding site (S-adenosyl-L-methionone): unstructured

The following sites are completely missing from the newly predicted protein:

- aa 353, Zinc Metal binding
- aa 356 Zinc metal binding
- aa 414, Zinc metal binding
- aa 418 Zinc metal binding

This means that, while some binding sites are not functional in the short DNMT1 isoform (due to being unstructured or completely missing), critically, the active site and surrounding binding sites are normally folded and therefore potentially functional.

Additionally, a comparative genomics analysis shows that the domain structure of the predicted short DNMT1 isoform would be similar to homologous proteins seen in bacteria [158], see fig. 5.5. Comparing this structure to the bacterial DNMT1 structure in the protein database (PDB ID 5LOD, chain A), using the results in high structural alignment: RMSD 3.03, TM-score 0.5, sequence identity 22%, equivalent residues 296, reference coverage 92%, and target coverage 55%.

5.4.5 *DNMT1* conclusions

These experiments or data sources in itself only give an indication each that the putative TSS actually produces a protein. But integrating all information, they all point in the same direction: the possible transcription and translation of the novel *DNMT1* isoform. It is therefore likely that a protein is produced, but potentially only at low levels and/or only in very specific cell types or cell states. But even a single active short DNMT1 protein could have a significant influence on a cell: Unchecked methylation can lead to substantial changes in a transcription program, altering the functionality of the cell. It is so far unclear whether the putative protein is active or not and why this protein might be useful for immune function.

5.5 Additional EpiNN detections

The EpiNN findings are not always clearly immune system-relevant, and the model might not have picked up other regions despite clear immune system relation. There are at least three difficult cases:

First, genes, where the immune-related regulatory signals look clear but are shared with other "immune-like" cell types. Regrouping all CD34 cells, spleen, and thymus together with immune cells has helped with that. In some cases, rectal mucosa and gastric samples would also need to be grouped with the "immune block", as these tissues have large populations of infiltrating lymphocytes, likely bringing about the immune-like characteristics. Example genes: *LY75-CD302*, *RIPOR2*, *ANKRD44*, *RASGRP1*, *LEPROTL1*, *CCDC88B*.

Second, very short genes flanked by long housekeeping genes (or even several) are challenging since the immune-relevant signal is dwarfed by the ubiquitous one. A potential solution would be to narrow the window or train two models: one with a narrow window (e.g. 20,000 bp) and another with a wide window (e.g. 100,000 bp). Example genes: *GZMM*, *CD79A*, *IL4I1*, *TTC24*.

Third, genes that are not immune-specific but differentially regulated in the immune cells. Some of these cases are clearly enhancers rather than proper genes

(*PDZD4*, *SYTL2*, *PRKX*, *CASK*, *FLT3LG*, *KLF12*, *MYBL1*, *LINS1*, *RPS6KA5*, *ARHGEF3*, *TMIGD1*, maybe *GCSAM*), while others may have immune-specific splicing variants (*TMEM229B*, *CARMIL2*, maybe *ARHGEF3*). Such regulatory aberrations would be interesting per se but might significantly undermine the model's performance.

Having optimised different settings, it is clear that no perfect model can be created, and some cases remain difficult but can be vastly reduced in numbers.

6 Discussion

Contents

6.1 Summary **84**
6.2 Implications for future work **85**
6.3 Limitations and future work **87**

6.1 Summary

In this thesis, I described a supervised classification algorithm called EpiNN that leverages a low-dimensional encoding of genome-wide epigenetic data produced by the ChromHMM model [1] to detect AIIS-relevant regions. I trained EpiNN using a handcrafted list of 888 IS loci and applied the model to scan the human genome for new IS regions. I built a genome-wide annotation of predicted IS relevance, detecting 2,765 loci with a high (> 0.5) EpiNN score. A large fraction of these loci did not have an associated Gene Ontology term, while the remaining were enriched for immune-related function. To evaluate EpiNN's efficacy in detecting genomic regions enriched for heritability in IS traits, I built a genome-wide EpiNN annotation and performed stratified LD score regression analysis using association summary statistics for 176 traits (average $N = 262k$). I observed a significant enrichment and heritability effect size τ^* for 20/26 IS traits, which remained significant after

conditioning on 97 other functional and evolutionary annotations. To confirm the specificity of the detected enrichments for IS phenotypes, I performed a meta-analysis of 63 independent traits, which revealed that IS-related traits were 4.45× (SE 0.09) more enriched for heritability than non-IS traits. Finally, we used S-LDXR [130] to test for the depletion of squared trans-ancestry genetic correlation in European and East Asian association summary statistics for 15 IS traits (average N_{EA} = 83k, N_{EU} = 235k). The EpiNN annotation was among the most significantly depleted annotations, indicating ancestry-specific effects that may be driven by gene-environment interactions at IS loci impacted by recent adaptation.

In chapter 5, I discussed several genes predicted by EpiNN to be immune system-relevant. These included validation genes, novel loci, novel enhancer connections, and a novel transcription start site. We performed PCR and Western Blot analyses for a constitutively expressed gene, *DNMT1*, which was flagged as AIIS-relevant by EpiNN, obtaining evidence for the presence of a novel immune system-specific transcription start site. Despite being noisy, the analysis conducted on *DNMT1* provides the necessary experimental validation for EpiNN and paves the way for future experimental work, such as investigating the novel findings for *TAGAP* and *IL2RB* as uncovered by EpiNN.

6.2 Implications for future work

Machine learning algorithms trained on regulatory annotations across tissues and cell types have recently shown great promise in predicting chromatin data and SNPs associated with a trait (e.g. [111, 112, 159, 160]). However, when conditioned on currently available coding, evolutionary, and regulatory data, these annotations were observed to only contain limited novel information for complex traits and diseases [129]. My analyses suggest that supervised classification algorithms trained on handcrafted sets of genes with known functional roles may be used to obtain insights into the architecture of a specific class of traits or diseases. Thanks to substantial recent development, CNNs are now easily adopted and applied, but I did briefly experiment with other, simpler models, such as logistic regression applied

to each window, and observed that CNNs performed sliightly better. This gap in performance, however, is not very large, and I believe that several classification models, including models that are simpler than CNNs, are likely to lead to good performance. As data sets for this analysis become larger, it is likely that more complex models will lead to improved performance by allowing to learn more complex features of the data. In our analysis of 176 complex traits and diseases, I found heritability effects for the EpiNN annotation to be specific to known immune traits, or respiratory and dermatological traits for which an immune component is likely to exist. This suggests that this approach may allow the identification of traits that would not be otherwise known to have an immune-related component. In addition, our observation of decreased cross-ancestry genetic correlation in regions predicted to be AIIS-relevant provides further evidence for selection acting on immune-related regions and may inform approaches that aim to maximise portability of polygenic scores across ancestries [161, 162]. At the same time, the negative results of the brain and epithelial genes indicate that the success of EpiNN relies heavily on the large data availability found for immune system cells, as the prediction accuracy could not be replicated with epithelial or brain genes.

Evidence for a novel TSS in *DNMT1* and heritability effects observed for EpiNN in conditional S-LDSC analyses also highlight the potential of methods that utilise epigenetic information to complement transcription-based approaches. This complementarity might be driven by the presence of a larger fraction of non-coding or microRNA genes predicted by EpiNN to be AIIS-relevant. This is likely due to many microRNAs losing their poly-A tail very early in their lifespan and, therefore, not being detected using transcription-based approaches, which rely on the poly-A tail for detection. The prediction of many long non-coding RNAs (lncRNAs) by EpiNN is consistent with their newly recognised roles in inflammatory diseases [81]. lncRNAs play an important role in the dynamic changes in the expression of genes in immune cells, which are necessary to react to infections. The functional consequences of disease-causing SNPs in lncRNAs and their biological pathways

are under-discovered. Still, approaches such as EpiNN provide an avenue to future analyses of lncRNAs with potential immune system relevance.

6.3 Limitations and future work

There are several limitations to the EpiNN approach and directions for future work. First, EpiNN uses a supervised training procedure that relies on a small set of hand-crafted genomic regions related to a specific class of well-studied genes. This inevitably leads to predicting genes exhibiting similar epigenetic profiles, potentially excluding AIIS-relevant genes with different activation patterns from the resulting EpiNN predictions. Reassuringly, however, I observed that EpiNN also flagged various genes active in different subcultures of immune system cells and are exclusively immune system relevant or only immune system-specific splicing variants. Interestingly, EpiNN obtained larger conditional heritability effects compared to the training list. Second, while the experimental results of DNMT1 suggest the existence of a novel isoform in immune cells as we see a new isoform of the predicted size in the Western Blot, some of the other bands are still unexplained, suggesting the need for further validation. Third, although the analyses primarily leveraged ChromHMM data, the EpiNN model may be extended to consider additional sources of information, such as Hi-C [163] or chromatin accessibility data [164]. Despite these limitations and avenues of future work, I believe that the EpiNN model and genome-wide annotation provide a valuable tool for analysing genomic regions involved in the regulatory mechanisms of the immune system.

Appendices

A Appendix

Contents

A.1 Applying S-LDSC to evolutionary annotations: CoalNN 89

A.1 Applying S-LDSC to evolutionary annotations: CoalNN

In this appendix, I briefly describe an additional S-LDSC analysis I performed in a collaborative project with co-authors Juba Nait Saada (J.N.S.) and Zoi Tsangalidou (Z.T.) and Pier Palamara (P.P.) [165]. The paper develops CoalNN, an algorithm to infer pairwise coalescence times and allele ages using convolutional neural networks. My contribution to this project was to perform S-LDSC analyses, testing an evolutionary annotation derived from CoalNN for heritability enrichment across complex traits.

Testing involved 26 different annotations for each of the 1000 Genome Project populations [118], one for each population super-group and one encompassing all 26 populations. Because allele frequency and allele age are correlated, annotations were quantile-normalised within 10 minor allele frequency bins. Overall, I tested 32 allele age annotations built by my co-authors J.N.S. and Z.T., evaluating the effect of these

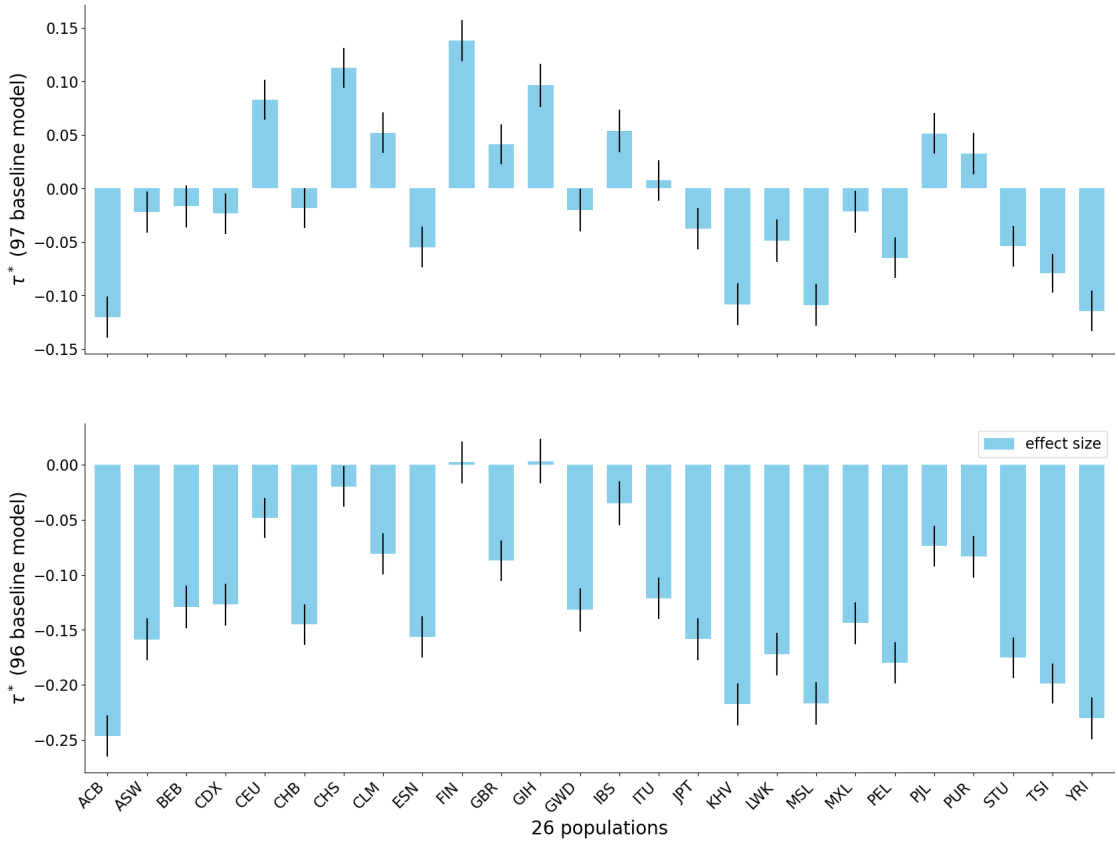


Figure A.1: S-LDSC effect sizes for CoalNN. Meta-analysed effect sizes of 63 independent complex traits for the CoalNN annotations of the 26 populations in the 1000 Genome Project. First graph with conditioning on the normal 97 baseline model (including ARGweaver), second graph with conditioning on the 96 baseline model (excluding ARGweaver). Between the two analyses, one can see the collinearity effects between CoalNN and ARGweaver vanishing: The unexpected sign shift for some of the populations disappears when the highly correlated ARGweaver is taken out of the multiple regression model.

CoalNN annotations on complex trait heritability by applying S-LDSC to summary association statistics from the same 63 independent diseases and complex traits I used for EpiNN, described in table B.2. Unlike the analysis of the EpiNN results, I did not discern between the 63 traits and meta-analysed all of them together.

Also, unlike in the EpiNN analysis, there were already annotations in the baseline model that captured similar evolutionary signal: ARGweaver-predicted allele age [166], nucleotide diversity, background selection statistic (B-statistic) [167], average pairwise TMRCA ASMC_{avg} [168], recombination rate, level of L.D., and CpG content [169]. Therefore, particular care had to be taken to ensure that the

CoalNN annotation captured signal orthogonal to these existing annotations. To this end, we evaluated only the S-LDSC effect size and not heritability enrichment and, as a first step, checked the correlation between the annotations.

To understand the differences between the baseline and CoalNN annotations, we calculated the correlation between annotations. As expected, the correlation between CoalNN annotation and evolutionary annotations was high since they capture similar information. The highest correlation was with ARGweaver MAF-adjusted allele age estimates (average across 26 CoalNN populations $r = 0.56$, S.E. $= 0.01$), followed by MAF-adjusted LLD-AFR ($r = 0.3$, S.E. $= 0.01$), MAF-adjusted ASMCavg ($r = 0.28$, S.E. $= 0.0$), and nucleotide diversity ($r = 0.16$, S.E. $= 0.0$). This presents additional challenges for this analysis because S-LDSC relies on linear regression, and the presence of collinear predictors may lead to unreliable effect size estimates [170]. E.g. calculating the effect size in a multi-linear regression with two highly correlated variables with a correlation approaching 1, then the coefficient estimates can misleadingly tend toward infinite magnitudes in opposite directions, even if the variables' actual effects are small and of the same sign [171]. This indeed happened for some CoalNN annotations when I calculated the CoalNN effect size on the full 97 annotation baseline model to investigate the ability of the CoalNN annotations to capture heritability signal (see fig. A.1): Most of the 26 populations had a significant effect size but some in the opposite direction, e.g. (e.g. $\tau^*_{\text{A.C.B.}} = -0,12$, SE=0.02, and $\tau^*_{\text{C.H.S.}} = 0.11$, SE=0.02). To circumvent this problem, I removed ARGweaver (the most correlated) from the baseline model and calculated the CoalNN effect size again on a 96 baseline model, this time resulting in consistent effect sizes across populations (see fig. A.1).

To better compare the effect of CoalNN and ARGweaver on complex trait heritability, J.N.S. and Z.T. aggregated age estimates across all 26 populations to build one overall CoalNN annotation. To avoid issues with collinearity, in order to compare the ARGweaver and CoalNN annotation, I performed the same tests once with ARGweaver in the model and once with CoalNN. This allows me to compare the two annotations.

First, when removing all evolutionary annotations from the baseline model, only using the 90 non-evolutionary annotations as the baseline model, CoalNN (τ^* = -1.01; SE=0.02) obtained a larger effect size estimate compared to ARGweaver (τ^* = -0.88; SE=0.02) in a meta-analysis.

Second, when using the 96 baseline model (all evolutionary annotations but ARGweaver), CoalNN ($\tau^* = -0.22$; SE=0.02) again obtained a larger effect size estimate compared to ARGweaver ($\tau^* = -0.15$; SE=0.01), resulting in significant effect size for both allele age estimates, indicating that both allele age annotations provide information that is orthogonal to that provided by other evolutionary annotations. In addition, I observed larger effect sizes for evolutionary annotations when including ages inferred by CoalNN than ARGweaver, which suggests that ARGWeaver has more overlapping information with the other baseline evolutionary annotations.

Overall, these analyses indicate that the CoalNN annotation contains new orthogonal signal not included in ARGWeaver or other evolutionary annotations in the baseline model. The negative sign of the effect size indicates that recent mutations are enriched for heritability, which is in line with the hypothesis that negative selection has had less time to eliminate recent pathogenic mutations, resulting in younger variants being more likely to have a significant deleterious effect than older variants.

B
Appendix

In the appendix:

- Amino acid sequence of proposed short DNMT1 protein isoform
- Technical and biological replicates of DNMT1 Western Blot (fig. B.1)
- Technical and biological replicates of *DNMT1* PCR (fig. B.2)
- List of EpiNN training genes (table B.1)
- List of 63 complex traits used in S-LDSC meta analysis (table B.2)
- List of S-LDSC results for 63 complex traits (table B.3)
- List of S-LDSC restults for 176 complez traits (table B.4)
- S-LDXR results (table B.5)
- *DNMT1* CAGEseq raw data (table B.6)

Short DNMT1 protein amino acids

The proposed amino acid chain of the short DNMT1 protein is as follows:

MGGP NRFYFLEAYN AKSKSFEDPP NHARSPGNKG KGKGKGKGKP KSQACEPSEP EIEIKLPKLR TLDVFSGCGG LSEGFHQAGI SDTLWAIEMW DPAAQAFRLN NPGSTVFTED CNILLKLVMA GETTNSRGQR LPQKGDVEML CGGPPCQGFS GMNRFNSRTY SKFKNSLVVS FLSYCDYYRP RFFLLENVRN FVSFKRSMVL KLTLRCLVRM GYQCTFGVLQ AGQYGVAQTR RRAIILAAAP GEKLPLFPEP LHVFAPRACQ LSVVVDDKKF VSNITRLSSG PFRTITVRDT MSDLPEVRNG ASALEISYNG EPQSWFQRQL RGAQYQPILR DHICKDMSAL VAARMRHIPL APGSDWRDLP NIEVRLSDGT MARKLRYTHH DRKNGRSSSG ALRGVCSCVE AGKACDPAAR QFNTLIPWCL PHTGNRHNHW AGLYGRLEWD GFFSTTVTNP EPMGKQGRVL HPEQHRVVSV RECARSQGFP DTYRLFGNIL DKHRQVGNAV PPPLAKAIGL EIKLCMLAKA RESASAKIKE EEAAKD

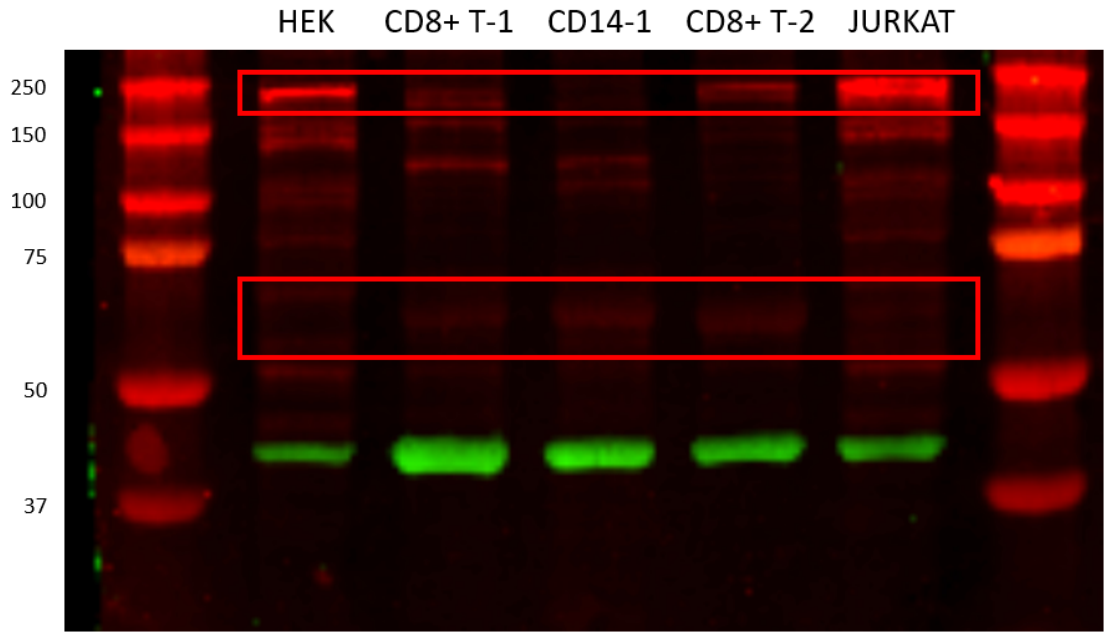


Figure B.1: DNMT1 Western Blot replicates. Biological replicates of two additional donors T-1 and T-2. In total, there are 3 biological and 2 technical replicates showing similar levels of signal strength.

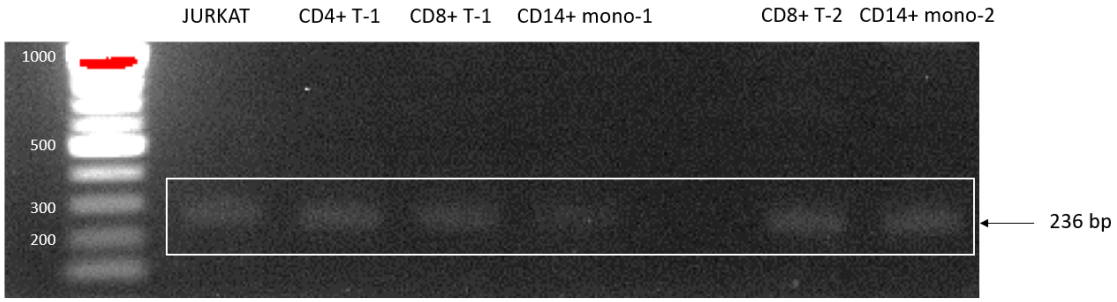


Figure B.2: *DNMT1* PCR replicate. Biological replicates of two additional donors T-1 and T-2. In total, there are 3 biological and 2 technical replicates showing similar levels of signal strength.

Table B.1: List of EpiNN training genes.

