Deciphering the Tissue-specific Genetic Architecture of Two Complex Diseases

Daniel SikWai Ho

Under the supervision of

Professor Justin O'Sullivan

The Liggins Institute

A thesis submitted in fulfilment of the requirements for the degree of
Doctor of Philosophy in Biomedical Science, The University of Auckland, 2021

Abstract

Complex diseases impact millions of people worldwide, caused by a variety of genetic and environmental factors. Genome-wide association studies have done an adequate job of identifying genetic variants associated with these complex diseases. Subsequently, polygenic risk models have been used to predict the disease risk of individuals with meaningful accuracy. However, the association studies and risk modelling cannot determine nor predict the underlying genetic architecture of the associated variants. In this thesis, I have developed a computational approach that integrates complex disease variants, their related tissue-specific gene regulation information, and individual genotype data. The essential information was selected from the combined data by the Mann Whitney U test and machine learning regularization. This information was then evaluated by a series of logistic regression predictor models to predict individual disease risk. With validation across multiple genotyped populations, the best predictor model was used to identify the most predictive regulatory elements conferring the complex disease risk. Applying this computational approach to study T1D and PD, my regularized predictor models revealed tissue-specific gene regulation impacting T1D and PD disease risk. The regularized logistic regression models supported a clear platform for interpreting the molecular mechanisms underlying the genetic components of the predictor model. These analyses implicate important insights into the mechanisms acting on different tissues to modulate T1D and PD onset and development.

The novelties of the regularised predictor modelling approach are the ability to distinguish trans and cis eQTL regulatory effects of disease-associated SNPs across tissues. Using Mann Whitney U Test filtering controlled by Benjamini Yekutieli FDR and machine learning regularisation, I can establish the curated associations of the eQTL regulatory effects in different tissues. Furthermore, my predictor models can estimate the risk contribution of each tissue-specific eQTL regulatory effect for identifying the crucial tissues and their essential SNP modulated eQTL elements.

This thesis is dedicated to

my Lord and my saviour

Jesus Christ

and

my most respected supervisor

Professor Justin O'Sullivan

Acknowledgements

My