Chromosome	Start bp	End bp	Gene identifier	Strand
chr10	101770129	101780369	NM_001206389.1	-
chr10	112446987	112818744	NM_001318203.1,NM_001318205.1,NM_145206.3	+
chr10	117241072	117279430	NM_003054.4	+
chr10	12349542	12826464	NM_020397.3	+
chr10	12388981	12826464	NM_001351032.1	+
chr10	133230273	133231558	NM_003577.2	+
chr10	14878782	14904315	NM_001193424.1,NM_001193426.1	+
chr10	14878865	14904315	NM_001193425.1,NM_001193427.1	+
chr10	14878899	14904315	NM_024670.3	+
chr10	16590615	16817454	NM_012425.3,NM_152724.2	-
chr10	17589958	17617374	NM_014241.3	-
chr10	28050992	28282097	NM_173496.4	-
chr10	28050992	28303066	NM_001318170.1	-
chr10	47300223	47313583	NM_004962.4	+
chr10	48909482	48914235	NM_001006939.3	-
chr10	6010693	6062370	NM_000417.2,NM_001308242.1,NM_001308243.1	-
chr10	60778330	60794852	NM_001320918.1	+
chr10	60778453	60794852	NM_001786.4,NM_033379.4	+
chr10	62193085	62268863	NM_145307.3	-
chr10	62811995	62816380	NM_000399.4,NM_001136177.2,NM_001136179.2,NM_001321037.1	-
chr10	69220302	69267559	NM_025130.3	+
chr10	70597347	70602788	NM_001083116.2,NM_005041.5	-
chr10	7818503	8015053	NM_031923.3	+
chr10	8054703	8075201	NM_001002295.1,NM_002051.2	+
chr10	84328588	84518521	NM_001284240.1,NM_001351290.1,NM_001351292.1,NM_018999.3	+
chr10	88819898	88851975	NM_144590.2	-
chr10	93893969	93903824	NM_001134658.2,NM_001345993.1,NM_153226.3	+
chr1	100345014	100499465	NM_001319211.1	+
chr1	100352466	100485370	NM_033313.2	+
chr1	100352466	100499465	NM_001319212.1,NM_033312.2	+
chr1	100352466	100520277	NM_001319210.1,NM_003672.3	+
chr1	107571159	107688504	NM_001079874.1	-
chr1	107571159	107964923	NM_006113.4	-
chr1	108560348	108639327	NM_001010883.2	+
chr1	108648287	108661122	NM_001102592.1	-
chr1	108648287	108661526	NM_144584.2	-
chr11	1012823	1036706	NM_005961.2	-
chr1	110871198	110899936	NM_001040033.1	+
chr1	110873099	110899936	NM_000560.3,NM_001320638.1	+
chr1	111227658	111243440	NM_001025197.1,NM_004000.2	+
chr1	111229710	111243440	NM_001025199.1	+
chr11	117986390	118001483	NM_001558.3	+
chr11	118304579	118316175	NM_000733.3	+
chr11	118339073	118342744	NM_000732.4,NM_001040651.1	-
chr11	118344343	118353782	NM_000073.2	+
chr11	122838497	122872639	NM_019604.3	+
chr11	122862310	122872639	NM_001304782.1	+
chr11	123430462	123435719	NM_001354837.1	+
chr11	128838013	128842534	NM_000220.4	-
chr11	128838013	128867373	NM_153764.2,NM_153765.2,NM_153766.2,NM_153767.3	-
chr1	113813810	113871759	NM_001193431.2,NM_001308297.1,NM_012411.5,NM_015967.6	-
chr1	113876813	113887547	NM_001010922.2	-
chr11	14643722	14872059	NM_000922.3	+
chr1	115048010	115089500	NM_001308315.1,NM_001308316.1,NM_005725.5	-
chr11	1584341	1585283	NM_001005922.1	-
chr1	116514533	116571093	NM_001779.2	-
chr1	116518699	116571093	NM_001144822.1	-
chr1	116754429	116769229	NM_001328609.1,NM_001767.4	+
chr1	117001749	117036551	NM_001256106.2,NM_001256109.2,NM_001256111.2,NM_004258.5	+
chr11	1732409	1750594	NM_001170820.3	-
chr11	17788047	18013162	NM_012139.3	-
chr11	19117144	19176420	NM_001001483.2,NM_019028.2	+
chr1	1203507	1206709	NM_004195.2,NM_148901.1,NM_148902.1	-
chr1	12063376	12144207	NM_001243.4,NM_001281430.2	+
chr11	2295639	2301913	NM_001142946.2,NM_001329958.1	-
chr11	2444990	2849110	NM_000218.2	+
chr11	2461453	2849110	NM_181798.1	+
chr11	298199	299526	NM_001025295.2	-
chr11	31432398	31509628	NM_001304274.1,NM_144981.2	-
chr11	417929	442011	NM_001012302.2,NM_001347882.1	-
chr1	151800263	151805442	NM_001004432.3	-
chr11	5225465	5227071	NM_000518.4	-
chr1	153330119	153348841	NM_020393.3	-
chr1	156581364	156587728	NM_001105669.3	+
chr1	157676480	157700872	NM_001320333.1	-
chr1	157678187	157700857	NM_052939.3	-
chr1	158289772	158294774	NM_001765.2	+
chr1	158327949	158331531	NM_001764.2	-
chr1	158931546	158977059	NM_152501.4,NM_198928.4,NM_198929.4,NM_198930.3	+
chr1	159062483	159076901	NM_001348247.1,NM_004833.2	-
chr1	159800510	159816257	NM_001284217.1	+
chr1	159802382	159816257	NM_001004310.2	+
chr1	159826688	159837492	NM_001330741.1,NM_020125.2	+
chr1	160485029	160523262	NM_001184714.1,NM_001184715.1,NM_001184716.1,NM_052931.4	-
chr1	160541093	160579516	NM_001184879.1,NM_001184881.1,NM_001184882.1, NM_001330742.1,NM_003874.3	-
chr1	160608099	160647311	NM_001330754.1,NM_003037.4	-
chr1	160678745	160711851	NM_001778.3	-
chr1	160680421	160711851	NM_001256030.1	-
chr1	160739056	160754821	NM_001282588.1,NM_001282589.1,NM_001282590.1, NM_001282591.1,NM_001282592.1,NM_001282593.1, NM_001282594.1,NM_001282595.1,NM_001282596.1,NM_021181.4	+
chr1	160796073	160802746	NM_001033667.2	+
chr1	160796073	160828255	NM_001261456.1,NM_001261457.1,NM_002348.3	+
chr1	160830159	160862902	NM_001166663.1,NM_001166664.1,NM_016382.3	-
chr11	60971640	61020377	NM_001254750.1,NM_001254751.1,NM_006725.4	+
chr1	161046941	161069970	NM_001025598.1,NM_001287600.1,NM_001287602.1,NM_181720.2	-
chr11	61093962	61127852	NM_001346456.1	+
chr11	61102457	61127852	NM_014207.3	+
chr1	161541758	161550103	NM_000569.7,NM_001127592.2,NM_001329122.1	-
chr1	161541758	161550623	NM_001127593.1,NM_001127595.1,NM_001127596.1	-
chr1	161541758	161550737	NM_001329120.1	-
chr1	161623195	161631047	NM_001271036.1,NM_001271037.1	-
chr1	161623195	161631462	NM_001244753.1,NM_001271035.1	-
chr1	161623195	161631963	NM_000570.4	-
chr11	63552769	63563383	NM_017878.1	-
chr11	64340217	64357534	NM_032251.5	+
chr11	64824188	64844569	NM_017525.2	-
chr1	167052835	167090631	NM_005814.2	-
chr1	167430639	167518610	NM_000734.3,NM_198053.2	-
chr1	168540764	168543997	NM_003175.3	-

chr11	68736131	68751520	NM_001039656.1	-
chr1	169588848	169630139	NM_003005.3	-
chr1	169690664	169711702	NM_000655.4	-
chr1	172659007	172666873	NM_000639.2,NM_001302746.1	+
chr11	73950318	73970287	NM_153614.3	+
chr11	78435960	78574864	NM_001243251.1,NM_024678.5	-
chr11	82901694	82934659	NM_145018.3	+
chr11	85694220	85719180	NM_001289610.1,NM_206929.3	-
chr11	85694220	85719350	NM_001162952.2,NM_001289609.1,NM_206930.3	-
chr11	85694220	85758069	NM_001162951.2,NM_001162953.2	-
chr11	85694220	85811159	NM_001289608.1,NM_032943.4	-
chr1	192575726	192580029	NM_002922.3	+
chr1	197734008	197775493	NM_001195216.1	-
chr1	198523221	198540945	NM_001320218.1,NM_133262.2,NM_133326.1	-
chr1	198638967	198695171	NM_001267798.1	+
chr1	198638967	198757476	NM_002838.4,NM_080921.3	+
chr1	202147012	202159991	NM_080588.2	-
chr1	202147012	202161588	NM_001199797.1,NM_002832.3	-
chr1	203765155	203776372	NM_001136190.1,NM_017773.3	+
chr1	203769637	203776372	NM_001282878.1	+
chr1	206767602	206772494	NM_000572.2	-
chr1	206798869	206842981	NM_153758.2	+
chr1	206828876	206842981	NM_013371.3	+
chr1	207454299	207489895	NM_001006658.2,NM_001877.4	+
chr1	209756031	209782323	NM_001320144.1,NM_025228.3	+
chr1	209758497	209782323	NM_001320143.1	+
chr1	209768487	209779657	NM_001287754.1	+
chr12	1027747	1495933	NM_001301248.1	+
chr12	10304445	10317251	NM_001114396.2	+
chr12	10372352	10390054	NM_007360.3	-
chr12	10372352	10410146	NM_001199805.1	-
chr12	10407383	10409757	NM_013431.2	-
chr12	104456913	104762014	NM_001173982.1,NM_018413.5	+
chr12	107318419	107659642	NM_001018072.1,NM_001347943.1	+
chr12	107320587	107659642	NM_001347944.1	+
chr12	107580856	107659642	NM_001017523.1	+
chr12	10880960	11171623	NM_001291314.1,NM_001291315.1	-
chr12	10937404	11171623	NM_001316893.1	-
chr12	10995961	10998304	NM_176889.3	-
chr12	11030386	11031407	NM_176885.2	-
chr12	111931281	112013219	NM_001193453.1,NM_001193531.1,NM_001294314.1,NM_138341.2	-
chr1	211326614	211374946	NM_004619.3	+
chr1	211326805	211374946	NM_001033910.2,NM_001319207.1	+
chr1	211346363	211374946	NM_145759.2	+
chr12	119989843	120094495	NM_207311.2	+
chr12	121651386	121669654	NM_173855.4	-
chr12	14914026	14929125	NM_001300784.1	-
chr12	14914026	14938549	NM_152321.3	-
chr12	15546343	15598331	NM_030668.2,NM_030669.2,NM_030670.2,NM_030671.2	+
chr1	223220818	223364202	NM_017982.3	-
chr1	223234909	223364202	NM_001037175.2	-
chr1	226231148	226309503	NM_001270409.1,NM_173083.3	-
chr1	226231148	226309748	NM_001270410.1	-
chr12	28190426	28550166	NM_001330367.1,NM_001352078.1,NM_001352079.1, NM_001352080.1,NM_001352081.1,NM_001352082.1,NM_001352083.1, NM_001352084.1,NM_001352086.1,NM_001352087.1,NM_018318.4	+
chr1	231626814	231724074	NM_001012958.1,NM_001164550.1,NM_001164552.1,NM_001164553.1	+
chr1	231626814	231750619	NM_001164551.1,NM_001164555.1	+
chr1	231626814	231771662	NM_001164549.1	+
chr1	231626814	231819244	NM_001164544.1,NM_001164545.1,NM_001164548.1,NM_001164556.1	+
chr1	231626814	231855380	NM_001164539.1	+
chr1	231626814	231866708	NM_001012959.1	+
chr1	231626814	231959503	NM_001164542.1,NM_001164546.1	+
chr1	231626814	232009654	NM_001164538.1,NM_001164541.1,NM_001164547.1	+
chr1	231626814	232041272	NM_001012957.1,NM_001164537.1,NM_001164540.1,NM_018662.2	+
chr1	232984135	233295713	NM_014801.3	-
chr12	3615327	3753200	NM_001144958.1	-
chr1	23627333	23640568	NM_001348075.1	+
chr12	3647942	3753200	NM_032680.3	-
chr12	39551219	39620041	NM_005164.3	-
chr12	44508274	44875661	NM_001145109.1	-
chr12	44508274	44876464	NM_001145108.1,NM_006159.2	-
chr12	44508274	44876850	NM_001145107.1	-
chr12	44508274	44913928	NM_001145110.1	-
chr1	24502350	24537020	NM_001251979.1,NM_001251984.1	+
chr1	24502896	24537020	NM_001251977.1,NM_001251981.1,NM_001251983.1,NM_013441.3	+
chr1	24507593	24537020	NM_001251978.1	+
chr1	24514313	24537020	NM_001251980.1,NM_001251982.1,NM_001251985.1	+
chr12	47079602	47236663	NM_001281429.1,NM_138371.2	+
chr1	247416155	247449108	NM_001079821.2,NM_001127461.2	+
chr1	247418048	247449108	NM_001127462.2,NM_001243133.1,NM_004895.4,NM_183395.2	+
chr12	48904109	48921576	NM_001286957.1,NM_033124.4	+
chr1	24899510	24930279	NM_004350.2	-
chr1	24899510	24965010	NM_001031680.2,NM_001320672.1	-
chr12	51281037	51324198	NM_001290008.1,NM_001290009.1,NM_016293.3	-
chr12	51281037	51324680	NM_001290007.1	-
chr12	52585588	52601538	NM_001146225.1,NM_001146226.1,NM_080747.2	-
chr12	52607569	52618559	NM_175068.2	-
chr12	52644557	52652175	NM_000423.2	-
chr12	54561443	54579239	NM_001165975.2,NM_001315534.1,NM_001315535.1 NM_001261844.1,NM_001351148.1,NM_001351150.1,	+
chr12	54948302	54973732	NM_001351151.1,NM_001351152.1,NM_001351153.1, NM_001351154.1,NM_001351155.1,NM_014796.2	-
chr12	54948302	54981838	NM_001098815.2	-
chr12	54948302	54984746	NM_001136030.2,NM_001351149.1	-
chr12	55636891	55637833	NM_206899.1	+
chr12	57028516	57050117	NM_005379.3	-
chr12	57028516	57050765	NM_001256041.1	-
chr12	57460134	57472268	NM_001160045.1	+
chr12	57460150	57472268	NM_001167609.1,NM_005269.2	+
chr12	57762333	57767193	NM_000785.3	-
chr12	59596039	59789855	NM_001270622.1,NM_001270623.1	+
chr12	59689336	59789855	NM_004731.4	+
chr1	26361633	26372794	NM_001114759.2,NM_001307925.1,NM_173574.3	-
chr12	64610512	64697567	NM_178169.3	+
chr12	6618834	6631649	NM_001142961.1	-
chr12	6618834	6636131	NM_020400.5	-
chr12	6789471	6820810	NM_000616.4,NM_001195014.2,NM_001195015.2, NM_001195016.2,NM_001195017.2	+
chr12	68154769	68159741	NM_000619.2	-
chr12	68248244	68253501	NM_020525.4	-
chr12	69738685	69823204	NM_022456.4,NM_175624.3	+
chr12	69739389	69823204	NM_175623.3,NM_175625.3	+
chr1	27612288	27626240	NM_001042747.1	-
chr1	27612288	27626595	NM_001042729.1	-
chr1	27612288	27635216	NM_005248.2	-
chr12	8950043	9010747	NM_001329101.1,NM_001329102.1,NM_001329103.1	+
chr12	8989624	9010747	NM_001329099.1,NM_005810.3	+
chr12	94148722	94307675	NM_005761.2	+

chr12	94368387	94459988	NM_001346460.1,NM_001346461.1,NM_001346462.1	-
chr12	95516559	95548272	NM_001278393.1	-
chr12	95516559	95548844	NM_001042403.2,NM_001347936.1,NM_001347937.1	-
chr12	95516559	95551490	NM_032147.4	-
chr12	9595273	9607901	NM_002258.2	-
chr12	9669707	9699555	NM_001004419.4,NM_001197317.2,NM_001197318.2, NM_001197319.2,NM_013269.5	+
chr12	9752485	9760901	NM_001781.2	-
chr1	32251238	32286167	NM_001330468.1,NM_005356.4	+
chr1	32274110	32286167	NM_001042771.2	+
chr13	28300345	28495128	NM_002019.4	-
chr13	28368096	28495128	NM_001160030.1	-
chr13	28385550	28495128	NM_001159920.1	-
chr13	28399043	28495128	NM_001160031.1	-
chr13	30713477	30764428	NM_001204406.1	+
chr13	42562735	42608013	NM_033012.3	+
chr13	42574154	42608013	NM_003701.3	+
chr13	51813346	51845152	NM_001351003.1,NM_001351005.1,NM_001351006.1	-
chr13	73686011	74133929	NM_007249.4	-
chr14	105049388	105065550	NM_001278694.1,NM_001278695.1,NM_001278696.1,NM_013345.3	-
chr14	23117305	23119611	NM_001805.3	-
chr14	24606479	24609720	NM_001270780.1,NM_001270781.1,NM_033423.4	-
chr14	24630953	24634269	NM_001346011.1,NM_004131.5	-
chr14	31025105	31096450	NM_001254729.1	+
chr14	31025469	31096450	NM_001128126.2	+
chr14	31026642	31096450	NM_001254728.1	+
chr14	35122320	35274078	NM_001256678.1,NM_001256679.1,NM_014672.3	+
chr14	35122551	35274078	NM_001256680.1,NM_001256681.1	+
chr14	45135938	45200890	NM_001308133.1,NM_020937.3	+
chr14	51860303	51969800	NM_001243773.1,NM_001243774.1,NM_053064.4	+
chr14	52267712	52276724	NM_000953.2,NM_001281469.1	+
chr14	53036739	53153328	NM_001160147.1,NM_001160148.1,NM_030637.2	-
chr14	61321442	61550980	NM_006255.4	+
chr14	64465498	64474503	NM_004857.3	+
chr14	65410591	65744121	NM_004480.4,NM_178155.2,NM_178156.2	+
chr14	67470265	67488732	NM_001348541.1,NM_001348542.1,NM_001348543.1	-
chr14	67470265	67515304	NM_001348544.1,NM_182526.2	-
chr14	67470265	67515485	NM_001348548.1,NM_001348549.1	-
chr14	70641786	70675360	NM_015351.1	+
chr14	88005123	88014811	NM_003608.3	+
chr14	89156171	89619150	NM_001085471.1	-
chr14	90868121	91060134	NM_001322237.1,NM_001322238.1	-
chr14	90868121	91060649	NM_001322227.1,NM_001322228.1,NM_001322229.1, NM_001322230.1,NM_001322231.1,NM_001322232.1, NM_001322233.1,NM_001322234.1,NM_001322235.1, NM_001322236.1,NM_004755.3	-
chr14	90893973	91060649	NM_182398.2	-
chr14	91232531	91244508	NM_001177676.1	-
chr14	91232531	91253880	NM_001348437.1,NM_003485.3	-
chr14	91779750	91836526	NM_001128595.2,NM_001289134.1,NM_152332.5	-
chr14	91779750	91867536	NM_001128596.2	-
chr14	96797346	96881614	NM_003384.2	+
chr14	99169286	99271713	NM_001282237.1,NM_001282238.1,NM_022898.2,NM_138576.3	-
chr15	100569222	100602240	NM_001040616.2,NM_001352507.1,NM_001352508.1	-
chr15	25678711	25863202	NM_024490.3	-
chr1	52633343	52657065	NM_001042693.2	+
chr15	35520272	35546203	NM_001141972.1	-
chr15	38488096	38564806	NM_001306086.1	-
chr15	38488100	38564806	NM_001128602.1,NM_005739.3	-
chr15	38696597	38700038	NM_207444.2	+
chr15	40287896	40307973	NM_001284297.1,NM_001284298.1,NM_004573.2	-
chr15	40300666	40307973	NM_001284299.1	-
chr15	44665731	44676888	NM_001145112.1,NM_001330283.1	-
chr15	45114324	45118106	NM_207581.3	+
chr15	48139427	48178361	NM_001301210.1,NM_016132.4	-
chr15	76994679	77037475	NM_001321137.1	+
chr15	76995084	77037475	NM_001321135.1,NM_001321136.1,NM_003978.4	+
chr15	79960889	79971301	NM_001114735.1,NM_004049.3	-
chr15	88877284	88895626	NM_001307952.1,NM_178232.3	-
chr15	90717326	90815462	NM_000057.3,NM_001287246.1,NM_001287247.1,NM_001287248.1	+
chr16	165973	166768	NM_001003938.3	+
chr16	23835978	24220611	NM_002738.6,NM_212535.2	+
chr16	27402161	27452042	NM_181078.2,NM_181079.4	+
chr16	27427257	27452042	NM_021798.3	+
chr16	2964215	2968383	NM_001253725.1,NM_001253726.1,NM_024507.3,NM_172229.2	+
chr16	29742462	29746019	NM_175900.3	-
chr16	3088889	3099317	NM_001282415.1,NM_001282416.1,NM_032805.2	-
chr16	31215961	31217093	NM_152901.3	-
chr16	50548329	50640860	NM_033119.4	+
chr16	50666299	50681353	NM_001144972.1,NM_153337.2	-
chr16	50671624	50681353	NM_182854.3	-
chr16	50697138	50733077	NM_022162.2	+
chr16	50699349	50733077	NM_001293557.1	+
chr16	57542636	57577189	NM_001304376.2,NM_001318481.1,NM_153837.3	+
chr16	67644939	67657569	NM_001013838.2,NM_001317026.2	+
chr16	85098351	85112508	NM_198491.2	-
chr16	88638592	88640474	NM_013278.3	+
chr16	88677681	88686474	NM_178310.3	-
chr16	88940036	88951524	NM_001290330.2	+
chr17	1279662	1300987	NM_172367.2	+
chr17	27631147	27649560	NM_001330163.1,NM_002308.3,NM_009587.2	+
chr17	2776920	3037741	NM_001330058.1	+
chr17	2796437	3037741	NM_001100398.1,NM_015085.4	+
chr17	29673911	29930228	NM_001282130.1	-
chr17	30316347	30334059	NM_001319942.1,NM_206832.2	-
chr17	35192519	35194393	NM_152462.2	-
chr17	35435095	35448837	NM_144682.5	-
chr17	35474922	35487739	NM_001195790.1	-
chr17	35871490	35880373	NM_001278736.1,NM_002985.2	-
chr17	36064271	36071481	NM_002988.3	+
chr17	3673226	3696404	NM_001204519.1,NM_001204520.1,NM_002561.3,NM_175080.2	-
chr17	39063302	39151649	NM_020405.4	-
chr17	3923868	3964464	NM_005173.3,NM_174953.2,NM_174954.2,NM_174955.2, NM_174956.2,NM_174957.2,NM_174958.2	-
chr17	39757714	39778225	NM_001284516.1	-
chr17	39757714	39864188	NM_001257408.1,NM_001257409.1,NM_001257410.1, NM_001257411.1,NM_001257412.1,NM_001257413.1, NM_001257414.1,NM_001284514.1,NM_001284515.1, NM_012481.4,NM_183228.2,NM_183229.2,NM_183230.2, NM_183231.2,NM_183232.2	-
chr17	41301825	41302851	NM_001257309.1	-
chr17	45254392	45262112	NM_152343.2	-
chr17	47209061	47223679	NM_001002841.1	+
chr17	47209347	47223679	NM_002476.2	+
chr17	47733243	47746119	NM_013351.1	+
chr17	48133439	48430232	NM_001075099.1,NM_003726.3	-
chr17	49206233	49209381	NM_031498.2	-
chr17	49206233	49209400	NM_001198755.1,NM_001198756.1	-
chr17	49206233	49210574	NM_001198754.1	-

chr17	58155153	58156086	NM_012374.1	+
chr17	58169655	58170579	NM_001004707.3	+
chr17	63998350	64004285	NM_001164257.1,NM_001191029.1,NM_001191030.1,NM_001191031.1	+
chr17	67377280	67697263	NM_012417.3,NM_181671.2	+
chr17	68267025	68291264	NM_001174166.1,NM_004694.4	-
chr1	77088980	77219447	NM_005482.2	-
chr17	74274209	74314884	NM_001172810.2,NM_001353167.1,NM_023036.5	+
chr17	74466372	74484798	NM_001330456.1,NM_001330457.1	+
chr17	74466382	74484798	NM_001256841.1,NM_007261.3	+
chr17	76624755	76643838	NM_001289107.1,NM_018414.4	-
chr17	78130777	78142968	NM_152468.4	+
chr1	7847611	7853044	NM_006786.3	-
chr1	7847611	7853491	NM_021995.2	-
chr1	78620381	78646145	NM_006820.3	+
chr17	8797109	8799349	NM_152599.3	-
chr17	8878915	8912517	NM_014308.3	-
chr17	8878915	8965712	NM_001142633.2,NM_001251851.1,NM_001251852.1, NM_001251853.1,NM_001251855.1	-
chr1	7915870	7943165	NM_001561.5	-
chr18	24460628	24479957	NM_001143828.1,NM_001160166.1,NM_021624.3	+
chr18	26226882	26391685	NM_001293725.1,NM_005640.2	+
chr1	84077974	84205439	NM_207578.2	+
chr1	84077974	84238498	NM_002731.3	+
chr18	45726952	45752520	NM_001146037.1	+
chr18	49041473	49460733	NM_001353210.1,NM_001353211.1,NM_001353212.1, NM_001353213.1,NM_001353214.1,NM_001353215.1, NM_001353216.1,NM_017653.4	-
chr18	56651384	57029805	NM_015285.2,NM_052834.2	+
chr18	69862955	69947853	NM_001303618.1	-
chr18	69862955	69957176	NM_001303619.1,NM_006566.3	-
chr18	70630518	70650857	NM_001278515.1	-
chr18	74630962	74804627	NM_001146190.1	+
chr18	74630962	74921058	NM_001146189.1	+
chr18	74630962	75065672	NM_017757.2	+
chr1	89258949	89272861	NM_001134486.2,NM_052942.3	-
chr1	8945833	8974444	NM_001270500.1	+
chr1	8945833	8975092	NM_001215.3,NM_001270501.1,NM_001270502.1	+
chr1	89633084	89719535	NM_032270.4	+
chr19	10512741	10517992	NM_001166215.1	-
chr19	11574659	11577741	NM_001611.4	-
chr19	11574659	11578194	NM_001322023.1	-
chr19	14031739	14052919	NM_004843.3	+
chr1	91500846	91525764	NM_001134419.1,NM_003503.3	+
chr1	91501107	91525764	NM_001134420.1	+
chr19	17795109	17813576	NM_014256.3	+
chr19	18058994	18087003	NM_001290023.1,NM_005535.2	-
chr19	18058994	18098816	NM_001290024.1	-
chr19	18070561	18087003	NM_153701.2	-
chr19	19538264	19546659	NM_153221.2	+
chr1	92080304	92147840	NM_183242.3	+
chr19	23762943	23828136	NM_001355283.1,NM_001355287.1	+
chr1	92473044	92483954	NM_001127215.2	-
chr19	3136031	3163769	NM_002068.3	+
chr19	35449714	35451767	NM_005306.2	+
chr19	35704486	35717038	NM_001316902.1,NM_001316903.1,NM_014383.2	+
chr19	3777968	3801812	NM_002378.3	-
chr19	38587639	38618035	NM_001042600.2,NM_007181.5	-
chr19	39704305	39709448	NM_020129.2,NM_203471.1	+
chr19	41114431	41128381	NM_000774.4	+
chr19	41549517	41586844	NM_001288773.2	+
chr19	41576178	41586844	NM_001098506.3,NM_001290113.1,NM_033543.5	+
chr19	41618970	41627074	NM_001817.3	-
chr19	41877119	41881372	NM_001783.3,NM_021601.3	+
chr19	4292226	4302431	NM_001169126.1,NM_001308232.1,NM_144615.2	-
chr19	45162927	45178243	NM_001270891.1,NM_001270892.1,NM_001270893.1,NM_024108.2	-
chr19	45668243	45683724	NM_000164.3,NM_001308418.1	+
chr19	49289631	49325217	NM_014037.2	-
chr19	49388217	49417999	NM_144688.4	+
chr19	49474208	49486231	NM_001204502.1,NM_001278637.1,NM_001278638.1,NM_001459.3	+
chr19	49474560	49486231	NM_001204503.1	+
chr19	49889653	49896890	NM_152899.1	-
chr19	50358471	50365830	NM_004851.2	-
chr19	54353623	54365114	NM_001289023.2,NM_001289025.2,NM_002287.5,NM_021706.4	-
chr19	54353623	54370556	NM_001289026.2,NM_001289027.2	-
chr19	544033	549920	NM_001258351.1,NM_005317.3	+
chr19	54874092	54891420	NM_002000.3,NM_133269.3,NM_133271.3	+
chr19	55349701	55354814	NM_144613.4	-
chr19	6585838	6591152	NM_001252.4	-
chr19	6772667	6857366	NM_001258206.1,NM_001258207.1,NM_005428.3	+
chr19	6887548	6940453	NM_001256252.1,NM_001256253.1,NM_001256254.1, NM_001256255.1,NM_001974.4	+
chr1	97077742	97921059	NM_000110.3	-
chr19	7669085	7670454	NM_001193374.1,NM_020415.3	+
chr19	7677056	7679833	NM_174918.2	+
chr1	97719757	97921059	NM_001160301.1	-
chr19	8490055	8502654	NM_032152.4	-
chr19	9853717	9913705	NM_001304347.1	-
chr20	13389388	13638936	NM_001323602.1,NM_001323603.1,NM_001323604.1,NM_017714.2	-
chr20	1564382	1588697	NM_001330639.1	-
chr20	1629151	1657779	NM_001039508.1,NM_018556.3,NM_080816.2	-
chr20	21511016	21514026	NM_002509.3	-
chr20	36612518	36646216	NM_032214.3,NM_175077.2	-
chr20	5750386	5864407	NM_001303477.1,NM_001303478.1,NM_152504.3	+
chr20	57603845	57620576	NM_001160417.1,NM_001160418.1,NM_030776.2	-
chr20	57612884	57620576	NM_001160419.2,NM_001323966.1	-
chr20	59191019	59259113	NM_178457.2	+
chr20	63553019	63556708	NM_024059.3	+
chr20	63639704	63652766	NM_001276310.1	-
chr20	63639704	63653610	NM_015894.3	-
chr2	101991843	102024756	NM_001261419.1	+
chr2	102418789	102452565	NM_003853.3	+
chr2	108894470	108989372	NM_022336.3	-
chr2	112915192	112918882	NM_173205.1	+
chr2	112978018	112985672	NM_001278568.1,NM_019618.3	+
chr21	14485227	14546360	NM_001286523.1,NM_022136.4	-
chr21	14485227	14583402	NM_001256370.1	-
chr2	119759630	119984899	NM_002830.3	+
chr21	29077470	29109791	NM_001286623.1	+
chr21	31118417	31344354	NM_001353686.1,NM_001353687.1,NM_001353694.1	-
chr21	31118417	31559087	NM_001353688.1,NM_001353689.1,NM_001353690.1, NM_001353691.1,NM_001353692.1,NM_001353693.1, NM_003253.3	-
chr21	42371886	42389149	NM_032404.2	-
chr21	42371886	42396091	NM_001256317.1,NM_024022.2	-
chr21	42381875	42396091	NM_032405.1	-
chr21	42403861	42447680	NM_001001895.2,NM_001243467.1,NM_018961.3	+

chr21	42653751	42775509	NM_001001567.1,NM_001001568.1,NM_001001569.1, NM_001001570.1,NM_001001571.1,NM_001001572.1, NM_001001573.1,NM_001001574.1,NM_001001575.1, NM_001001576.1,NM_001001577.1,NM_001001578.1, NM_001001579.1,NM_001001580.1,NM_001001581.1, NM_001001582.1,NM_001001583.1,NM_001001584.2, NM_001001585.1,NM_001315533.1,NM_002606.2	+
chr21	42917250	42925598	NM_001308491.1	-
chr2	143129329	143768352	NM_018460.3	+
chr21	44285837	44298219	NM_000383.3	+
chr21	44627122	44628378	NM_198690.2	+
chr21	44654212	44654659	NM_198698.1	-
chr21	44657931	44658341	NM_198697.2	+
chr2	157414618	157444092	NM_004288.4	-
chr2	159768627	159904756	NM_001198759.1,NM_001198760.1	-
chr2	159803354	159904756	NM_002349.3	-
chr2	168456248	168775134	NM_001256126.1,NM_203463.2	+
chr2	174072446	174248670	NM_001011708.2,NM_001328688.1,NM_013341.4	-
chr2	174747591	174764472	NM_000079.3,NM_001039523.2	-
chr2	178829756	179050059	NM_173648.3	-
chr2	178873020	179050059	NM_001316745.1	-
chr2	190506892	190534742	NM_001142645.1	-
chr2	191029575	191151260	NM_003151.3	-
chr2	191029575	191151596	NM_001243835.1	-
chr2	195656807	195737702	NM_001127257.1	+
chr2	195657127	195737702	NM_020342.2	+
chr2	196986661	197310797	NM_001195144.1	-
chr2	197099470	197310797	NM_153697.2	-
chr2	197804701	198149884	NM_006226.3	+
chr2	202635177	202769757	NM_173511.3	+
chr2	202773149	202871472	NM_001288623.1	-
chr2	202773149	202871985	NM_001288622.1,NM_138468.5	-
chr2	203706474	203738912	NM_001243077.1,NM_001243078.1,NM_006139.3	+
chr2	203867787	203873960	NM_001037631.2,NM_005214.4	+
chr2	203936747	203961577	NM_012092.3	+
chr2	20684013	20823130	NM_001282719.1,NM_001282720.1,NM_001282721.1, NM_001282722.1,NM_001282723.1,NM_001282724.1,NM_021925.3	-
chr2	207821287	208025560	NM_001080475.2	-
chr2	212999683	213150334	NM_016260.2	-
chr2	212999683	213151609	NM_001079526.1	-
chr22	17178789	17219585	NM_001282225.1,NM_001282226.1,NM_001282228.1,NM_001282229.1	-
chr22	17178789	17221854	NM_001282227.1	-
chr2	218880532	218893929	NM_025216.2	+
chr2	219161458	219170095	NM_001144889.1,NM_001144890.1,NM_144712.4	-
chr2	221418026	221572358	NM_001304537.1,NM_004438.4	-
chr2	221418026	221574202	NM_001304536.1	-
chr22	22547695	22557754	NM_001291715.1,NM_001291716.1,NM_001291719.1	-
chr22	22547695	22559361	NM_001291717.1,NM_006115.4,NM_206953.2, NM_206954.2,NM_206955.2,NM_206956.2	-
chr22	23608451	23632321	NM_016449.3	-
chr22	26599397	26618027	NM_001887.3	-
chr2	230225729	230239078	NM_001005176.2	+
chr2	230225729	230313215	NM_001278451.1,NM_001278452.1,NM_001278453.1,NM_007237.4	+
chr22	30262827	30265817	NM_001319108.1	-
chr2	230907327	230925226	NM_005683.3	-
chr2	231523159	231530471	NM_006056.4	-
chr22	32190434	32203477	NM_001098527.2	-
chr22	32190434	32204731	NM_001159545.1	-
chr22	36860987	36878017	NM_000631.4,NM_013416.3	+
chr22	36913632	36940439	NM_000395.2	+
chr22	37125837	37149990	NM_000878.4	-
chr22	37125837	37175056	NM_001346223.1	-
chr22	37125837	37175118	NM_001346222.1	-
chr22	37204236	37210689	NM_001278687.2	-
chr22	37204236	37212376	NM_001051.4	-
chr22	37225260	37244299	NM_002872.4	-
chr22	37282382	37315348	NM_001318024.1,NM_013385.4	+
chr22	37469062	37486387	NM_001166343.1	-
chr22	37469062	37486440	NM_002405.3	-
chr22	39077004	39087743	NM_001349436.1,NM_001349437.1,NM_021822.3	+
chr22	39097223	39102605	NM_001166004.2	+
chr22	39097223	39104067	NM_001166002.2,NM_001166003.2,NM_181773.4	+
chr22	39901081	39973342	NM_004810.3	+
chr22	39926590	39973342	NM_001291824.1,NM_001291826.1	+
chr22	39926627	39973342	NM_001291828.1	+
chr22	39946816	39973342	NM_001291825.1	+
chr2	240605407	240631259	NM_001195381.1,NM_001195382.1	+
chr2	241849880	241858906	NM_005018.2	-
chr22	44172955	44208469	NM_001137605.2	+
chr22	44180889	44208469	NM_022141.6	+
chr22	49620455	49657542	NM_001289922.2	-
chr22	50546202	50551023	NM_138433.4	+
chr2	26308172	26319102	NM_001145169.1	-
chr2	30447225	30644225	NM_001002257.2,NM_001304445.1,NM_001304446.1,NM_182551.4	+
chr2	42494568	42756947	NM_001282755.1	+
chr2	60451166	60553498	NM_018014.3,NM_138559.1	-
chr2	60457193	60553498	NM_022893.3	-
chr2	63050801	63057831	NM_014562.3	+
chr2	68365189	68397453	NM_002664.2	+
chr2	84821665	84881245	NM_001080824.2,NM_001277053.1	-
chr2	84835918	84881245	NM_001307978.1	-
chr2	85694290	85698851	NM_001302758.1,NM_006433.4,NM_012483.3	+
chr2	86784604	86808396	NM_001145873.1	-
chr2	86815336	86861924	NM_001178100.1,NM_172101.3,NM_172102.3,NM_172213.3	-
chr2	86841565	86861924	NM_004931.4	-
chr2	96123849	96138436	NM_001002036.3	-
chr2	97713567	97739860	NM_001079.3	+
chr2	97734405	97739860	NM_207519.1	+
chr2	99547252	100105583	NM_001025108.1	-
chr2	99547252	100142575	NM_002285.2	-
chr3	100114929	100178632	NM_001167924.1	+
chr3	101324188	101513241	NM_001077203.2,NM_001282801.1,NM_001282802.1,NM_001282803.1,NM_020654.4	-
chr3	101371852	101513241	NM_001282804.1	-
chr3	108822697	108855005	NM_001317747.1,NM_016388.3	+
chr3	111542078	111607078	NM_001318889.1	+
chr3	111542078	111652359	NM_005816.4,NM_198196.2	+
chr3	112086334	112118197	NM_024616.2	+
chr3	112086334	112118226	NM_001171747.1	+
chr3	112120840	112133305	NM_001190259.1,NM_001190260.1,NM_152785.4	-
chr3	112463965	112499561	NM_001085357.1,NM_181780.3	-
chr3	112922684	112975090	NM_138806.3,NM_170780.2	-
chr3	112928917	112975090	NM_138939.2,NM_138940.2	-
chr3	113286929	113441514	NM_001164496.1	-
chr3	113362864	113441514	NM_018338.3	-
chr3	113532370	113629575	NM_001308350.1,NM_001322294.1,NM_001322295.1, NM_001322296.1,NM_001322297.1,NM_001322298.1, NM_001322299.1,NM_001322300.1,NM_017699.2	+
chr3	114234630	114237578	NM_007136.3	-
chr3	114293985	114310288	NM_173799.3	+
chr3	119821952	120094417	NM_001146156.1,NM_001354596.1,NM_002093.3	-

chr3	121894323	121944187	NM_001145998.1,NM_021082.3	+
chr3	122577601	122639047	NM_001113523.2,NM_001308320.1	+
chr3	151197831	151203201	NM_013308.3	-
chr3	160256985	160383766	NM_001190242.1	-
chr3	160256985	160399532	NM_001190241.1,NM_020800.2	-
chr3	16933089	17090606	NM_015184.5	+
chr3	183487527	183555712	NM_130446.3	-
chr3	184249656	184258759	NM_032331.3	+
chr3	195746764	195811973	NM_018406.6	-
chr3	27715948	27722715	NM_001278182.1,NM_001278183.1,NM_005442.3	-
chr3	32238678	32370325	NM_001320308.1,NM_178868.4	+
chr3	32951573	32954911	NM_005508.4	+
chr3	39263493	39280036	NM_001337.3	-
chr3	39263493	39280439	NM_001171172.1	-
chr3	39263493	39281273	NM_001171171.1	-
chr3	39263493	39281735	NM_001171174.1	-
chr3	39329705	39333680	NM_005201.3	+
chr3	42489298	42537573	NM_001251882.1	+
chr3	42502611	42537573	NM_001251883.1,NM_001251884.1,NM_001251885.1,NM_004624.3	+
chr3	45823315	45915724	NM_001276378.1,NM_001276379.1	-
chr3	45886503	45903175	NM_001256369.1,NM_006641.3,NM_031200.2	+
chr3	46353743	46360922	NM_001123041.2,NM_001123396.2	+
chr3	46370141	46376206	NM_000579.3,NM_001100168.1	+
chr3	52221079	52226163	NM_017442.3	-
chr3	52433034	52445085	NM_020163.2	-
chr3	56727417	56916471	NM_001289698.1	-
chr3	56727417	57079308	NM_001128615.1	-
chr3	59747277	60081127	NM_001320901.1	-
chr3	66220741	66378927	NM_001164796.1,NM_001350993.1	+
chr3	70954713	71543557	NM_001244816.1	-
chr3	70954713	71583753	NM_001244808.1,NM_001349338.1,NM_001349340.1,NM_001349341.1	-
chr3	70954713	71583989	NM_001244810.1,NM_032682.5	-
chr3	71197881	71583989	NM_001012505.1	-
chr3	72749276	72848447	NM_018130.2	-
chr4	102251040	102345253	NM_001135147.1	-
chr4	102261663	102324312	NM_001135148.1	-
chr4	102261663	102345191	NM_022154.5	-
chr4	102261663	102345498	NM_001135146.1	-
chr4	108047544	108166797	NM_001166119.1	-
chr4	108047544	108168956	NM_001130713.2,NM_001130714.2,NM_016269.4	-
chr4	109560198	109687716	NM_017918.4	+
chr4	112145396	112189081	NM_152400.2	+
chr4	121131407	121164340	NM_024873.5	-
chr4	121131407	121216627	NM_001128843.2	-
chr4	121131407	121227466	NM_001244764.1	-
chr4	122451469	122456495	NM_000586.3	-
chr4	122612627	122621057	NM_001207006.2,NM_021803.3	-
chr4	13367722	13484365	NM_001017979.2,NM_001159601.1,NM_004249.3	-
chr4	142023159	142846535	NM_001101669.2,NM_003866.3	-
chr4	142028955	142846535	NM_001331040.1	-
chr4	150264658	151015267	NM_001199282.2	-
chr4	150264658	151015497	NM_006726.4	-
chr4	152626113	152680165	NM_152680.2	-
chr4	15778264	15853243	NM_001775.3	+
chr4	15960239	15963236	NM_031950.3	-
chr4	173168752	173323967	NM_017423.2	+
chr4	177309836	177362938	NM_018248.2	+
chr4	25747426	25863031	NM_015187.4	-
chr4	25747426	25863595	NM_001297592.1,NM_001297594.1	-
chr4	26860690	27025381	NM_001169117.1,NM_001169118.1,NM_020860.3	+
chr4	38966743	39032421	NM_001303228.1,NM_024943.2	-
chr4	40191010	40244764	NM_001278368.1	+
chr4	40192966	40244764	NM_001278364.1,NM_001278365.1,NM_001278366.1,NM_001278367.1	+
chr4	40193648	40244764	NM_001278369.1	+
chr4	40196906	40244764	NM_001278359.1,NM_001278360.1,NM_001278361.1, NM_001278362.1,NM_001278363.1,NM_004310.4	+
chr4	48066392	48134256	NM_003328.2	-
chr4	6026198	6200591	NM_001099433.1	-
chr4	6053754	6194619	NM_001306133.1,NM_001306134.1	-
chr4	6053754	6200591	NM_144720.3	-
chr4	67558696	67607337	NM_001317769.1	+
chr4	67558696	67607341	NM_012108.3	+
chr4	69726967	69760712	NM_014465.3	-
chr4	70993547	71030914	NM_000788.2	+
chr4	83090047	83109859	NM_001130716.1	-
chr4	83090047	83114758	NM_016619.2	-
chr4	83094153	83114758	NM_001130715.1	-
chr4	83292460	83334813	NM_001199830.1	-
chr4	83292460	83334881	NM_001098540.2	-
chr4	83292460	83335153	NM_001166498.2,NM_006665.5	-
chr4	89244277	89308010	NM_198281.2	-
chr4	99816823	99870190	NM_014395.2	+
chr4	99816832	99870190	NM_001306151.1	+
chr5	100806934	100903285	NM_005668.5	-
chr5	100885310	100903285	NM_175052.2	-
chr5	111223652	111494886	NM_001323375.1	+
chr5	111223949	111494886	NM_001323374.1	+
chr5	111224236	111494886	NM_001323376.1,NM_001744.5	+
chr5	111224541	111494886	NM_001323377.1	+
chr5	119333174	119394604	NM_001286814.1	+
chr5	127042557	127073492	NM_001164479.1	-
chr5	127051303	127073492	NM_001164478.1	-
chr5	131264008	131394690	NM_001038702.1,NM_020240.2	+
chr5	132692626	132737578	NM_001300791.1,NM_001300792.1,NM_007054.6	-
chr5	14581781	14616180	NM_019018.2	+
chr5	149374266	149379275	NM_014443.2	-
chr5	149960736	149987400	NM_000112.3	+
chr5	154474971	154478264	NM_004821.2	-
chr5	157085831	157109237	NM_032782.4	-
chr5	157180895	157255099	NM_005546.3	+
chr5	169637246	170083382	NM_004946.2	+
chr5	169861305	169980757	NM_001129891.2,NM_001346304.1	-
chr5	175959530	175968315	NM_032361.3	-
chr5	177384430	177390513	NM_001167579.1	+
chr5	177384430	177398848	NM_003052.4	+
chr5	35617886	35671574	NM_144722.3	+
chr5	35617886	35814611	NM_024867.3	+
chr5	39105251	39219606	NM_001349333.1,NM_001465.5,NM_199335.4	-
chr5	39105251	39270657	NM_001243093.1	-
chr5	43376644	43412391	NM_001301873.1,NM_001301874.1	-
chr5	43379192	43397264	NM_001301875.1	-
chr5	43379192	43412391	NM_148672.3	-
chr5	43486700	43515171	NM_198566.3	-
chr5	50396196	50441400	NM_198449.2	-
chr5	50665898	50846522	NM_001178055.1	+
chr5	50666937	50846522	NM_001178056.1,NM_024615.3	+
chr5	50667545	50846522	NM_001331028.1	+
chr5	55024278	55034132	NM_002104.2	+
chr5	55102645	55110252	NM_006144.3	+

chr5	56099679	56233359	NM_024669.2	-
chr5	64768917	65018763	NM_001297644.1,NM_005869.3	+
chr5	67182274	67196789	NM_005582.2	-
chr5	75027463	75030899	NM_016591.2	-
chr5	80649999	80651035	NM_001190470.1	-
chr5	80960688	81230162	NM_006909.2	+
chr5	81412806	81751253	NM_001256732.2,NM_001256733.2,NM_001256734.2, NM_001256735.2,NM_001256736.2,NM_001345886.1,NM_012446.4	-
chr5	95555120	95605102	NM_198150.2	+
chr6	106184475	106325435	NM_001286106.1	-
chr6	106184475	106325820	NM_001286107.1,NM_001286108.1,NM_001286111.1,NM_004849.3	-
chr6	107065180	107114432	NM_001080450.2	-
chr6	107702156	107732396	NM_001286409.1	-
chr6	107702156	107824317	NM_001286408.1,NM_198081.4	-
chr6	11183297	11382348	NM_001142393.1	-
chr6	121079493	121334500	NM_152730.5	-
chr6	127708193	127855188	NM_001318531.1	-
chr6	127708193	127901081	NM_001010923.2,NM_001164685.1	-
chr6	127708193	127918631	NM_001164687.1	-
chr6	130144301	130215885	NM_001277185.1	-
chr6	130144301	130222954	NM_001017373.3	-
chr6	130144301	130365425	NM_001258275.2	-
chr6	131250158	131283535	NM_004842.3,NM_138633.2	+
chr6	132743869	132758008	NM_001242350.2,NM_004665.4	-
chr6	132743869	132763459	NM_078488.2	-
chr6	135181314	135219173	NM_001130172.1,NM_001130173.1,NM_001161656.1, NM_001161657.1,NM_001161658.1,NM_001161659.1, NM_001161660.1,NM_005375.3	+
chr6	136822563	136913934	NM_000288.3	+
chr6	136922240	136925660	NM_001008783.2	+
chr6	144150516	144191939	NM_003764.3	+
chr6	149447629	149485012	NM_207360.2	-
chr6	154039410	154246867	NM_001008503.2	+
chr6	154154482	154330081	NM_001130699.1	-
chr6	154154482	154356766	NM_001130700.1,NM_015553.2	-
chr6	15522800	15663058	NM_032122.4	-
chr6	15522806	15663058	NM_001271667.1,NM_001271668.1,NM_001271669.1	-
chr6	15523562	15663058	NM_183040.2	-
chr6	159034467	159042353	NM_001278733.1	-
chr6	159034467	159045152	NM_054114.4,NM_152133.2	-
chr6	159800248	159820704	NM_001271862.1,NM_173516.2	+
chr6	167111806	167139141	NM_004367.5	+
chr6	167122752	167139141	NM_031409.3	+
chr6	20099683	20212464	NM_001080480.2	-
chr6	20534456	21232404	NM_017774.3	+
chr6	24797320	24798896	NM_001282492.1	-
chr6	24839358	25042288	NM_001286446.2	-
chr6	25279427	25620530	NM_001173977.1,NM_017640.5	+
chr6	30070265	30075849	NM_170769.2	-
chr6	30070265	30075851	NM_025236.3	-
chr6	31572098	31574324	NM_001159740.2	+
chr6	31572293	31574324	NM_000595.3	+
chr6	31575566	31578336	NM_000594.3	+
chr6	31586178	31588909	NM_205838.2,NM_205839.2	+
chr6	31586698	31588909	NM_205837.2	+
chr6	31587199	31588909	NM_001166538.1,NM_007161.3,NM_205840.2	+
chr6	31588882	31592985	NM_001145466.1,NM_147130.2	-
chr6	31589273	31592985	NM_001145467.1	-
chr6	31706868	31717837	NM_001353334.1	+
chr6	32659463	32666689	NM_001243961.1,NM_002123.4	-
chr6	32934628	32941070	NM_002118.4	-
chr6	391738	411443	NM_001195286.1,NM_002460.3	+
chr6	41189748	41201187	NM_024807.3	-
chr6	41934932	42048894	NM_001136017.3,NM_001136126.2,NM_001287434.1	-
chr6	5108418	5260950	NM_001164841.2,NM_020408.5	-
chr6	5136173	5260950	NM_001318782.1	-
chr6	5186599	5260950	NM_001164840.2	-
chr6	52186386	52190638	NM_002190.2	+
chr6	5261017	5771592	NM_001318872.1	+
chr6	5261325	5771592	NM_006567.4	+
chr6	6588700	6654983	NM_004271.3	+
chr6	70667770	70862015	NM_001044305.2,NM_001281439.1,NM_021940.4	+
chr6	70668529	70862015	NM_001281440.1	+
chr6	82892397	83065841	NM_001304437.1,NM_001350602.1,NM_001350603.1, NM_001350604.1,NM_198920.2	-
chr6	87322587	87342328	NM_001042493.2,NM_001287445.1,NM_020425.5	+
chr6	87344852	87365463	NM_001010868.2	+
chr6	88609896	88963629	NM_001286426.1,NM_001286428.1,NM_003800.4	-
chr6	89926527	90296908	NM_001170794.1,NM_021813.3	-
chr6	97142160	97283277	NM_001350599.1,NM_001350600.1,NM_198468.3	-
chr7	100159241	100168750	NM_024637.4	-
chr7	100177723	100214387	NM_001282716.1	+
chr7	100177742	100214387	NM_001282717.1,NM_001282718.1,NM_012447.3	+
chr7	106865277	106908978	NM_001282426.1,NM_001282427.1	+
chr7	106865478	106908978	NM_002649.3	+
chr7	107201743	107564514	NM_006348.3,NM_181733.2	-
chr7	107207852	107564514	NM_001161520.1	-
chr7	111091005	111125454	NM_001099658.1,NM_001099660.1,NM_018334.4	+
chr7	112819146	112939877	NM_152556.2	-
chr7	132784861	133082158	NM_001317177.1,NM_017812.3	-
chr7	133253066	133484197	NM_001037126.1	+
chr7	135361794	135510127	NM_001190847.1,NM_001190849.1,NM_001190850.1,NM_013316.3	-
chr7	135387069	135510127	NM_001008225.2,NM_001190848.1	-
chr7	140404042	140426250	NM_001008749.2	+
chr7	141708352	141731271	NM_001105558.1	+
chr7	148807371	148883509	NM_001203249.1	-
chr7	148807371	148884349	NM_001203247.1,NM_001203248.1,NM_004456.4,NM_152998.2	-
chr7	150450629	150479395	NM_175571.3	+
chr7	150991055	151014599	NM_000603.4	+
chr7	150993803	151003935	NM_001160109.1,NM_001160110.1,NM_001160111.1	+
chr7	151381120	151410038	NM_001284260.1,NM_198285.2	-
chr7	151381120	151410705	NM_001284261.1,NM_001284262.1	-
chr7	151428831	151440170	NM_001308292.1,NM_144727.2	-
chr7	156680875	156893208	NM_001350953.1,NM_001350954.1,NM_001350955.1, NM_001350956.1,NM_001350957.1,NM_001350958.1,NM_022458.3	-
chr7	21428033	21514822	NM_001326542.1,NM_001326543.1,NM_003112.4	+
chr7	25134696	25180198	NM_138811.3	-
chr7	26152226	26187137	NM_004289.6	+
chr7	2906074	3043945	NM_001324281.1,NM_032415.5	-
chr7	36512940	36724549	NM_001177506.1,NM_001177507.1,NM_001637.3	-
chr7	44962660	44979105	NM_033054.2	-
chr7	50304082	50405101	NM_001291839.1	+
chr7	50304668	50369444	NM_001291845.1	+
chr7	50304668	50369839	NM_001291846.1	+
chr7	50304668	50405101	NM_001220765.2,NM_006060.5	+
chr7	50308679	50369839	NM_001291847.1	+
chr7	50308679	50405101	NM_001291837.1,NM_001291838.1	+
chr7	50319047	50405101	NM_001220767.2,NM_001220768.2,NM_001220770.2, NM_001220771.2,NM_001291840.1,NM_001291841.1, NM_001291842.1,NM_001291843.1,NM_001291844.1	+

chr7	549196	712524	NM_002735.2	-
chr7	549196	712940	NM_001164762.1	-
chr7	549196	713207	NM_001164761.1	-
chr7	549196	727341	NM_001164760.1	-
chr7	549196	727676	NM_001164758.1,NM_001164759.1	-
chr7	72574506	72828234	NM_001145440.2	-
chr8	116644815	116755823	NM_003756.2	-
chr8	120445397	120523635	NM_022045.4	+
chr8	132709944	132760668	NM_001145153.1,NM_144649.2	-
chr8	132866959	133134899	NM_003235.4	+
chr8	133036727	133103066	NM_001045556.2,NM_001045557.2,NM_001282964.1, NM_001282965.1	-
chr8	134477787	134713049	NM_001174157.1,NM_001289394.1,NM_020863.3	-
chr8	139727724	140457762	NM_001160372.3,NM_001321646.1	-
chr8	139727724	140458579	NM_031466.7	-
chr8	141207165	141254629	NM_001286646.1	-
chr8	142834800	142846846	NM_002066.2	+
chr8	143368854	143384220	NM_052924.2	+
chr8	19404160	19602545	NM_001130518.1,NM_001354491.1,NM_001354492.1, NM_001354494.1,NM_001354495.1,NM_001354496.1, NM_001354497.1,NM_001354498.1,NM_001354499.1	-
chr8	19404160	19682583	NM_018371.4	-
chr8	19404160	19757083	NM_001354490.1	-
chr8	19404160	19757306	NM_001354483.1,NM_001354484.1,NM_001354485.1, NM_001354487.1,NM_001354488.1,NM_001354489.1	-
chr8	19404160	19758029	NM_001354475.1,NM_001354476.1,NM_001354477.1, NM_001354480.1,NM_001354481.1	-
chr8	22042752	22048809	NM_001304478.1,NM_003867.3	+
chr8	22687660	22692389	NM_001199880.1	-
chr8	22687660	22692580	NM_001199881.1	-
chr8	22687660	22693302	NM_004430.2	-
chr8	23102813	23117437	NM_003841.4	+
chr8	30095405	30137706	NM_001128208.1	+
chr8	30131670	30144684	NM_001100916.1	-
chr8	42697375	42737407	NM_001347717.1	+
chr8	42697418	42737407	NM_000749.4	+
chr8	58805417	59119208	NM_014729.2	-
chr8	66562174	66613249	NM_001080416.3,NM_001144755.2,NM_001294282.1	-
chr8	71841541	71844496	NM_005098.3	-
chr8	73420368	73746927	NM_001164380.1,NM_001164381.1,NM_001164382.1, NM_001164383.1	-
chr8	94371959	94436952	NM_001205263.1	-
chr8	94371959	94475115	NM_012415.3	-
chr8	94427711	94475115	NM_001205262.2	-
chr8	99013265	99209787	NM_015243.2	+
chr8	99013265	99877586	NM_017890.4,NM_152564.4	+
chr9	114901144	114930595	NM_001244.3	-
chr9	123121619	123268564	NM_018387.4	-
chr9	123121619	123268576	NM_001171137.1	-
chr9	124942607	125143559	NM_001144877.2,NM_173690.4	-
chr9	20341664	20621987	NM_001286691.1	-
chr9	20341664	20622543	NM_004529.3	-
chr9	20658308	20995955	NM_017794.4	+
chr9	214864	465259	NM_203447.3	+
chr9	273047	465259	NM_001190458.1,NM_001193536.1	+
chr9	3218296	3526001	NM_001282116.1,NM_134428.2	-
chr9	3247036	3395596	NM_002919.3	-
chr9	3270928	3489487	NM_001282117.1	-
chr9	37120471	37358149	NM_001289119.1,NM_032226.2	+
chr9	37120597	37358149	NM_001289120.1,NM_001289121.1	+
chr9	5299863	5304228	NM_001329191.1	-
chr9	5299863	5304611	NM_005059.3,NM_134441.2	-
chr9	5334931	5339873	NM_006911.3	-
chr9	5357965	5437937	NM_018465.3	-
chr9	6757349	6815149	NM_001304341.2	+
chr9	75890643	76195889	NM_006200.5	+
chr9	75890643	76362339	NM_001190482.1	+
chr9	85546538	85742029	NM_001286715.1,NM_001286717.1,NM_001330701.1,NM_015239.2	-
chr9	92947318	93036236	NM_001083536.1	+
chr9	92963960	93036236	NM_001286993.1	+
chr9	92974435	93036236	NM_033086.2	+
chr9	96385940	96418387	NM_153695.3	-
chr9	97926790	97938393	NM_197978.2	-
chr9	97926790	97944915	NM_018437.4	-
chr9	98807698	98850081	NM_024642.4	+
chr9	99821854	99866891	NM_006981.3,NM_173200.2	+
chrX	108044969	108079184	NM_001170553.1,NM_182607.4	+
chrX	118495908	118686160	NM_144658.3	+
chrX	124346281	124373160	NM_001114937.2,NM_002351.4	+
chrX	129738969	129769549	NM_003399.5	+
chrX	129779915	129795201	NM_018990.3	+
chrX	132023216	132075943	NM_001042453.1,NM_016542.3	+
chrX	132023607	132075943	NM_001042452.1	+
chrX	135520625	135582535	NM_001351601.1,NM_001351603.1,NM_001351604.1, NM_001351605.1,NM_001351606.1,NM_182540.5	+
chrX	136648176	136660390	NM_000074.2	+
chrX	1462571	1537144	NM_178129.4	-
chrX	152988823	152992217	NM_001103150.1,NM_001103151.1,NM_052926.2	-
chrX	153802165	153829896	NM_001303515.1,NM_001303516.1	-
chrX	153802165	153830567	NM_001303512.1,NM_001303513.1,NM_001303514.1,NM_032512.3	-
chrX	153905074	153907166	NM_000054.4,NM_001146151.1	+
chrX	153907375	153926260	NM_001164741.1,NM_001666.4	-
chrX	154295673	154330363	NM_001145933.1,NM_012253.3	+
chrX	154305056	154330363	NM_001145934.1	+
chrX	154675248	154751083	NM_001081573.2,NM_001282283.1,NM_080612.3	-
chrX	15738288	15787625	NM_007220.3	+
chrX	19533964	19799799	NM_001024666.2	-
chrX	19533964	19833151	NM_001353893.1,NM_001353894.1,NM_001353895.1	-
chrX	19533964	19887626	NM_001353890.1,NM_001353891.1,NM_001353892.1,NM_031892.2	-
chrX	24465226	24534425	NM_005391.4	+
chrX	24465226	24550466	NM_001142386.2	+
chrX	3604342	3713634	NM_005044.4	-
chrX	41514933	41923034	NM_001126054.2,NM_001126055.2,NM_003688.3	-
chrX	41688972	41697277	NM_001097579.1,NM_005300.3	+
chrX	48683796	48691427	NM_000377.2	+
chrX	49250435	49264826	NM_001114377.1,NM_014009.3	-
chrX	71103888	71106788	NM_001025265.2	-
chrX	71107403	71111631	NM_000206.2	-
chrX	78945331	78963727	NM_001324218.1,NM_001324221.1,NM_001324225.1, NM_014499.3,NM_198333.2	+
chrX	79170971	79172229	NM_032553.1	+
chrY	1462571	1537144	NM_178129.4	-
chrY	2935476	2982506	NM_001145275.1,NM_003411.3	+

Table B.2: List of 63 complex traits.

Trait Identifier	Trait	h2g	Z	N	Reference	AIIS
PASS_ADHD_Demontis2018	ADHD	248	11.634	55374	Demontis et al. 2019 Nat Genet	0
PASS_AgeFirstBirth	Age first birth	62	15.073	222037	Barban et al., 2016 Nat Genet	0
PASS_Alzheimers_Jansen2019	Alzheimer's Disease	14	7.833	433886	Jansen et al. 2019 Nat Genet	0
PASS_Anorexia	Anorexia	236	8.274	32143	Boraska et al., 2014 Mol Psych	0
PASS_AtrialFibrillation_Nielsen2018	Atrial Fibrilation	22	16.769	1030836	Nielsen et al. 2018 AJHG	0
PASS_Autism	Autism Spectrum	622	6.191	10263	PGC Cross-Disorder Group, 2013 Lancet	0
PASS_BDSCZ_Ruderfer2018	Bipolar and/or Schizophrenia	334	21.656	10762	Ruderfer et al. 2018 Cell	0
PASS_BMI1	BMI (Speliotes 2010)	146	16.233	122033	Speliotes et al., 2010 Nat Genet	0
PASS_CigarettesPerDay_Liu2019	Cigarattes per Day	58	8.657	257073	Liu et al. 2019 Nat Genet	0
PASS_Coronary_Artery_Disease	Coronary Artery Disease	75	6.016	7721	Schunkert et al., 2011 Nat Genet	0
PASS_DrinksPerWeek_Liu2019	Drinks per Week	46	14.438	527299	Liu et al. 2019 Nat Genet	0
PASS_Ever_Smoked	Ever Smoked	85	7.482	74035	TAG Consortium, 2010 Nat Genet	0
PASS_GeneralRiskTolerance_KarlssonLinner2019	General Risk Tolerance	51	18.357	466571	Karlsson Linner et al. 2019 Nat Genet	0
PASS_HDL	HDL	125	9.259	97749	Teslovich et al., 2010 Nature	0
PASS_Height1	Height (Lango Allen 2010)	222	17.512	131547	Lango Allen et al., 2010 Nature	0
PASS_IBD_deLange2017	IBD	321	13.193	59957	de Lange et al. 2017 Nat Genet	1
PASS_Insomnia_Jansen2019	Insomnia	44	17.6	383948	Jansen et al. 2019 Nat Genet	0
PASS_Intelligence_SavageJansen2018	Intelligence	186	29.047	264193	Savage et al. 2018 Nat Genet	0
PASS_IschemicStroke_Malik2018	Ischemic Stroke	17	8.35	440328	Malik et al. 2018 Nat Genet	0
PASS_LDL	LDL	103	8.821	93354	Teslovich et al., 2010 Nature	0
PASS_MDD_Wray2018	Major depressive disorder	73	12.339	170229	Wray et al. 2018 Nat Genet	0
PASS_MedicationUse_Wu2019	Medication Use	157	14.036	78808	Wu et al. 2019 Nat Comm	0
PASS_NumberChildrenEverBorn	Number children even born	23	9.542	318863	Barban et al., 2016 Nat Genet	0
PASS_ProstateCancer	Prostate Cancer	131	7.016	72729	Schumacher et al. 2018 Nat Genet	0
PASS_ReactionTime_Davies2018	Reaction Time	81	22.556	300486	Davies et al. 2018 Nat Comm	0
PASS_Rheumatoid_Arthritis	Rheumatoid Arthritis	187	7.51	37681	Okada et al., 2014 Nature	1
PASS_SCZvsBD_Ruderfer2018	Schizophrenia vs Bipolar	265	9.495	38855	Ruderfer et al. 2018 Cell	0
PASS_SleepDuration_Dashti2019	Sleep Duration	71	19.667	446118	Dashti et al. 2019 Nat Comm	0
PASS_Type_2_Diabetes	Type 2 Diabetes	95	6.463	60786	Morris et al., 2012 Nat Genet	0
PASS_Years_of_Education1	Years of Education	77	9.354	126559	Rietveld et al., 2013 Science	0
UKB_460K.biochemistry_AlkalinePhosphatase	AlkalinePhosphatase	235	14.675	433862	UKBiobank - 460K	0
UKB_460K.biochemistry_AspartateAminotransferase	AspartateAminotransferase	114	18.306	430982	UKBiobank - 460K	0
UKB_460K.biochemistry_Cholesterol	Cholesterol	131	11.67	435137	UKBiobank - 460K	0
UKB_460K.biochemistry_Creatinine	Creatinine	221	21.911	434158	UKBiobank - 460K	0
UKB_460K.biochemistry_IGF1	IGF1	288	18.682	432292	UKBiobank - 460K	0
UKB_460K.biochemistry_Phosphate	Phosphate	129	14.198	397561	UKBiobank - 460K	0
UKB_460K.biochemistry_Testosterone_Male	TestosteroneMale	184	16.577	196813	UKBiobank - 460K	0
UKB_460K.biochemistry_TotalBilirubin	TotalBilirubin	84	11.066	429423	UKBiobank - 460K	0
UKB_460K.biochemistry_TotalProtein	TotalProtein	181	24.147	397652	UKBiobank - 460K	0
UKB_460K.biochemistry_VitaminD	VitaminD	86	12.464	4157	UKBiobank - 460K	0
UKB_460K.blood_PLATELET_COUNT	Platelet Count	351	21.242	444382	UKBiobank - 460K	0
UKB_460K.blood_RBC_DISTRIB_WIDTH	Red Blood Cell Distribution Width	217	17.803	4427	UKBiobank - 460K	0
UKB_460K.blood_RED_COUNT	Red Blood Cell Count	261	21.775	445174	UKBiobank - 460K	0
UKB_460K.blood_WHITE_COUNT	White Blood Cell Count	227	27.386	444502	UKBiobank - 460K	1
UKB_460K.bmd_HEEL_TSCOREz	Heel T Score	355	20.514	445921	UKBiobank - 460K	0
UKB_460K.body_BALDING1	Balding Type I	223	14.697	208336	UKBiobank - 460K	0
UKB_460K.body_BMIz	BMI (UK Biobank)	278	41.478	457824	UKBiobank - 460K	0
UKB_460K.body_HEIGHTz	Height (UK Biobank)	674	25.824	458303	UKBiobank - 460K	0
UKB_460K.body_WHRadjBMIz	Waist-hip Ratio	171	23.681	458417	UKBiobank - 460K	0
UKB_460K.bp_DIASTOLICadjMEDz	Diastolic Blood Pressure	233	27.738	422771	UKBiobank - 460K	0
UKB_460K.cancer_BREAST	Breast Cancer (UKBB)	18	6.846	459324	UKBiobank - 460K	0
UKB_460K.cancer_PROSTATE	Prostate Cancer (UKBB)	17	7	459324	UKBiobank - 460K	0
UKB_460K.cov_EDU_YEARS	Years of Education (UK Biobank)	136	35.789	454813	UKBiobank - 460K	0
UKB_460K.disease_ALLERGY_ECZEMA_DIAGNOSED	Eczema	85	14.603	458699	UKBiobank - 460K	1
UKB_460K.disease_HYPOTHYROIDISM_SELF_REP	Hypothyroidism	55	15.167	459324	UKBiobank - 460K	1
UKB_460K.lung_FEV1FVCzSMOKE	FEV1-FVC Ratio	276	23.615	371949	UKBiobank - 460K	0
UKB_460K.lung_FVCzSMOKE	Forced Vital Capacity (FVC)	234	32.444	371949	UKBiobank - 460K	0
UKB_460K.mental_NEUROTICISM	Neuroticism (UK Biobank)	114	28.475	372066	UKBiobank - 460K	0
UKB_460K.other_MORNINGPERSON	Morning Person	106	24.674	41052	UKBiobank - 460K	0
UKB_460K.pigment_SUNBURN	Sunburn Occasion	73	9.618	344229	UKBiobank - 460K	0
UKB_460K.repro_MENARCHE_AGE	Age at Menarche (UK Biobank)	253	24.563	242278	UKBiobank - 460K	0
UKB_460K.repro_MENOPAUSE_AGE	Age at Menopause (UK Biobank)	109	11.04	143025	UKBiobank - 460K	0
UKB_460K.repro_NumberChildrenEverBorn_Pooled	Number children even born (UKBB, pooled)	39	16.783	4565	UKBiobank - 460K	0

Table B.3: List of S-LDSC results for 63 complex traits.

Trait Identifier	Prop. SNPs	Prop. h2	Prop. h2 std error	Enrichment	Enrichment std error	Enrichment p	Coefficient	Coefficient std error	Coefficient z-score	TauStar	TauStar se	Tau z value
PASS_ADHD_Demontis2018	0.03565	0.033046	0.010001	0.926973	0.280538	0.794481	-3.59E-09	1.31E-08	-0.27302	-0.02571	0.094179	-0.27302
PASS_AgeFirstBirth	0.03565	0.040618	0.008719	1.139366	0.244586	0.568072	1.61E-09	2.86E-09	0.563792	0.046267	0.082064	0.563792
PASS_Alzheimers_Jansen2019	0.03565	0.041906	0.024117	1.175503	0.676498	0.793454	2.69E-10	1.78E-09	0.151122	0.034294	0.226929	0.151122
PASS_Anorexia	0.03565	0.044572	0.014068	1.250275	0.394604	0.52361	1.07E-08	1.75E-08	0.612775	0.08087	0.131974	0.612775
PASS_AtrialFibrillation_Nielsen2018	0.03565	0.012626	0.009997	0.354154	0.280427	0.022484	-2.75E-09	1.17E-09	-2.34931	-0.22231	0.094629	-2.34931
PASS_Autism	0.03565	-0.00038	0.018777	-0.0106	0.526714	0.042522	-1.20E-07	5.93E-08	-2.02843	-0.33928	0.167261	-2.02843
PASS_BDSCZ_Ruderfer2018	0.03565	0.038974	0.006244	1.093249	0.175135	0.594401	5.01E-09	1.11E-08	0.45238	0.026635	0.058877	0.45238
PASS_BMI1	0.03565	0.02485	0.006262	0.697072	0.175663	0.086639	-8.80E-09	4.90E-09	-1.79403	-0.10654	0.059388	-1.79403
PASS_CigarettesPerDay_Liu2019	0.03565	0.034532	0.016551	0.968634	0.464267	0.946	-4.43E-10	5.11E-09	-0.0868	-0.01355	0.156097	-0.0868
PASS_Coronary_Artery_Disease	0.03565	0.022711	0.025215	0.637071	0.707305	0.6022	-5.17E-09	9.93E-09	-0.52093	-0.12276	0.235654	-0.52093
PASS_DrinksPerWeek_Liu2019	0.03565	0.037772	0.010078	1.059515	0.282705	0.833289	3.80E-10	2.49E-09	0.152748	0.014556	0.095297	0.152748
PASS_Ever_Smoked	0.03565	0.061577	0.019903	1.727261	0.558279	0.200725	1.17E-08	9.12E-09	1.277403	0.243599	0.190699	1.277403
PASS_GeneralRiskTolerance_KarlssonLinner2019	0.03565	0.03919	0.007783	1.099303	0.218319	0.649514	8.98E-10	2.13E-09	0.421922	0.030976	0.073416	0.421922
PASS_HDL	0.03565	0.01633	0.015448	0.458065	0.433333	0.211259	-1.35E-08	1.00E-08	-1.34791	-0.19507	0.144717	-1.34791
PASS_Height1	0.03565	0.018198	0.007431	0.510474	0.208442	0.021862	-2.20E-08	8.95E-09	-2.45456	-0.1745	0.071092	-2.45456
PASS_IBD_deLange2017	0.03565	0.154501	0.024402	4.333845	0.684482	1.79E-06	1.96E-07	4.07E-08	4.830536	1.101221	0.227971	4.830536
PASS_Insomnia_Jansen2019	0.03565	0.044712	0.009444	1.254196	0.264901	0.334148	2.04E-09	2.19E-09	0.932677	0.082374	0.08832	0.932677
PASS_Intelligence_SavageJansen2018	0.03565	0.033268	0.005878	0.933198	0.164895	0.685094	-2.51E-09	5.78E-09	-0.43503	-0.02398	0.055125	-0.43503
PASS_IschemicStroke_Malik2018	0.03565	0.006185	0.016853	0.173486	0.472749	0.072905	-2.68E-09	1.47E-09	-1.82342	-0.28123	0.154231	-1.82342
PASS_LDL	0.03565	0.011158	0.019824	0.313001	0.556069	0.214745	-1.47E-08	1.08E-08	-1.35206	-0.25196	0.186349	-1.35206
PASS_MDD_Wray2018	0.03565	0.055543	0.011326	1.558015	0.317711	0.076115	7.55E-09	4.30E-09	1.756101	0.185228	0.105477	1.756101
PASS_MedicationUse_Wu2019	0.03565	0.041816	0.009088	1.172966	0.254923	0.499492	4.89E-09	7.62E-09	0.641507	0.055039	0.085796	0.641507
PASS_NumberChildrenEverBorn	0.03565	-0.00386	0.014755	-0.10829	0.413879	0.005313	-4.86E-09	1.73E-09	-2.81472	-0.37301	0.132522	-2.81472
PASS_ProstateCancer	0.03565	0.037891	0.019152	1.062868	0.537222	0.90664	6.08E-10	1.32E-08	0.046042	0.008305	0.18038	0.046042
PASS_ReactionTime_Davies2018	0.03565	0.03969	0.006876	1.113343	0.192868	0.556874	1.56E-09	2.97E-09	0.526719	0.034174	0.064881	0.526719
PASS_Rheumatoid_Arthritis	0.03565	0.143014	0.033005	4.011642	0.92582	0.000468	1.03E-07	2.95E-08	3.504227	0.993677	0.283565	3.504227
PASS_SCZvsBD_Ruderfer2018	0.03565	0.032691	0.01345	0.91699	0.377285	0.826137	-4.93E-09	1.90E-08	-0.25865	-0.03289	0.127164	-0.25865
PASS_SleepDuration_Dashti2019	0.03565	0.038825	0.005931	1.089056	0.166367	0.591743	1.08E-09	2.23E-09	0.483206	0.027027	0.055932	0.483206
PASS_Type_2_Diabetes	0.03565	0.039149	0.020809	1.098147	0.583703	0.866391	1.39E-09	1.04E-08	0.133407	0.026205	0.196432	0.133407
PASS_Years_of_Education1	0.03565	0.055198	0.014364	1.548348	0.402915	0.165669	7.89E-09	5.73E-09	1.378243	0.182906	0.13271	1.378243
UKB_460K.biochemistry_AlkalinePhosphatase	0.03565	0.020413	0.010551	0.572607	0.295973	0.144314	-2.13E-08	1.30E-08	-1.63639	-0.1604	0.09802	-1.63639
UKB_460K.biochemistry_AspartateAminotransferase	0.03565	0.023487	0.010161	0.658814	0.285028	0.232269	-8.38E-09	6.17E-09	-1.35657	-0.13073	0.096365	-1.35657
UKB_460K.biochemistry_Cholesterol	0.03565	0.015685	0.015061	0.439977	0.422483	0.190209	-1.49E-08	1.06E-08	-1.41021	-0.20313	0.144044	-1.41021
UKB_460K.biochemistry_Creatinine	0.03565	0.022547	0.007286	0.632445	0.204366	0.072466	-1.69E-08	8.60E-09	-1.96098	-0.13497	0.068828	-1.96098
UKB_460K.biochemistry_IGF1	0.03565	0.03602	0.014778	1.010376	0.414546	0.979857	-2.77E-10	2.25E-08	-0.0123	-0.00171	0.138687	-0.0123
UKB_460K.biochemistry_Phosphate	0.03565	0.018212	0.010734	0.510853	0.301093	0.117807	-1.27E-08	7.70E-09	-1.65308	-0.17429	0.105431	-1.65308
UKB_460K.biochemistry_Testosterone_Male	0.03565	0.042588	0.010898	1.194621	0.305687	0.523656	5.94E-09	1.07E-08	0.557281	0.057179	0.102603	0.557281
UKB_460K.biochemistry_TotalBilirubin	0.03565	0.075509	0.033943	2.118063	0.95212	0.229451	1.73E-08	1.48E-08	1.170966	0.365134	0.311823	1.170966
UKB_460K.biochemistry_TotalProtein	0.03565	0.087866	0.014752	2.464704	0.413813	0.000516	4.84E-08	1.42E-08	3.415222	0.476059	0.139393	3.415222
UKB_460K.biochemistry_VitaminD	0.03565	0.025892	0.020094	0.726292	0.563638	0.617428	-4.98E-09	8.92E-09	-0.55775	-0.10221	0.183259	-0.55775
UKB_460K.blood_PLATELET_COUNT	0.03565	0.068915	0.012803	1.93311	0.359145	0.010444	5.82E-08	2.39E-08	2.433801	0.294578	0.121036	2.433801
UKB_460K.blood_RBC_DISTRIB_WIDTH	0.03565	0.04124	0.02007	1.156807	0.562988	0.779411	4.62E-09	2.32E-08	0.199287	0.037669	0.189017	0.199287
UKB_460K.blood_RED_COUNT	0.03565	0.056616	0.014209	1.588112	0.398563	0.145131	2.67E-08	1.99E-08	1.339784	0.180997	0.135094	1.339784
UKB_460K.blood_WHITE_COUNT	0.03565	0.069654	0.011492	1.953834	0.322368	0.003114	3.90E-08	1.38E-08	2.825905	0.304112	0.107616	2.825905
UKB_460K.bmd_HEEL_TSCOREz	0.03565	0.017258	0.008351	0.484091	0.234239	0.027512	-3.69E-08	1.57E-08	-2.34477	-0.18378	0.078379	-2.34477
UKB_460K.body_BALDING1	0.03565	0.026651	0.01394	0.747585	0.391023	0.515555	-1.21E-08	1.64E-08	-0.73663	-0.09558	0.129755	-0.73663
UKB_460K.body_BMIz	0.03565	0.03716	0.004039	1.042348	0.113292	0.708751	1.47E-09	5.97E-09	0.245774	0.009376	0.03815	0.245774
UKB_460K.body_HEIGHTz	0.03565	0.016157	0.00555	0.453226	0.155689	0.000536	-7.40E-08	1.97E-08	-3.74856	-0.19472	0.051946	-3.74856
UKB_460K.body_WHRadjBMIz	0.03565	0.008368	0.005828	0.234721	0.163483	5.25E-06	-2.56E-08	5.30E-09	-4.83116	-0.26534	0.054922	-4.83116
UKB_460K.bp_DIASTOLICadjMEDz	0.03565	0.026041	0.006888	0.730478	0.193209	0.162692	-1.30E-08	8.53E-09	-1.52445	-0.09892	0.064888	-1.52445
UKB_460K.cancer_BREAST	0.03565	0.008195	0.015563	0.229866	0.436539	0.085348	-2.75E-09	1.52E-09	-1.80663	-0.2704	0.149671	-1.80663
UKB_460K.cancer_PROSTATE	0.03565	0.056113	0.030439	1.573993	0.853838	0.49444	1.72E-09	2.63E-09	0.656195	0.184227	0.280751	0.656195
UKB_460K.cov_EDU_YEARS	0.03565	0.037966	0.004965	1.064976	0.139263	0.641252	1.52E-09	3.60E-09	0.422388	0.019803	0.046883	0.422388
UKB_460K.disease_ALLERGY_ECZEMA_DIAGNOSED	0.03565	0.146626	0.019756	4.112962	0.554156	4.27E-06	4.88E-08	1.05E-08	4.668972	1.03282	0.221209	4.668972
UKB_460K.disease_HYPOTHYROIDISM_SELF_REP	0.03565	0.221424	0.041861	6.211095	1.174235	5.30E-05	5.24E-08	1.28E-08	4.094165	1.734029	0.423537	4.094165
UKB_460K.lung_FEV1FVCzSMOKE	0.03565	0.022427	0.006066	0.629083	0.170153	0.029454	-2.05E-08	8.87E-09	-2.30845	-0.1313	0.056879	-2.30845
UKB_460K.lung_FVCzSMOKE	0.03565	0.028102	0.00551	0.788266	0.154565	0.171023	-1.03E-08	6.83E-09	-1.50152	-0.07782	0.051828	-1.50152
UKB_460K.mental_NEUROTICISM	0.03565	0.028737	0.005125	0.806093	0.143764	0.178289	-4.33E-09	3.11E-09	-1.39362	-0.0674	0.048364	-1.39362
UKB_460K.other_MORNINGPERSON	0.03565	0.034691	0.005477	0.973094	0.153645	0.861019	-7.47E-10	3.09E-09	-0.24135	-0.01247	0.05167	-0.24135
UKB_460K.pigment_SUNBURN	0.03565	0.077398	0.044092	2.171064	1.236804	0.343726	1.57E-08	1.70E-08	0.922141	0.381449	0.413656	0.922141
UKB_460K.repro_MENARCHE_AGE	0.03565	0.034436	0.006716	0.965962	0.188383	0.856782	-2.31E-09	9.06E-09	-0.25518	-0.01621	0.063537	-0.25518
UKB_460K.repro_MENOPAUSE_AGE	0.03565	0.040961	0.016607	1.148994	0.465849	0.747607	1.83E-09	9.73E-09	0.187662	0.029442	0.156888	0.187662
UKB_460K.repro_NumberChildrenEverBorn_Pooled	0.03565	0.029812	0.008575	0.836244	0.240542	0.494737	-1.23E-09	1.76E-09	-0.69748	-0.05632	0.080746	-0.69748

Table B.4: List of S-LDSC results for 176 complex traits.

Trait Identifier	IS trait	Prop. SNPs	Prop. h2	Prop. h2 std error	Enrichment	Enrichment std error	Enrichment p	TauStar	TauStar se	Tau z value	p-value
PASS_ADHD_Demontis2018	0	0.03565	0.033046	0.010001	0.926973	0.280538	0.794481	-0.02571	0.094179	-0.27302	6.08E-01
PASS_AdultOnsetAsthma_Ferreira2019	1	0.03565	0.15351	0.029342	4.306041	0.823048	9.50E-06	1.098167	0.245021	4.481936	3.70E-06
PASS_AgeFirstBirth	0	0.03565	0.040618	0.008719	1.139366	0.244586	0.568072	0.046267	0.082064	0.563792	2.86E-01
PASS_AgeOfInitiation_Liu2019	0	0.03565	0.054366	0.016685	1.525001	0.468022	0.257588	0.170039	0.156465	1.086758	1.39E-01
PASS_Alzheimer	0	0.03565	0.025597	0.029774	0.718016	0.835176	0.735907	-0.11791	0.280806	-0.4199	6.63E-01
PASS_Alzheimers_Jansen2019	0	0.03565	0.041906	0.024117	1.175503	0.676498	0.793454	0.034294	0.226929	0.151122	4.40E-01
PASS_AlzheimersMaternal_Marioni2018	0	0.03565	-0.01285	0.074036	-0.36056	2.076766	0.512924	-0.47781	0.703562	-0.67913	7.51E-01
PASS_AlzheimersPaternal_Marioni2018	0	0.03565	0.03067	0.060724	0.860313	1.703335	0.934556	-0.05057	0.571672	-0.08845	5.35E-01
PASS_AlzheimersProxy_Marioni2018	0	0.03565	0.026229	0.026245	0.73574	0.736196	0.721847	-0.11299	0.250186	-0.4516	6.74E-01
PASS_Anorexia	0	0.03565	0.044572	0.014068	1.250275	0.394604	0.52361	0.08087	0.131974	0.612775	2.70E-01
PASS_AnorexiaNervosa_Duncan_2017	0	0.03565	0.032795	0.033068	0.919923	0.927574	0.930805	-0.02045	0.310101	-0.06594	5.26E-01
PASS_AtrialFibrillation_Nielsen2018	0	0.03565	0.012626	0.009997	0.354154	0.280427	0.022484	-0.22231	0.094629	-2.34931	9.91E-01
PASS_Autism	0	0.03565	-0.00038	0.018777	-0.0106	0.526714	0.042522	-0.33928	0.167261	-2.02843	9.79E-01
PASS_Autism_Grove2019	0	0.03565	0.046958	0.017694	1.317214	0.496325	0.511069	0.108035	0.162219	0.665985	2.53E-01
PASS_BDSCZ_Ruderfer2018	0	0.03565	0.038974	0.006244	1.093249	0.175135	0.594401	0.026635	0.058877	0.45238	3.25E-01
PASS_BIP_Stahl2019	0	0.03565	0.041007	0.00875	1.15028	0.245441	0.53982	0.044874	0.082575	0.543435	2.93E-01
PASS_Bipolar_Disorder	0	0.03565	0.056177	0.021581	1.575789	0.605349	0.331413	0.186428	0.199104	0.936337	1.75E-01
PASS_BipolarDisorder_Ruderfer2018	0	0.03565	0.041279	0.011049	1.157898	0.309937	0.609755	0.04643	0.103952	0.446643	3.28E-01
PASS_BMI1	0	0.03565	0.02485	0.006262	0.697072	0.175663	0.086639	-0.10654	0.059388	-1.79403	9.64E-01
PASS_Bone_Mineral_Density_Kemp_2017	0	0.03565	0.018603	0.008351	0.521828	0.234261	0.041576	-0.1699	0.078428	-2.16636	9.85E-01
PASS_BreastCancer	0	0.03565	0.006836	0.011773	0.191762	0.330235	0.019778	-0.28103	0.115759	-2.42775	9.92E-01
PASS_CardioembolicStroke_Malik2018	0	0.03565	0.052043	0.068871	1.459846	1.93187	0.809398	0.167278	0.643029	0.26014	3.97E-01
PASS_CD_deLange2017	1	0.03565	0.163153	0.027114	4.576556	0.760575	4.77E-06	1.184072	0.256086	4.623726	1.88E-06
PASS_Celiac	1	0.03565	0.234031	0.059079	6.564713	1.657198	0.000317	1.854008	0.510985	3.6283	1.43E-04
PASS_ChildOnsetAsthma_Ferreira2019	1	0.03565	0.155341	0.032753	4.35741	0.918752	0.000369	1.112082	0.311764	3.56707	1.80E-04
PASS_CigarettesPerDay_Liu2019	0	0.03565	0.034532	0.016551	0.968634	0.464267	0.946	-0.01355	0.156097	-0.0868	5.35E-01
PASS_Coronary_Artery_Disease	0	0.03565	0.022711	0.025215	0.637071	0.707305	0.6022	-0.12276	0.235654	-0.52093	6.99E-01
PASS_Coronary_Artery_Disease_Howson_2017	0	0.03565	0.02715	0.025757	0.761575	0.722513	0.741171	-0.08115	0.245127	-0.33107	6.30E-01
PASS_Crohns_Disease	1	0.03565	0.168431	0.035976	4.724591	1.009163	9.77E-05	1.23135	0.315456	3.903402	4.74E-05
PASS_DepressedAffect_Nagel2018	0	0.03565	0.036567	0.005984	1.025727	0.167851	0.878181	0.007354	0.056418	0.130348	4.48E-01
PASS_Depression_Nagel2018	0	0.03565	0.04039	0.009843	1.132968	0.276115	0.629493	0.043304	0.092816	0.466556	3.20E-01
PASS_DrinksPerWeek_Liu2019	0	0.03565	0.037772	0.010078	1.059515	0.282705	0.833289	0.014556	0.095297	0.152748	4.39E-01
PASS_DS	0	0.03565	0.027688	0.019441	0.776679	0.545337	0.681151	-0.07473	0.183088	-0.40815	6.58E-01
PASS_ENIGMA2_MeanPutamen	0	0.03565	0.012479	0.027527	0.35005	0.77215	0.392434	-0.22704	0.255356	-0.88911	8.13E-01
PASS_Epilepsy_Anney_2014	0	0.03565	0.067149	0.028611	1.88356	0.802565	0.249514	0.295322	0.257307	1.147742	1.26E-01
PASS_Ever_Smoked	0	0.03565	0.061577	0.019903	1.727261	0.558279	0.200725	0.243599	0.190699	1.277403	1.01E-01
PASS_FastingGlucose_Manning	0	0.03565	0.001926	0.025589	0.05404	0.717798	0.180596	-0.33217	0.237004	-1.40155	9.19E-01
PASS_FetalBirthWeight_Warrington2019	0	0.03565	0.016968	0.010052	0.475955	0.28197	0.063978	-0.18549	0.095073	-1.95098	9.74E-01
PASS_GeneralRiskTolerance_KarlssonLinner2019	0	0.03565	0.03919	0.007783	1.099303	0.218319	0.649514	0.030976	0.073416	0.421922	3.37E-01
PASS_HbA1C	0	0.03565	0.08439	0.035184	2.367205	0.986923	0.143679	0.45506	0.31445	1.447163	7.39E-02
PASS_HDL	0	0.03565	0.01633	0.015448	0.458065	0.433333	0.211259	-0.19507	0.144717	-1.34791	9.11E-01
PASS_Height1	0	0.03565	0.018198	0.007431	0.510474	0.208442	0.021862	-0.1745	0.071092	-2.45456	9.93E-01
PASS_IBD	1	0.03565	0.146588	0.025686	4.111875	0.720512	2.34E-05	1.026962	0.241646	4.249858	1.07E-05
PASS_IBD_deLange2017	1	0.03565	0.154501	0.024402	4.333845	0.684482	1.79E-06	1.101221	0.227971	4.830536	6.81E-07
PASS_Insomnia_Jansen2019	0	0.03565	0.044712	0.009444	1.254196	0.264901	0.334148	0.082374	0.08832	0.932677	1.75E-01
PASS_Intelligence_SavageJansen2018	0	0.03565	0.033268	0.005878	0.933198	0.164895	0.685094	-0.02398	0.055125	-0.43503	6.68E-01
PASS_IschemicStroke_Malik2018	0	0.03565	0.006185	0.016853	0.173486	0.472749	0.072905	-0.28123	0.154231	-1.82342	9.66E-01
PASS_LargeArteryStroke_Malik2018	0	0.03565	0.045669	0.037981	1.281042	1.065381	0.791147	0.086866	-0.35787	-0.24273	5.96E-01
PASS_LDL	0	0.03565	0.011158	0.019824	0.313001	0.556069	0.214745	-0.25196	0.186349	-1.35206	9.12E-01
PASS_Lean_Body_Mass_Zillikens_2017	0	0.03565	0.052291	0.026386	1.466803	0.740145	0.528628	0.15117	0.248834	0.607513	2.72E-01
PASS_LongSleepDuration_Dashti2019	0	0.03565	0.042975	0.012402	1.205489	0.347872	0.554785	0.064565	0.116892	0.552346	2.90E-01
PASS_Lupus	1	0.03565	0.10799	0.03863	3.029193	1.083593	0.037617	0.658731	0.326128	2.019856	2.17E-02
PASS_MaternalBirthWeight_Warrington2019	0	0.03565	0.016757	0.010794	0.470053	0.302776	0.082904	-0.1864	0.102074	-1.82616	9.66E-01
PASS_MDD_Howard2019	0	0.03565	0.037152	0.006138	1.042147	0.172177	0.806587	0.013566	0.05793	0.234173	4.07E-01
PASS_MDD_Wray2018	0	0.03565	0.055543	0.011326	1.558015	0.317711	0.076115	0.185228	0.105477	1.756101	3.95E-02
PASS_MedicationUse_Wu2019	0	0.03565	0.041816	0.009088	1.172966	0.254923	0.499492	0.055039	0.085796	0.641507	2.61E-01
PASS_Menarche2017	0	0.03565	0.038573	0.007962	1.082009	0.223335	0.713348	0.021838	0.075185	0.290454	3.86E-01
PASS_Multiple_sclerosis	1	0.03565	0.823202	0.426079	23.09133	11.9518	2.71E-08	7.376602	1.285021	5.740453	4.72E-09
PASS_Neuroticism	0	0.03565	0.035793	0.009144	1.004023	0.256497	0.987312	0.002031	0.084964	0.023906	4.90E-01
PASS_Neuroticism_Nagel2018	0	0.03565	0.030688	0.005126	0.860826	0.143778	0.335986	-0.04793	0.048685	-0.98456	8.38E-01
PASS_NumberChildrenEverBorn	0	0.03565	-0.00386	0.014755	-0.10829	0.413879	0.005313	-0.37301	0.132522	-2.81472	9.98E-01
PASS_OvarianCancer	0	0.03565	0.103238	0.044443	2.895891	1.246643	0.136995	0.624764	0.426655	1.464331	7.16E-02
PASS_Parkinsons_23andMe_Corces2020	0	0.03565	0.086263	0.044201	2.419731	1.239854	0.255095	0.442133	0.418293	1.056994	1.45E-01
PASS_Primary_biliary_cirrhosis	1	0.03565	0.186973	0.040308	5.24472	1.130663	7.18E-06	1.398125	0.309726	4.514073	3.18E-06
PASS_ProstateCancer	0	0.03565	0.037891	0.019152	1.062868	0.537222	0.90664	0.008305	0.18038	0.046042	4.82E-01
PASS_ReactionTime_Davies2018	0	0.03565	0.03969	0.006876	1.113343	0.192868	0.556874	0.034174	0.064881	0.526719	2.99E-01
PASS_Rheumatoid_Arthritis	1	0.03565	0.143014	0.033005	4.011642	0.92582	0.000468	0.993677	0.283565	3.504227	2.29E-04
PASS_Schizophrenia	0	0.03565	0.037559	0.007013	1.053558	0.196728	0.785383	0.01519	0.066131	0.2297	4.09E-01
PASS_Schizophrenia_Pardinas2018	0	0.03565	0.030852	0.006515	0.865416	0.182754	0.461936	-0.05082	0.061569	-0.82536	7.95E-01

PASS_Schizophrenia_Ruderfer2018	0	0.03565	0.031963	0.007446	0.896584	0.208878	0.620656	-0.03735	0.070426	-0.53039	7.02E-01
PASS_SCZvsBD_Ruderfer2018	0	0.03565	0.032691	0.01345	0.91699	0.377285	0.826137	-0.03289	0.127164	-0.25865	6.02E-01
PASS_ShortSleepDuration_Dashti2019	0	0.03565	0.049165	0.007746	1.379113	0.217272	0.07919	0.127292	0.072404	1.758096	3.94E-02
PASS_SleepDuration_Dashti2019	0	0.03565	0.038825	0.005931	1.089056	0.166367	0.591743	0.027027	0.055932	0.483206	3.14E-01
PASS_SmallVesselStroke_Malik2018	0	0.03565	0.056809	0.092204	1.593515	2.586375	0.812202	0.239119	-0.83878	-0.28508	6.12E-01
PASS_SmokingCessation_Liu2019	0	0.03565	0.029186	0.014951	0.818692	0.419396	0.665631	-0.06623	0.141282	-0.46879	6.80E-01
PASS_SmokingInitiation_Liu2019	0	0.03565	0.031116	0.006484	0.872832	0.181867	0.483255	-0.04662	0.061155	-0.76225	7.77E-01
PASS_Stroke_Malik2018	0	0.03565	0.056809	0.092204	1.593515	2.586375	0.812202	0.239119	-0.83878	-0.28508	6.12E-01
PASS_SWB	0	0.03565	0.03008	0.023785	0.843775	0.667197	0.81275	-0.06052	0.221177	-0.27364	6.08E-01
PASS_Triglycerides	0	0.03565	0.016628	0.013568	0.466428	0.380604	0.166836	-0.19317	0.130097	-1.48484	9.31E-01
PASS_Type_1_Diabetes	1	0.03565	0.390524	0.15704	10.95445	4.405074	6.17E-06	3.305423	0.720676	4.586559	2.25E-06
PASS_Type_2_Diabetes	0	0.03565	0.039149	0.020809	1.098147	0.583703	0.866391	0.026205	0.196432	0.133407	4.47E-01
PASS_UC_deLange2017	1	0.03565	0.134046	0.020979	3.760073	0.588471	8.47E-06	0.909259	0.20257	4.488621	3.58E-06
PASS_Ulcerative_Colitis	1	0.03565	0.166865	0.031998	4.68068	0.897559	2.37E-05	1.216765	0.285301	4.26484	1.00E-05
PASS_VerbalNumericReasoning_Davies2018	0	0.03565	0.033701	0.005863	0.945326	0.164454	0.739143	-0.02009	0.054965	-0.36546	6.43E-01
PASS_Worry_Nagel2018	0	0.03565	0.026865	0.005264	0.753588	0.14766	0.096898	-0.08497	0.049774	-1.70715	9.56E-01
PASS_Years_of_Education1	0	0.03565	0.055198	0.014364	1.548348	0.402915	0.165669	0.182906	0.13271	1.378243	8.41E-02
PASS_Years_of_Education2	0	0.03565	0.042549	0.007028	1.193521	0.197154	0.327341	0.063223	0.066493	0.95082	1.71E-01
UKB_460K.biochemistry_AlanineAminotransferase	0	0.03565	0.008483	0.008342	0.237942	0.233992	0.001665	-0.26926	0.080738	-3.33504	1.00E+00
UKB_460K.biochemistry_Albumin	0	0.03565	0.033718	0.00883	0.945823	0.247695	0.826638	-0.03103	0.082725	-0.37512	6.46E-01
UKB_460K.biochemistry_AlkalinePhosphatase	0	0.03565	0.020413	0.010551	0.572607	0.295973	0.144314	-0.1604	0.09802	-1.63639	9.49E-01
UKB_460K.biochemistry_ApolipoproteinA	0	0.03565	0.037483	0.013618	1.05142	0.381995	0.891741	0.006051	0.126176	0.047955	4.81E-01
UKB_460K.biochemistry_ApolipoproteinB	0	0.03565	0.002859	0.016621	0.080192	0.466238	0.06214	-0.3303	0.169018	-1.95422	9.75E-01
UKB_460K.biochemistry_AspartateAminotransferase	0	0.03565	0.023487	0.010161	0.658814	0.285028	0.232269	-0.13073	0.096365	-1.35657	9.13E-01
UKB_460K.biochemistry_Calcium	0	0.03565	0.029368	0.011372	0.823805	0.318985	0.57604	-0.07438	0.106091	-0.70108	7.58E-01
UKB_460K.biochemistry_Cholesterol	0	0.03565	0.015685	0.015061	0.439977	0.422483	0.190209	-0.20313	0.144044	-1.41021	9.21E-01
UKB_460K.biochemistry_CreactiveProtein	1	0.03565	0.043732	0.031451	1.226723	0.882234	0.79581	0.064329	0.294783	0.218224	4.14E-01
UKB_460K.biochemistry_Creatinine	0	0.03565	0.022547	0.007286	0.632445	0.204366	0.072466	-0.13497	0.068828	-1.96098	9.75E-01
UKB_460K.biochemistry_CystatinC	0	0.03565	0.015716	0.010922	0.440849	0.306362	0.075268	-0.19967	0.105194	-1.89809	9.71E-01
UKB_460K.biochemistry_DirectBilirubin	0	0.03565	0.072319	0.031626	2.028608	0.887127	0.228603	0.332254	0.286465	1.15984	1.23E-01
UKB_460K.biochemistry_GammaGlutamyltransferase	0	0.03565	0.020103	0.008424	0.563897	0.236303	0.068157	-0.15482	0.080088	-1.93312	9.73E-01
UKB_460K.biochemistry_Glucose	0	0.03565	0.024117	0.010964	0.676504	0.307537	0.289012	-0.11321	0.102451	-1.10498	8.65E-01
UKB_460K.biochemistry_HbA1c	0	0.03565	0.057914	0.015975	1.624526	0.448121	0.163908	0.200569	0.151285	1.325772	9.25E-02
UKB_460K.biochemistry_HDLcholesterol	0	0.03565	0.040703	0.014162	1.14174	0.39725	0.719416	0.03778	0.131989	0.286238	3.87E-01
UKB_460K.biochemistry_IGF1	0	0.03565	0.03602	0.014778	1.010376	0.414546	0.979857	-0.00171	0.138687	-0.0123	5.05E-01
UKB_460K.biochemistry_LDLdirect	0	0.03565	0.003562	0.016123	0.099906	0.452249	0.054343	-0.32239	0.158501	-2.03401	9.79E-01
UKB_460K.biochemistry_LipoproteinA	0	0.03565	-0.00456	0.07407	-0.12779	2.077717	0.593628	-0.40392	0.719833	-0.56113	7.13E-01
UKB_460K.biochemistry_Phosphate	0	0.03565	0.018212	0.010734	0.510853	0.301093	0.117807	-0.17429	0.105431	-1.65308	9.51E-01
UKB_460K.biochemistry_SHBG	0	0.03565	0.031528	0.010629	0.884369	0.298155	0.697291	-0.04829	0.099987	-0.483	6.85E-01
UKB_460K.biochemistry_Testosterone_Male	0	0.03565	0.042588	0.010898	1.194621	0.305687	0.523656	0.057179	0.102603	0.557281	2.89E-01
UKB_460K.biochemistry_TotalBilirubin	0	0.03565	0.075509	0.033943	2.118063	0.95212	0.229451	0.365134	0.311823	1.170966	1.21E-01
UKB_460K.biochemistry_TotalProtein	0	0.03565	0.087866	0.014752	2.464704	0.413813	0.000516	0.476059	0.139393	3.415222	3.19E-04
UKB_460K.biochemistry_Triglycerides	0	0.03565	0.014175	0.010948	0.397605	0.307112	0.042005	-0.21064	0.098949	-2.12875	9.83E-01
UKB_460K.biochemistry_Urate	0	0.03565	0.009315	0.007011	0.261304	0.196673	0.000436	-0.25968	0.069968	-3.71148	1.00E+00
UKB_460K.biochemistry_Urea	0	0.03565	0.046253	0.017132	1.297416	0.480574	0.535152	0.093191	0.161298	0.577754	2.82E-01
UKB_460K.biochemistry_VitaminD	0	0.03565	0.025892	0.020094	0.726292	0.563638	0.617428	-0.10221	0.183259	-0.55775	7.11E-01
UKB_460K.blood_EOSINOPHIL_COUNT	1	0.03565	0.166246	0.020475	4.663316	0.574344	1.04E-08	1.212278	0.206186	5.879525	2.06E-09
UKB_460K.blood_HIGH_LIGHT_SCATTER_RETICULOCYTE_COUNT	0	0.03565	0.045819	0.012474	1.28526	0.349899	0.408379	0.085363	0.116046	0.735602	2.31E-01
UKB_460K.blood_LYMPHOCYTE_COUNT	1	0.03565	0.157567	0.020124	4.419868	0.56448	3.64E-08	1.134833	0.201405	5.634576	8.77E-09
UKB_460K.blood_MEAN_CORPUSCULAR_HEMOGLOBIN	0	0.03565	0.08491	0.019951	2.381792	0.559647	0.014647	0.443346	0.188583	2.350935	9.36E-03
UKB_460K.blood_MEAN_PLATELET_VOL	0	0.03565	0.079597	0.02044	2.232742	0.573363	0.032461	0.396541	0.192432	2.060679	1.97E-02
UKB_460K.blood_MEAN_SPHERED_CELL_VOL	0	0.03565	0.098344	0.020547	2.758603	0.576344	0.002989	0.572485	0.196926	2.907115	1.82E-03
UKB_460K.blood_MONOCYTE_COUNT	1	0.03565	0.101758	0.025561	2.854377	0.716993	0.012079	0.601616	0.246413	2.441497	7.31E-03
UKB_460K.blood_PLATELET_COUNT	0	0.03565	0.068915	0.012803	1.93311	0.359145	0.010444	0.294578	0.121036	2.433801	7.47E-03
UKB_460K.blood_PLATELET_DISTRIB_WIDTH	0	0.03565	0.076668	0.019891	2.150596	0.557947	0.036929	0.370995	0.184344	2.012516	2.21E-02
UKB_460K.blood_RBC_DISTRIB_WIDTH	0	0.03565	0.04124	0.02007	1.156807	0.562988	0.779411	0.037669	0.189017	0.199287	4.21E-01
UKB_460K.blood_RED_COUNT	0	0.03565	0.056616	0.014209	1.588112	0.398563	0.145131	0.180997	0.135094	1.339784	9.02E-02
UKB_460K.blood_WHITE_COUNT	1	0.03565	0.069654	0.011492	1.953834	0.322368	0.003114	0.304112	0.107616	2.825905	2.36E-03
UKB_460K.bmd_HEEL_TSCOREz	0	0.03565	0.017258	0.008351	0.484091	0.234239	0.027512	-0.18378	0.078379	-2.34477	9.90E-01
UKB_460K.body_BALDING1	0	0.03565	0.026651	0.01394	0.747585	0.391023	0.515555	-0.09558	0.129755	-0.73663	7.69E-01
UKB_460K.body_BALDING4	0	0.03565	0.021904	0.015627	0.614413	0.438359	0.370921	-0.1378	0.14399	-0.95704	8.31E-01
UKB_460K.body_BMIz	0	0.03565	0.03716	0.004039	1.042348	0.113292	0.708751	0.009376	0.03815	0.245774	4.03E-01
UKB_460K.body_HEIGHTz	0	0.03565	0.016157	0.00555	0.453226	0.155689	0.000536	-0.19472	0.051946	-3.74856	1.00E+00
UKB_460K.body_LEFT_HANDED	0	0.03565	0.053161	0.031438	1.491189	0.881859	0.571235	0.168789	0.293856	0.574394	2.83E-01
UKB_460K.body_WHRadjBMIz	0	0.03565	0.008368	0.005828	0.234721	0.163483	5.25E-06	-0.26534	0.054922	-4.83116	1.00E+00
UKB_460K.bp_DIASTOLICadjMEDz	0	0.03565	0.026041	0.006888	0.730478	0.193209	0.162692	-0.09892	0.064888	-1.52445	9.36E-01
UKB_460K.bp_SYSTOLICadjMEDz	0	0.03565	0.020806	0.005842	0.583626	0.163866	0.01162	-0.14794	0.055102	-2.68487	9.96E-01
UKB_460K.cancer_ALL	0	0.03565	0.055836	0.022236	1.566244	0.623723	0.361589	0.176623	0.208153	0.848522	1.98E-01
UKB_460K.cancer_BREAST	0	0.03565	0.008195	0.015563	0.229866	0.436539	0.085348	-0.2704	0.149671	-1.80663	9.65E-01
UKB_460K.cancer_MELANOMA	0	0.03565	0.039377	0.097671	1.10456	2.739729	0.957384	0.022164	0.652117	0.033988	4.86E-01
UKB_460K.cancer_PROSTATE	0	0.03565	0.056113	0.030439	1.573993	0.853838	0.49444	0.184227	0.280751	0.656195	2.56E-01
UKB_460K.cov_EDU_COLLEGE	0	0.03565	0.034609	0.004772	0.970811	0.133854	0.827481	-0.01237	0.045026	-0.27476	6.08E-01
UKB_460K.cov_EDU_YEARS	0	0.03565	0.037966	0.004965	1.064976	0.139263	0.641252	0.019803	0.046883	0.422388	3.36E-01
UKB_460K.cov_SMOKING_STATUS	0	0.03565	0.029315	0.004678	0.822311	0.131228	0.180179	-0.0619	0.044608	-1.38759	9.17E-01
UKB_460K.disease_AID_ALL	1	0.03565	0.187756	0.035423	5.266688	0.993628	4.10E-05	1.42124	0.341937	4.156434	1.62E-05

UKB_460K.disease_AID_SURE	1	0.03565	0.124703	0.032867	3.497987	0.921939	0.002336	0.826882	0.27301	3.028761	1.23E-03
UKB_460K.disease_ALLERGY_ECZEMA_DIAGNOSED	1	0.03565	0.146626	0.019756	4.112962	0.554156	4.27E-06	1.03282	0.221209	4.668972	1.51E-06
UKB_460K.disease_ASTHMA_DIAGNOSED	1	0.03565	0.131847	0.020634	3.6984	0.578794	2.06E-05	0.892126	0.207316	4.303208	8.42E-06
UKB_460K.disease_CARDIOVASCULAR	0	0.03565	0.033635	0.008108	0.943479	0.227429	0.803209	-0.02816	0.076399	-0.36854	6.44E-01
UKB_460K.disease_DERMATOLOGY	0	0.03565	0.155751	0.034246	4.368916	0.96063	0.000159	1.118228	0.295202	3.788017	7.59E-05
UKB_460K.disease_DIABETES_ANY_DIAGNOSED	0	0.03565	0.025299	0.0086	0.709647	0.241248	0.225958	-0.10285	0.080724	-1.27414	8.99E-01
UKB_460K.disease_ENDOCRINE_DIABETES	0	0.03565	0.027423	0.009109	0.769227	0.255509	0.365615	-0.08401	0.085955	-0.97739	8.36E-01
UKB_460K.disease_HI_CHOL_SELF_REP	0	0.03565	0.024392	0.014275	0.684211	0.400421	0.428055	-0.11784	0.134379	-0.87694	8.10E-01
UKB_460K.disease_HYPERTENSION_DIAGNOSED	0	0.03565	0.033183	0.008627	0.930808	0.241991	0.774419	-0.03195	0.081156	-0.3937	6.53E-01
UKB_460K.disease_HYPOTHYROIDISM_SELF_REP	1	0.03565	0.221424	0.041861	6.211095	1.174235	5.30E-05	1.734029	0.423537	4.094165	2.12E-05
UKB_460K.disease_PSORIASIS	1	0.03565	0.14764	0.052607	4.141383	1.475652	0.008274	1.031949	0.397632	2.59524	4.73E-03
UKB_460K.disease_RESPIRATORY_ENT	0	0.03565	0.12692	0.02006	3.560174	0.562698	3.08E-05	0.846744	0.201098	4.210598	1.27E-05
UKB_460K.disease_T2D	0	0.03565	0.031038	0.008414	0.870649	0.236024	0.582608	-0.04917	0.079037	-0.6221	7.33E-01
UKB_460K.disease_THYROID_ANY_SELF_REP	1	0.03565	0.230359	0.044855	6.461706	1.2582	9.68E-05	1.817128	0.460868	3.942836	4.03E-05
UKB_460K.impedance_BASAL_METABOLIC_RATEz	0	0.03565	0.028618	0.004713	0.802765	0.132196	0.135644	-0.07375	0.044115	-1.67167	9.53E-01
UKB_460K.lung_FEV1FVCzSMOKE	0	0.03565	0.022427	0.006066	0.629083	0.170153	0.029454	-0.1313	0.056879	-2.30845	9.90E-01
UKB_460K.lung_FVCzSMOKE	0	0.03565	0.028102	0.00551	0.788266	0.154565	0.171023	-0.07782	0.051828	-1.50152	9.33E-01
UKB_460K.mental_NEUROTICISM	0	0.03565	0.028737	0.005125	0.806093	0.143764	0.178289	-0.0674	0.048364	-1.39362	9.18E-01
UKB_460K.other_MORNINGPERSON	0	0.03565	0.034691	0.005477	0.973094	0.153645	0.861019	-0.01247	0.05167	-0.24135	5.95E-01
UKB_460K.pigment_HAIR	0	0.03565	0.058157	0.031017	1.631333	0.870037	0.470716	0.193308	0.293087	0.659558	2.55E-01
UKB_460K.pigment_HAIR_blackmale	0	0.03565	0.063458	0.047749	1.780029	1.339397	0.552457	0.236312	0.440187	0.536844	2.96E-01
UKB_460K.pigment_HAIR_blonde	0	0.03565	0.049658	0.031867	1.392941	0.893891	0.659885	0.110261	0.299196	0.368525	3.56E-01
UKB_460K.pigment_HAIR_darkbrown	0	0.03565	0.068705	0.048269	1.927216	1.353979	0.493367	0.289312	0.453619	0.637787	2.62E-01
UKB_460K.pigment_SKIN	0	0.03565	0.057563	0.053207	1.614692	1.492483	0.674763	0.183433	0.491618	0.373121	3.55E-01
UKB_460K.pigment_SUNBURN	0	0.03565	0.077398	0.044092	2.171064	1.236804	0.343726	0.381449	0.413656	0.922141	1.78E-01
UKB_460K.pigment_TANNING	0	0.03565	0.089191	0.06416	2.501859	1.799741	0.404195	0.48341	0.602159	0.802795	2.11E-01
UKB_460K.repro_AgeFirstBirth_Female	0	0.03565	0.029164	0.007041	0.81808	0.197497	0.359751	-0.06363	0.066768	-0.95305	8.30E-01
UKB_460K.repro_MENARCHE_AGE	0	0.03565	0.034436	0.006716	0.965962	0.188383	0.856782	-0.01621	0.063537	-0.25518	6.01E-01
UKB_460K.repro_MENOPAUSE_AGE	0	0.03565	0.040961	0.016607	1.148994	0.465849	0.747607	0.029442	0.156888	0.187662	4.26E-01
UKB_460K.repro_NumberChildrenEverBorn_Female	0	0.03565	0.040783	0.010227	1.143988	0.286876	0.615182	0.046086	0.096437	0.47789	3.16E-01
UKB_460K.repro_NumberChildrenEverBorn_Male	0	0.03565	0.014311	0.01545	0.401425	0.433384	0.160414	-0.20169	0.143321	-1.40725	9.20E-01
UKB_460K.repro_NumberChildrenEverBorn_Pooled	0	0.03565	0.029812	0.008575	0.836244	0.240542	0.494737	-0.05632	0.080746	-0.69748	7.57E-01

Table B.5: S-LDXR results.

	All traits								blood/immune traits						
ANNOT	PROP SNPS	HSQ1 ENR	HSQ1 ENR SE	HSQ2 ENR	HSQ2 ENR SE	GCORSQ ENR	GCORSQ ENR SE	GCORSQ ENR P	HSQ1 ENR	HSQ1 ENRSE	HSQ2 ENR	HSQ2 ENR SE	GCORSQ ENR	GCORSQ ENR SE	GCORSQ ENR P
UTR_5_UCSC	0.01	4.21	1.34	4.79	0.71	0.81	0.05	0.0001	5.87	1.97	6.09	1.22	0.63	0.08	2E-06
EpiNN	0.04	1.6	0.16	1.52	0.11	0.73	0.04	7.4E-10	2.32	0.22	2.11	0.16	0.67	0.05	3.9E-10
UTR_3_UCSC	0.01	4.5	0.7	3.59	0.44	0.91	0.04	0.0167	3.92	1.09	3.23	0.65	0.78	0.08	0.00455
Coding_UCSC	0.01	6.4	0.79	6.17	0.45	0.88	0.04	0.00122	6.88	1.33	7.69	0.79	0.8	0.05	5.9E-05
CTCF_Hoffman.extend.500	0.07	1.13	0.22	0.74	0.16	0.86	0.03	5.1E-06	0.96	0.41	0.6	0.3	0.8	0.05	5.4E-05
Promoter_UCSC.extend.500	0.04	3.68	0.41	3.52	0.29	0.87	0.03	0.00021	5.24	0.48	4.83	0.28	0.81	0.04	1.2E-05
Enhancer_Andersson	0	8.34	2.29	5.27	1.66	0.83	0.07	0.00971	16.89	3.62	15.58	2.11	0.82	0.09	0.0451
Promoter_UCSC	0.03	3.41	0.58	4.32	0.37	0.87	0.05	0.00662	4.03	0.71	5.51	0.41	0.85	0.06	0.0128
SuperEnhancer_Hnisz.extend.500	0.17	2.9	0.17	2.67	0.15	0.87	0.02	8.1E-09	3.41	0.1	3.13	0.08	0.86	0.02	3.5E-09
SuperEnhancer_Hnisz	0.17	2.84	0.18	2.67	0.15	0.88	0.02	1.2E-10	3.31	0.11	3.16	0.07	0.86	0.02	5.2E-10
Enhancer_Hoffman.extend.500	0.09	4.42	0.36	4.18	0.35	0.9	0.02	1.9E-08	5.73	0.33	5.5	0.19	0.88	0.02	5.4E-07
CTCF_Hoffman	0.02	-0.06	0.71	0.23	0.41	0.89	0.04	0.0105	-1.91	1.1	0.13	0.68	0.89	0.07	0.116
Enhancer_Andersson.extend.500	0.02	3.78	0.86	3.45	0.62	0.93	0.03	0.057	7.94	1.48	7.04	0.75	0.89	0.04	0.00685
TSS_Hoffman.extend.500	0.03	7.06	0.66	7.63	0.63	0.89	0.02	2.9E-06	8.8	0.63	9.48	0.43	0.89	0.03	0.00037
Coding_UCSC.extend.500	0.06	3.57	0.36	3.24	0.28	0.91	0.03	0.00158	4.4	0.34	3.75	0.21	0.89	0.04	0.00311
TSS_Hoffman	0.02	9	1.13	8.93	0.95	0.94	0.03	0.0263	13.11	1.42	13.93	0.82	0.92	0.03	0.0134
Conserved_LindbladToh	0.02	10.12	1	9.31	0.77	0.91	0.02	9.3E-06	7.07	1.23	8.96	1.18	0.92	0.04	0.0452
Intron_UCSC.extend.500	0.39	1.43	0.03	1.38	0.02	0.93	0.01	1.3E-07	1.48	0.04	1.46	0.02	0.93	0.02	0.00024
UTR_5_UCSC.extend.500	0.03	3.08	0.4	3.63	0.35	0.93	0.05	0.124	4.33	0.64	5.08	0.4	0.93	0.07	0.315
DGF_ENCODE.extend.500	0.54	1.53	0.09	1.56	0.07	0.98	0.02	0.248	1.64	0.1	1.82	0.04	0.93	0.02	0.0011
H3K4me1_Trynka.extend.500	0.6	1.57	0.05	1.47	0.04	0.95	0.01	2.4E-05	1.57	0.05	1.51	0.03	0.93	0.02	0.00023
H3K27ac_Hnisz	0.39	1.99	0.1	2.06	0.1	0.95	0.01	9.3E-06	2.25	0.07	2.39	0.04	0.94	0.02	0.00058
UTR_3_UCSC.extend.500	0.03	3.32	0.43	2.58	0.2	0.94	0.04	0.171	3.96	0.57	3.44	0.34	0.94	0.06	0.296
H3K9ac_Trynka.extend.500	0.23	2.72	0.17	2.65	0.15	0.93	0.02	3E-05	2.75	0.17	2.96	0.13	0.94	0.02	0.0121
H3K9ac_peaks_Trynka	0.04	5.23	0.61	5.31	0.38	0.96	0.02	0.0558	5.86	1.01	6.31	0.7	0.95	0.03	0.0545
H3K27ac_Hnisz.extend.500	0.42	1.98	0.09	1.91	0.08	0.94	0.01	2.8E-06	2.22	0.06	2.14	0.03	0.95	0.02	0.00673
H3K4me3_Trynka	0.13	3.68	0.25	3.9	0.27	0.97	0.02	0.0633	4.42	0.32	4.9	0.2	0.95	0.02	0.025
H3K4me1_Trynka	0.42	2.22	0.12	2.15	0.1	0.96	0.01	0.0146	2.49	0.12	2.41	0.07	0.95	0.02	0.0103
H3K4me3_peaks_Trynka	0.04	4.3	0.69	4.66	0.44	0.98	0.02	0.222	4.96	1.22	5.36	0.8	0.95	0.03	0.0974
RSsupp4	0.07	2.82	0.22	2.64	0.15	0.93	0.03	0.0142	2.31	0.31	2.36	0.17	0.96	0.05	0.341
Transcribed_Hoffman.extend.500	0.76	0.95	0.03	0.93	0.02	0.98	0.01	0.196	0.9	0.04	0.87	0.02	0.96	0.02	0.108
H3K9ac_Trynka	0.12	4.53	0.36	4.05	0.3	0.93	0.02	0.00025	4.62	0.47	4.34	0.31	0.96	0.03	0.161
Intron_UCSC	0.39	1.29	0.03	1.22	0.02	0.96	0.01	0.00924	1.33	0.05	1.24	0.03	0.96	0.02	0.122
Enhancer_Hoffman	0.04	5.64	0.51	4.6	0.37	0.93	0.02	0.0038	7.4	0.78	5.73	0.48	0.96	0.03	0.287
H3K4me3_Trynka.extend.500	0.25	2.5	0.14	2.54	0.15	0.95	0.02	0.00464	2.75	0.15	2.89	0.08	0.97	0.03	0.168
H3K27ac_PGC2	0.27	2.73	0.17	2.66	0.17	0.97	0.02	0.074	3.11	0.16	3.14	0.11	0.97	0.02	0.168
DHS_Trynka.extend.500	0.5	1.42	0.08	1.56	0.06	0.99	0.02	0.593	1.46	0.12	1.68	0.07	0.97	0.02	0.184
PromoterFlanking_Hoffman	0.01	0.24	1.24	2.68	0.92	0.97	0.1	0.777	1.98	2.04	3.66	1.57	0.98	0.14	0.862
TFBS_ENCODE	0.13	4.32	0.47	3.96	0.32	0.98	0.02	0.233	5.07	0.57	4.95	0.33	0.98	0.02	0.196
TFBS_ENCODE.extend.500	0.34	2	0.11	1.86	0.09	0.98	0.01	0.22	2.06	0.13	2.06	0.08	0.98	0.02	0.362
WeakEnhancer_Hoffman	0.02	2.38	0.76	2.58	0.42	0.92	0.04	0.0191	2.84	1.3	3.31	0.81	0.98	0.04	0.687
DGF_ENCODE	0.13	3.08	0.38	3.1	0.29	0.99	0.01	0.311	3.65	0.51	4.06	0.37	0.98	0.02	0.417
FetalDHS_Trynka.extend.500	0.28	2.05	0.16	2.24	0.14	0.99	0.02	0.581	2.17	0.21	2.69	0.1	0.99	0.02	0.532
H3K27ac_PGC2.extend.500	0.33	2.34	0.11	2.28	0.11	0.98	0.02	0.161	2.59	0.1	2.61	0.05	0.99	0.02	0.556
Transcribed_Hoffman	0.34	1.31	0.08	1.17	0.04	0.98	0.02	0.35	1.24	0.12	1.16	0.07	0.99	0.03	0.727
H3K4me1_peaks_Trynka	0.17	3.48	0.3	3.02	0.14	1	0.02	0.77	3.35	0.4	3.3	0.21	1	0.03	0.975
PromoterFlanking_Hoffman.extend.500	0.03	2.12	0.39	2.88	0.31	0.96	0.05	0.372	2.12	0.67	3.79	0.41	1	0.05	0.93
WeakEnhancer_Hoffman.extend.500	0.09	2.47	0.26	2.19	0.19	0.94	0.03	0.0741	3.06	0.34	2.84	0.21	1.01	0.03	0.678
Repressed_Hoffman	0.46	0.45	0.08	0.46	0.06	1.01	0.02	0.618	0.39	0.1	0.29	0.05	1.02	0.04	0.636
DHS_peaks_Trynka	0.11	3.38	0.52	2.99	0.26	1.01	0.02	0.562	3.51	0.82	3.61	0.31	1.04	0.03	0.206
FetalDHS_Trynka	0.08	3.57	0.37	3.86	0.29	1.02	0.02	0.298	4.26	0.62	5.01	0.34	1.04	0.03	0.177
Conserved_LindbladToh.extend.500	0.33	1.83	0.08	1.81	0.06	1.01	0.01	0.321	1.77	0.14	1.88	0.08	1.04	0.02	0.121
DHS_Trynka	0.17	2.5	0.32	2.58	0.23	1.02	0.02	0.297	2.43	0.42	3.19	0.27	1.05	0.02	0.0478
Repressed_Hoffman.extend.500	0.72	0.37	0.05	0.46	0.05	1.09	0.02	5.1E-06	0.26	0.04	0.31	0.02	1.25	0.06	2E-05

Table B.6: *DNMT1* CAGEseq raw data in .bed file format.

Chromosome	Start bp	End bp	Mesurement	CAGEseq value	strand
chr19	10142848	10142849	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142852	10142853	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	2	-
chr19	10142868	10142869	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142888	10142889	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142891	10142892	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142892	10142893	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142893	10142894	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	5	-
chr19	10142897	10142898	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	6	-
chr19	10142899	10142900	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	4	-
chr19	10142900	10142901	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142903	10142904	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10142908	10142909	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor1	1	-
chr19	10143011	10143012	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	+
chr19	10143013	10143014	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	2	+
chr19	10142863	10142864	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142868	10142869	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142872	10142873	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	2	-
chr19	10142879	10142880	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142885	10142886	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142891	10142892	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142893	10142894	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	7	-
chr19	10142894	10142895	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	3	-
chr19	10142897	10142898	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	9	-
chr19	10142899	10142900	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	2	-
chr19	10142900	10142901	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142903	10142904	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142904	10142905	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	2	-
chr19	10142908	10142909	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10142910	10142911	CD4+CD25+CD45RA+_naive_regulatory_T_cells_donor2	1	-
chr19	10143013	10143014	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	+
chr19	10142852	10142853	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142868	10142869	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142870	10142871	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142872	10142873	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142893	10142894	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	7	-
chr19	10142894	10142895	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142897	10142898	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	2	-
chr19	10142898	10142899	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142899	10142900	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	4	-
chr19	10142900	10142901	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142901	10142902	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142902	10142903	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142908	10142909	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor1	1	-
chr19	10142868	10142869	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	+
chr19	10143013	10143014	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	+
chr19	10143015	10143016	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	+
chr19	10142844	10142845	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142858	10142859	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142868	10142869	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	2	-
chr19	10142872	10142873	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	2	-
chr19	10142874	10142875	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142876	10142877	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142887	10142888	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142890	10142891	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142891	10142892	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142892	10142893	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142893	10142894	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	12	-
chr19	10142896	10142897	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142897	10142898	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	10	-
chr19	10142899	10142900	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	12	-
chr19	10142900	10142901	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142901	10142902	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142902	10142903	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	2	-
chr19	10142903	10142904	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142905	10142906	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142907	10142908	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10142908	10142909	CD4+CD25-CD45RA-_memory_conventional_T_cells_donor2	1	-
chr19	10143001	10143002	CD8+_T_Cells_pluriselect_donor11	1	+
chr19	10143011	10143012	CD8+_T_Cells_pluriselect_donor11	1	+
chr19	10143013	10143014	CD8+_T_Cells_pluriselect_donor11	2	+
chr19	10142852	10142853	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142870	10142871	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142874	10142875	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142876	10142877	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142886	10142887	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142890	10142891	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor11	3	-
chr19	10142892	10142893	CD8+_T_Cells_pluriselect_donor11	2	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor11	5	-
chr19	10142894	10142895	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142895	10142896	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor11	15	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor11	2	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor11	8	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor11	3	-
chr19	10142904	10142905	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142905	10142906	CD8+_T_Cells_pluriselect_donor11	1	-
chr19	10142906	10142907	CD8+_T_Cells_pluriselect_donor11	2	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor11	5	-
chr19	10143011	10143012	CD8+_T_Cells_pluriselect_donor12	1	+
chr19	10142839	10142840	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142878	10142879	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor12	2	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor12	2	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor12	9	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor12	2	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142907	10142908	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142922	10142923	CD8+_T_Cells_pluriselect_donor12	1	-
chr19	10142852	10142853	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor13	4	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor13	4	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10142902	10142903	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor13	1	-
chr19	10143011	10143012	CD8+_T_Cells_pluriselect_donor21	1	+
chr19	10142851	10142852	CD8+_T_Cells_pluriselect_donor21	2	-
chr19	10142863	10142864	CD8+_T_Cells_pluriselect_donor21	1	-

chr19	10142867	10142868	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142874	10142875	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142879	10142880	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142885	10142886	CD8+_T_Cells_pluriselect_donor21	3	-
chr19	10142886	10142887	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142889	10142890	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142890	10142891	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor21	4	-
chr19	10142892	10142893	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor21	8	-
chr19	10142894	10142895	CD8+_T_Cells_pluriselect_donor21	2	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor21	7	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor21	3	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor21	5	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor21	5	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142901	10142902	CD8+_T_Cells_pluriselect_donor21	1	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor21	2	-
chr19	10143001	10143002	CD8+_T_Cells_pluriselect_donor22	1	+
chr19	10143016	10143017	CD8+_T_Cells_pluriselect_donor22	1	+
chr19	10142852	10142853	CD8+_T_Cells_pluriselect_donor22	2	-
chr19	10142855	10142856	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142864	10142865	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142867	10142868	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142872	10142873	CD8+_T_Cells_pluriselect_donor22	4	-
chr19	10142879	10142880	CD8+_T_Cells_pluriselect_donor22	2	-
chr19	10142883	10142884	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142885	10142886	CD8+_T_Cells_pluriselect_donor22	3	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142892	10142893	CD8+_T_Cells_pluriselect_donor22	3	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor22	9	-
chr19	10142894	10142895	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor22	10	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor22	9	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor22	3	-
chr19	10142901	10142902	CD8+_T_Cells_pluriselect_donor22	2	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor22	1	-
chr19	10142909	10142910	CD8+_T_Cells_pluriselect_donor22	2	-
chr19	10142852	10142853	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142871	10142872	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor31	5	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor31	2	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor31	2	-
chr19	10142906	10142907	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142907	10142908	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor31	1	-
chr19	10143010	10143011	CD8+_T_Cells_pluriselect_donor32	1	+
chr19	10143013	10143014	CD8+_T_Cells_pluriselect_donor32	2	+
chr19	10142866	10142867	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142870	10142871	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142871	10142872	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142872	10142873	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142884	10142885	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142890	10142891	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142892	10142893	CD8+_T_Cells_pluriselect_donor32	5	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor32	6	-
chr19	10142894	10142895	CD8+_T_Cells_pluriselect_donor32	3	-
chr19	10142895	10142896	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142897	10142898	CD8+_T_Cells_pluriselect_donor32	12	-
chr19	10142898	10142899	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor32	11	-
chr19	10142900	10142901	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142901	10142902	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142903	10142904	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142906	10142907	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142907	10142908	CD8+_T_Cells_pluriselect_donor32	2	-
chr19	10142908	10142909	CD8+_T_Cells_pluriselect_donor32	3	-
chr19	10142909	10142910	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142910	10142911	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142911	10142912	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142951	10142952	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142994	10142995	CD8+_T_Cells_pluriselect_donor32	1	-
chr19	10142858	10142859	CD8+_T_Cells_pluriselect_donor33	1	-
chr19	10142891	10142892	CD8+_T_Cells_pluriselect_donor33	1	-
chr19	10142892	10142893	CD8+_T_Cells_pluriselect_donor33	1	-
chr19	10142893	10142894	CD8+_T_Cells_pluriselect_donor33	2	-
chr19	10142894	10142895	CD8+_T_Cells_pluriselect_donor33	1	-
chr19	10142896	10142897	CD8+_T_Cells_pluriselect_donor33	2	-
chr19	10142899	10142900	CD8+_T_Cells_pluriselect_donor33	2	-
chr19	10142901	10142902	CD8+_T_Cells_pluriselect_donor33	1	-
chr19	10142903	10142904	CD8+_T_Cells_pluriselect_donor33	2	-
chr19	10142909	10142910	CD8+_T_Cells_pluriselect_donor33	1	-

References

[1] Jason Ernst and Manolis Kellis. 'ChromHMM: automating chromatin-state discovery and characterization'. In: *Nature methods* 9.3 (2012), p. 215.

[2] AR Abbas et al. 'Immune response in silico (IRIS): immune-specific genes identified from a compendium of microarray expression data'. In: *Genes and immunity* 6.4 (2005), p. 319.

[3] Cosmas C Giallourakis et al. 'Genome-wide analysis of immune system genes by expressed sequence Tag profiling'. In: *The Journal of Immunology* 190.11 (2013), pp. 5578–5587.

[4] Anshul Kundaje et al. 'Integrative analysis of 111 reference human epigenomes'. In: *Nature* 518.7539 (2015), p. 317.

[5] Yukinori Okada et al. 'Genetics of rheumatoid arthritis contributes to biology and drug discovery'. In: *Nature* 506.7488 (2014), pp. 376–381.

[6] Calliope A Dendrou et al. 'Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity'. In: *Science translational medicine* 8.363 (2016), 363ra149–363ra149.

[7] Wenjuan Chen et al. 'The tumor necrosis factor receptor superfamily member 1B polymorphisms predict response to anti-TNF therapy in patients with autoimmune disease: A meta-analysis'. In: *International immunopharmacology* 28.1 (2015), pp. 146–153.

[8] Lorea Villanueva, Damiana Álvarez-Errico and Manel Esteller. 'The Contribution of Epigenetics to Cancer Immunotherapy'. In: *Trends in Immunology* (2020).

[9] Yukihide Momozawa et al. 'IBD risk loci are enriched in multigenic regulatory modules encompassing putative causative genes'. In: *Nature communications* 9.1 (2018), pp. 1–18.

[10] Luke Jostins et al. 'Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease'. In: *Nature* 491.7422 (2012), pp. 119–124.

[11] Michael F Seldin. 'The genetics of human autoimmune disease: a perspective on progress in the field and future directions'. In: *Journal of autoimmunity* 64 (2015), pp. 1–12.

[12] Jimmy Z Liu et al. 'Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations'. In: *Nature genetics* 47.9 (2015), pp. 979–986.

[13] Mario Roederer et al. 'The genetic architecture of the human immune system: a bioresource for autoimmunity and disease pathogenesis'. In: *Cell* 161.2 (2015), pp. 387–403.

[14] Kyle Kai-How Farh et al. 'Genetic and epigenetic fine mapping of causal autoimmune disease variants'. In: *Nature* 518.7539 (2015), pp. 337–343.

[15] Maria Gutierrez-Arcelus, Stephen S Rich and Soumya Raychaudhuri. 'Autoimmune diseases—connecting risk alleles with molecular traits of the immune system'. In: *Nature Reviews Genetics* 17.3 (2016), p. 160.

[16] CJ Zopf et al. 'Cell-cycle dependence of transcription dominates noise in gene expression'. In: *PLoS computational biology* 9.7 (2013), e1003161.

[17] Jenny Bryan. 'Problems in gene clustering based on gene expression data'. In: *Journal of Multivariate Analysis* 90.1 (2004), pp. 44–66.

[18] Juan M Vaquerizas et al. 'A census of human transcription factors: function, expression and evolution'. In: *Nature Reviews Genetics* 10.4 (2009), pp. 252–263.

[19] Kondethimmanahalli Chandramouli and Pei-Yuan Qian. 'Proteomics: challenges, techniques and possibilities to overcome biological sample complexity'. In: *Human genomics and proteomics: HGP* 2009 (2009).

[20] M Ryan Corces et al. 'Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution'. In: *Nature genetics* 48.10 (2016), p. 1193.

[21] Colleen M Lau et al. 'Epigenetic control of innate and adaptive immune memory'. In: *Nature immunology* 19.9 (2018), pp. 963–972.

[22] Francesca Meda et al. 'The epigenetics of autoimmunity'. In: *Cellular & molecular immunology* 8.3 (2011), pp. 226–236.

[23] Anshul Kundaje et al. 'Integrative analysis of 111 reference human epigenomes'. In: *Nature* 518.7539 (2015), pp. 317–330.

[24] Paul D Thomas et al. 'PANTHER: a library of protein families and subfamilies indexed by function'. In: *Genome research* 13.9 (2003), pp. 2129–2141.

[25] Brendan K Bulik-Sullivan et al. 'LD Score regression distinguishes confounding from polygenicity in genome-wide association studies'. In: *Nature genetics* 47.3 (2015), p. 291.

[26] Hilary K Finucane et al. 'Partitioning heritability by functional annotation using genome-wide association summary statistics'. In: *Nature genetics* 47.11 (2015), p. 1228.

[27] Steven Gazal et al. 'Linkage disequilibrium–dependent architecture of human complex traits shows action of negative selection'. In: *Nature genetics* 49.10 (2017), p. 1421.

[28] Douglas Hanahan and Robert A Weinberg. 'Hallmarks of cancer: the next generation. CELL'. In: *Journal Article* (2020), pp. 646–74.

[29] Geoffrey C Gurtner et al. 'Wound repair and regeneration'. In: *Nature* 453.7193 (2008), pp. 314–321.

[30] Robert E Sorge et al. 'Different immune cells mediate mechanical pain hypersensitivity in male and female mice'. In: *Nature neuroscience* 18.8 (2015), pp. 1081–1083.

[31] Mel Storey and Sue Jordan. 'An overview of the immune system.' In: *Nursing Standard* 23 (2008).

[32] Michael F Tosi. 'Innate immune responses to infection'. In: *Journal of Allergy and Clinical Immunology* 116.2 (2005), pp. 241–249.

[33] Jacqueline Parkin and Bryony Cohen. 'An overview of the immune system'. In: *The Lancet* 357.9270 (2001), pp. 1777–1789.

[34] EP Cronkite et al. 'Regulation of granulopoiesis'. In: *Progress in clinical and biological research* 184 (1985), pp. 129–144.

[35] J Vidya Sarma and Peter A Ward. 'The complement system'. In: *Cell and tissue research* 343.1 (2011), pp. 227–235.

[36] Christiane S Hampe. 'B cells in autoimmune diseases'. In: *Scientifica* 2012 (2012).

[37] *WHO: The top 10 causes of death*. URL: `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death` (visited on 01/08/2022).

[38] Julie R Brahmer et al. 'Safety and activity of anti–PD-L1 antibody in patients with advanced cancer'. In: *New England Journal of Medicine* 366.26 (2012), pp. 2455–2465.

[39] Philip W Kantoff et al. 'Sipuleucel-T immunotherapy for castration-resistant prostate cancer'. In: *New England Journal of Medicine* 363.5 (2010), pp. 411–422.

[40] Haidong Dong et al. 'Tumor-associated B7-H1 promotes T-cell apoptosis: a potential mechanism of immune evasion'. In: *Nature medicine* 8.8 (2002), pp. 793–800.

[41] Steven A Rosenberg, James C Yang and Nicholas P Restifo. 'Cancer immunotherapy: moving beyond current vaccines'. In: *Nature medicine* 10.9 (2004), pp. 909–915.

[42] Mark E Dudley et al. 'Cancer regression and autoimmunity in patients after clonal repopulation with antitumor lymphocytes'. In: *Science* 298.5594 (2002), pp. 850–854.

[43] Omid Hamid et al. 'Safety and tumor responses with lambrolizumab (anti–PD-1) in melanoma'. In: *New England Journal of Medicine* 369.2 (2013), pp. 134–144.

[44] David L Porter et al. 'Chimeric antigen receptor–modified T cells in chronic lymphoid leukemia'. In: *N engl j Med* 365 (2011), pp. 725–733.

[45] Richard A Morgan et al. 'Cancer regression in patients after transfer of genetically engineered lymphocytes'. In: *Science* 314.5796 (2006), pp. 126–129.

[46] Christine McCusker and Richard Warrington. 'Primary immunodeficiency'. In: *Allergy, Asthma & Clinical Immunology* 7.1 (2011), pp. 1–8.

[47] *Severe Combined Immunodeficiency (SCID) | NIH: National Institute of Allergy and Infectious Diseases*. en. URL: `https://www.niaid.nih.gov/diseases-conditions/severe-combined-immunodeficiency-scid` (visited on 01/08/2022).

[48] Donald YM Leung. 'Molecular basis of allergic diseases'. In: *Molecular genetics and metabolism* 63.3 (1998), pp. 157–167.

[49] Yan Xing and Kristin A Hogquist. 'T-cell tolerance: central and peripheral'. In: *Cold Spring Harbor perspectives in biology* 4.6 (2012), a006957.

[50] Pietro Invernizzi and M Eric Gershwin. 'The genetics of human autoimmune disease'. In: *Journal of autoimmunity* 33.3-4 (2009), pp. 290–299.

[51] Sebastien Viatte and Anne Barton. 'Genetics of rheumatoid arthritis susceptibility, severity, and treatment response'. In: *Seminars in immunopathology*. Vol. 39. 4. Springer. 2017, pp. 395–408.

[52] Peter N Taylor et al. 'Whole-genome sequence-based analysis of thyroid function'. In: *Nature communications* 6.1 (2015), pp. 1–11.

[53] Yul Hwangbo and Young Joo Park. 'Genome-wide association studies of autoimmune thyroid diseases, thyroid function, and thyroid cancer'. In: *Endocrinology and Metabolism* 33.2 (2018), pp. 175–184.

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

[55] Dennis A Carson. 'Genetic factors in the etiology and pathogenesis of autoimmunity 1'. In: *The FASEB journal* 6.10 (1992), pp. 2800–2805.

[56] Celi Sun et al. 'High-density genotyping of immune-related loci identifies new SLE risk variants in individuals with Asian ancestry'. In: *Nature genetics* 48.3 (2016), pp. 323–330.

[57] Xinli Hu et al. 'Additive and interaction effects at three amino acid positions in HLA-DQ and HLA-DR molecules drive type 1 diabetes risk'. In: *Nature genetics* 47.8 (2015), pp. 898–905.

[58] Nunzio Bottini et al. 'A functional variant of lymphoid tyrosine phosphatase is associated with type I diabetes'. In: *Nature genetics* 36.4 (2004), pp. 337–338.

[59] Ann B Begovich et al. 'A missense single-nucleotide polymorphism in a gene encoding a protein tyrosine phosphatase (PTPN22) is associated with rheumatoid arthritis'. In: *The American Journal of Human Genetics* 75.2 (2004), pp. 330–337.

[60] Jochen Hampe et al. 'A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1'. In: *Nature genetics* 39.2 (2007), pp. 207–211.

[61] N Laurin et al. 'No preferential transmission of paternal alleles at risk genes in attention-deficit hyperactivity disorder'. In: *Molecular psychiatry* 12.3 (2007), pp. 226–229.

[62] Richard H Duerr. 'Genome-wide association studies herald a new era of rapid discoveries in inflammatory bowel disease research'. In: *Gastroenterology* 132.5 (2007), pp. 2045–2049.

[63] Michele Cargill et al. 'A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes'. In: *The American Journal of Human Genetics* 80.2 (2007), pp. 273–290.

[64] Elaine F Remmers et al. 'STAT4 and the risk of rheumatoid arthritis and systemic lupus erythematosus'. In: *New England Journal of Medicine* 357.10 (2007), pp. 977–986.

[65] Peter Brian Medawar and Jean S Medawar. *Aristotle to zoos: a philosophical dictionary of biology*. Harvard University Press, 1983.

[66] Yuqi Chen et al. 'Epigenetic modification of nucleic acids: from basic studies to medical applications'. In: *Chemical Society Reviews* 46.10 (2017), pp. 2844–2872.

[67] Forest H Andrews, Brian D Strahl and Tatiana G Kutateladze. 'Insights into newly discovered marks and readers of epigenetic information'. In: *Nature chemical biology* 12.9 (2016), pp. 662–668.

[68] Yaser Atlasi and Hendrik G Stunnenberg. 'The interplay of epigenetic marks during stem cell differentiation and development'. In: *Nature Reviews Genetics* 18.11 (2017), pp. 643–658.

[69] A Ganesan. 'Epigenetic drug discovery: a success story for cofactor interference'. In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1748 (2018), p. 20170069.

[70] Dirk Schübeler. 'Function and information content of DNA methylation'. In: *Nature* 517.7534 (2015), pp. 321–326.

[71] Swaminathan Venkatesh and Jerry L Workman. 'Histone exchange, chromatin structure and the regulation of transcription'. In: *Nature reviews Molecular cell biology* 16.3 (2015), pp. 178–189.

[72] Armen R Karapetyan et al. 'Regulatory roles for long ncRNA and mRNA'. In: *Cancers* 5.2 (2013), pp. 462–490.

[73] Alvaro Rada-Iglesias. 'Is H3K4me1 at enhancers correlative or causative?' In: *Nature genetics* 50.1 (2018), pp. 4–5.

[74] Gangning Liang et al. 'Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome'. In: *Proceedings of the National Academy of Sciences* 101.19 (2004), pp. 7357–7362.

[75] Tony Kouzarides. 'Chromatin modifications and their function'. In: *Cell* 128.4 (2007), pp. 693–705.

[76] Karin J Ferrari et al. 'Polycomb-dependent H3K27me1 and H3K27me2 regulate active transcription and enhancer fidelity'. In: *Molecular cell* 53.1 (2014), pp. 49–62.

[77] Schraga Schwartz, Eran Meshorer and Gil Ast. 'Chromatin organization marks exon-intron structure'. In: *Nature structural & molecular biology* 16.9 (2009), pp. 990–995.

[78] Menno P Creyghton et al. 'Histone H3K27ac separates active from poised enhancers and predicts developmental state'. In: *Proceedings of the National Academy of Sciences* 107.50 (2010), pp. 21931–21936.

[79] *NIH Roadmap Epigenomics Web Portal*. URL: `https://egg2.wustl.edu/roadmap/web_portal/` (visited on 05/08/2022).

[80] Xin Zhou et al. 'The human epigenome browser at Washington University'. In: *Nature methods* 8.12 (2011), pp. 989–990.

[81] Maninjay K Atianand, Daniel R Caffrey and Katherine A Fitzgerald. 'Immunobiology of long noncoding RNAs'. In: *Annual review of immunology* 35 (2017), p. 177.

[82] Vinod Kumar et al. 'Human disease-associated genetic variation impacts large intergenic non-coding RNA expression'. In: *PLoS genetics* 9.1 (2013), e1003201.

[83] Wenqian Hu, Juan R Alvarez-Dominguez and Harvey F Lodish. 'Regulation of mammalian cell differentiation by long non-coding RNAs'. In: *EMBO reports* 13.11 (2012), pp. 971–983.

[84] Adam Frankish et al. 'GENCODE 2021'. In: *Nucleic acids research* 49.D1 (2021), pp. D916–D923.

[85] Valeria Ranzani et al. 'The long intergenic noncoding RNA landscape of human lymphocytes highlights the regulation of T cell differentiation by linc-MAF-4'. In: *Nature immunology* 16.3 (2015), pp. 318–325.

[86] I-Cheng Ho, David Lo and Laurie H Glimcher. 'c-maf promotes T helper cell type 2 (Th2) and attenuates Th1 differentiation by both interleukin 4–dependent and–independent mechanisms'. In: *The Journal of experimental medicine* 188.10 (1998), pp. 1859–1866.

[87] Kojiro Sato et al. 'Marked induction of c-Maf protein during Th17 cell differentiation and its implication in memory Th cell development'. In: *Journal of Biological Chemistry* 286.17 (2011), pp. 14963–14971.

[88] Zhonghan Li et al. 'The long noncoding RNA THRIL regulates TNFα expression through its interaction with hnRNPL'. In: *Proceedings of the National Academy of Sciences* 111.3 (2014), pp. 1002–1007.

[89] Bodu Liu et al. 'A cytoplasmic NF-κB interacting long noncoding RNA blocks IκB phosphorylation and suppresses breast cancer metastasis'. In: *Cancer cell* 27.3 (2015), pp. 370–381.

[90] Maninjay K Atianand et al. 'A long noncoding RNA lincRNA-EPS acts as a transcriptional brake to restrain inflammation'. In: *Cell* 165.7 (2016), pp. 1672–1685.

[91] Simon Hirschberger, Ludwig Christian Hinske and Simone Kreth. 'MiRNAs: dynamic regulators of immune cell functions in inflammation and cancer'. In: *Cancer letters* 431 (2018), pp. 11–21.

[92] Ludwig C Hinske et al. 'MiRIAD update: using alternative polyadenylation, protein interaction network analysis and additional species to enhance exploration of the role of intragenic miRNAs and their host genes'. In: *Database* 2017 (2017).

[93] Simone Kreth, Max Hübner and Ludwig Christian Hinske. 'MicroRNAs as clinical biomarkers and therapeutic tools in perioperative medicine'. In: *Anesthesia & Analgesia* 126.2 (2018), pp. 670–681.

[94] Viola Neudecker et al. 'Emerging roles for microRNAs in perioperative medicine'. In: *Anesthesiology* 124.2 (2016), pp. 489–506.

[95] Hiro-oki Iwakawa and Yukihide Tomari. 'The functions of microRNAs: mRNA decay and translational repression'. In: *Trends in cell biology* 25.11 (2015), pp. 651–665.

[96] Arnav Mehta and David Baltimore. 'MicroRNAs as regulatory elements in immune system logic'. In: *Nature Reviews Immunology* 16.5 (2016), pp. 279–294.

[97] Aurora Esquela-Kerscher and Frank J Slack. 'Oncomirs—microRNAs with a role in cancer'. In: *Nature reviews cancer* 6.4 (2006), pp. 259–269.

[98] J Craig Venter et al. 'The sequence of the human genome'. In: *science* 291.5507 (2001), pp. 1304–1351.

[99] 1958 Birth Cohort Controls Jones Richard W. 18 McArdle Wendy L. 18 Ring Susan M. 18 Strachan David P. 19 Pembrey Marcus 18 20, Type 1 Diabetes Clayton David G. 2 Dunger David B. 2 41 Nutland Sarah 2 Stevens Helen E. 2 Walker Neil M. 2 Widmer Barry 2 41 Todd John A. 2 et al. 'Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls'. In: *Nature* 447.7145 (2007), pp. 661–678.

[100] Abdel Abdellaoui et al. '15 years of GWAS discovery: Realizing the promise'. In: *The American Journal of Human Genetics* (2023).

[101] Adrian Liston and An Goris. 'The origins of diversity in human immunity'. In: *Nature Immunology* 19.3 (2018), pp. 209–210.

[102] Vasiliki Lagou et al. 'Genetic architecture of adaptive immune system identifies key immune regulators'. In: *Cell reports* 25.3 (2018), pp. 798–810.

[103] John C Marioni et al. 'RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays'. In: *Genome research* 18.9 (2008), pp. 1509–1517.

[104] Graham RS Ritchie et al. 'Functional annotation of noncoding sequence variants'. In: *Nature methods* 11.3 (2014), pp. 294–296.

[105] Martin Kircher et al. 'A general framework for estimating the relative pathogenicity of human genetic variants'. In: *Nature genetics* 46.3 (2014), pp. 310–315.

[106] Gosia Trynka et al. 'Disentangling the effects of colocalizing genomic annotations to functionally prioritize non-coding variants within complex-trait loci'. In: *The American Journal of Human Genetics* 97.1 (2015), pp. 139–152.

[107] Gleb Kichaev et al. 'Integrating functional data to prioritize causal variants in statistical fine-mapping studies'. In: *PLoS genetics* 10.10 (2014), e1004722.

[108] Denes Hnisz et al. 'Super-enhancers in the control of cell identity and disease'. In: *Cell* 155.4 (2013), pp. 934–947.

[109] Stephen CJ Parker et al. 'Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants'. In: *Proceedings of the National Academy of Sciences* 110.44 (2013), pp. 17921–17926.

[110] Suna Onengut-Gumuscu et al. 'Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers'. In: *Nature genetics* 47.4 (2015), pp. 381–386.

[111] Jian Zhou et al. 'Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk'. In: *Nature genetics* 50.8 (2018), pp. 1171–1179.

[112] David R Kelley et al. 'Sequential regulatory activity prediction across chromosomes with convolutional neural networks'. In: *Genome research* 28.5 (2018), pp. 739–750.

[113] Ronan Collobert and Jason Weston. 'A unified architecture for natural language processing: Deep neural networks with multitask learning'. In: *Proceedings of the 25th international conference on Machine learning*. 2008, pp. 160–167.

[114] Maria V Valueva et al. 'Application of the residue number system to reduce hardware costs of the convolutional neural network implementation'. In: *Mathematics and computers in simulation* 177 (2020), pp. 232–243.

[115] Tong Tong Wu et al. 'Genome-wide association analysis by lasso penalized logistic regression'. In: *Bioinformatics* 25.6 (2009), pp. 714–721.

[116] Steven Gazal et al. 'Reconciling S-LDSC and LDAK functional enrichment estimates'. In: *Nature genetics* 51.8 (2019), pp. 1202–1204.

[117] Peter M Visscher, William G Hill and Naomi R Wray. 'Heritability in the genomics era—concepts and misconceptions'. In: *Nature reviews genetics* 9.4 (2008), pp. 255–266.

[118] 1000 Genomes Project Consortium et al. 'A global reference for human genetic variation'. In: *Nature* 526.7571 (2015), p. 68.

[119] Omer Weissbrod et al. 'Functionally informed fine-mapping and polygenic localization of complex trait heritability'. In: *Nature genetics* 52.12 (2020), pp. 1355–1363.

[120] Hilary K Finucane et al. 'Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types'. In: *Nature genetics* 50.4 (2018), pp. 621–629.

[121] Brendan K Bulik-Sullivan et al. 'LD Score regression distinguishes confounding from polygenicity in genome-wide association studies'. In: *Nature genetics* 47.3 (2015), pp. 291–295.

[122] Luke M Evans et al. 'Comparison of methods that use whole genome data to estimate the heritability and genetic architecture of complex traits'. In: *Nature genetics* 50.5 (2018), pp. 737–745.

[123] Aaron M Newman et al. 'Robust enumeration of cell subsets from tissue expression profiles'. In: *Nature methods* 12.5 (2015), pp. 453–457.

[124] Daniel S Chen and Ira Mellman. 'Oncology meets immunology: the cancer-immunity cycle'. In: *immunity* 39.1 (2013), pp. 1–10.

[125] *MetaCore*. URL: `https://clarivate.com/products/biopharma/discovery-clinical-regulatory/early-research-intelligence-solutions/` (visited on 13/02/2021).

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

[127] Richard D Riley, Julian PT Higgins and Jonathan J Deeks. 'Interpretation of random effects meta-analyses'. In: *Bmj* 342 (2011).

[128] Steven Gazal et al. 'Combining SNP-to-gene linking strategies to pinpoint disease genes and assess disease omnigenicity'. In: *medRxiv* (2021).

[129] Kushal K Dey et al. 'Evaluating the informativeness of deep learning annotations for human complex diseases'. In: *Nature communications* 11.1 (2020), pp. 1–9.

[130] Huwenbo Shi et al. 'Population-specific causal disease effect sizes in functionally important regions impacted by selection'. In: *Nature communications* 12.1 (2021), pp. 1–15.

[131] Toshiyuki Shiraki et al. 'Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage'. In: *Proceedings of the National Academy of Sciences* 100.26 (2003), pp. 15776–15781.

[132] Rimantas Kodzius et al. 'CAGE: cap analysis of gene expression'. In: *Nature methods* 3.3 (2006), pp. 211–222.

[133] Marina Lizio et al. 'Gateways to the FANTOM5 promoter level mammalian expression atlas'. In: *Genome biology* 16.1 (2015), pp. 1–14.

[134] *HUGO Gene Nomenclature*. URL: `https://www.genenames.org/` (visited on 01/02/2021).

[135] Guo-Cui Wu et al. 'Emerging role of long noncoding RNAs in autoimmune diseases'. In: *Autoimmunity reviews* 14.9 (2015), pp. 798–805.

[136] Daniel Peltier et al. 'RNA-seq of human T cells after hematopoietic stem cell transplantation identifies Linc00402 as a regulator of T cell alloimmunity'. In: *Science Translational Medicine* 13.585 (2021).

[137] Masahiro Kanai et al. 'Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases'. In: *Nature genetics* 50.3 (2018), pp. 390–400.

[138] William J Astle et al. 'The allelic landscape of human blood cell trait variation and links to common complex disease'. In: *Cell* 167.5 (2016), pp. 1415–1429.

[139] Po-Ru Loh et al. 'Mixed-model association for biobank-scale datasets'. In: *Nature genetics* 50.7 (2018), pp. 906–908.

[140] Eli Eisenberg and Erez Y Levanon. 'Human housekeeping genes, revisited'. In: *TRENDS in Genetics* 29.10 (2013), pp. 569–574.

[141] Per-Henrik D Edqvist et al. 'Expression of human skin-specific genes defined by transcriptomics and antibody-based profiling'. In: *Journal of Histochemistry & Cytochemistry* 63.2 (2015), pp. 129–141.

[142] Fredrik Pontén, Karin Jirström and Matthias Uhlen. 'The Human Protein Atlas—a tool for pathology'. In: *The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland* 216.4 (2008), pp. 387–393.

[143] COVID-19 Host Genetics Initiative et al. 'The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic'. In: *European Journal of Human Genetics* 28.6 (2020), p. 715.

[144] Kyle Vogan. 'PIK3CD mutations cause immunodeficiency'. In: *Nature Genetics* 45.12 (2013), pp. 1417–1417.

[145] G Stelzer et al. 'GeneCards – the human gene database'. In: *www.genecards.org* (2019).

[146] Charles G Danko et al. 'Dynamic evolution of regulatory element ensembles in primate CD4+ T cells'. In: *Nature ecology & evolution* 2.3 (2018), pp. 537–548.

[147] ENCODE Project Consortium et al. 'An integrated encyclopedia of DNA elements in the human genome'. In: *Nature* 489.7414 (2012), p. 57.

[148] T Kirchhoff et al. 'Germline determinants of immune related adverse events (irAEs) in melanoma immunotherapy response'. In: *Annals of Oncology* 28 (2017), pp. v407–v408.

[149] Annalisa Buniello et al. 'The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019'. In: *Nucleic acids research* 47.D1 (2019), pp. D1005–D1012.

[150] Maria Arshad et al. 'Association of rs182429 variant in TAGAP with rheumatoid arthritis in Pakistani population'. In: *Meta Gene* 14 (2017), pp. 59–63.

[151] Gökhan Cildir et al. 'Genome-wide analyses of chromatin state in human mast cells reveal molecular drivers and mediators of allergic and inflammatory diseases'. In: *Immunity* 51.5 (2019), pp. 949–965.

[152] AS Parmar et al. 'Association of celiac disease genes with inflammatory bowel disease in Finnish and Swedish patients'. In: *Genes & Immunity* 13.6 (2012), pp. 474–480.

[153] Ali Jazaeri and Sadeq Vallian. 'Association of rs1738074 polymorphism of TAGAP gene with susceptibility to multiple sclerosis in the Iranian population'. In: *Neuroscience Letters* 648 (2017), pp. 66–69.

[154] B. Alberts et al. *Molecular biology of the cell. 4th edn.* 2002.

[155] Mutsumi Kanamori-Katayama et al. 'Unamplified cap analysis of gene expression on a single-molecule sequencer'. In: *Genome research* 21.7 (2011), pp. 1150–1159.

[156] Harry Towbin, Theophil Staehelin and Julian Gordon. 'Electrophoretic transfer of proteins from polyacrylamide gels to nitrocellulose sheets: procedure and some applications.' In: *Proceedings of the national academy of sciences* 76.9 (1979), pp. 4350–4354.

[157] Anthony Scott et al. 'Regulation of protein stability of DNA methyltransferase 1 by post-translational modifications'. In: *Acta Biochim Biophys Sin* 46.3 (2014), pp. 199–203.

[158] Timothy H Bestor. 'The DNA methyltransferases of mammals'. In: *Human molecular genetics* 9.16 (2000), pp. 2395–2402.

[159] Babak Alipanahi et al. 'Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning'. In: *Nature biotechnology* 33.8 (2015), pp. 831–838.

[160] Jian Zhou and Olga G Troyanskaya. 'Predicting effects of noncoding variants with deep learning–based sequence model'. In: *Nature methods* 12.10 (2015), p. 931.

[161] Allison J Shultz and Timothy B Sackton. 'Immune genes are hotspots of shared positive selection across birds and mammals'. In: *Elife* 8 (2019), e41815.

[162] Tiffany Amariuta et al. 'Improving the trans-ancestry portability of polygenic risk scores by prioritizing variants in predicted cell-type-specific regulatory elements'. In: *Nature genetics* 52.12 (2020), pp. 1346–1354.

[163] Erez Lieberman-Aiden et al. 'Comprehensive mapping of long-range interactions reveals folding principles of the human genome'. In: *science* 326.5950 (2009), pp. 289–293.

[164] Alan P Boyle et al. 'High-resolution mapping and characterization of open chromatin across the genome'. In: *Cell* 132.2 (2008), pp. 311–322.

[165] Juba Nait Saada et al. 'Inference of genealogical relatedness and variant age using convolutional neural networks'. In: *Under Review* (2022).

[166] Matthew D Rasmussen et al. 'Genome-wide inference of ancestral recombination graphs'. In: *PLoS Genetics* 10.5 (2014), e1004342.

[167] Graham McVicker et al. 'Widespread genomic signatures of natural selection in hominid evolution'. In: *PLoS genetics* 5.5 (2009), e1000471.

[168] Pier Francesco Palamara et al. 'High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability'. In: *Nature Genetics* 50 (Sept. 2018).

[169] Le Zhang et al. 'CpG-island-based annotation and analysis of human housekeeping genes'. In: *Briefings in bioinformatics* 22.1 (2021), pp. 515–525.

[170] Yiming Hu et al. 'Leveraging functional annotations in genetic risk prediction for human complex diseases'. In: *PLoS computational biology* 13.6 (2017), e1005589.

[171] Arturs Kalnins. 'Multicollinearity: How common factors cause Type 1 errors in multivariate regression'. In: *Strategic Management Journal* 39.8 (2018), pp. 2362–2385.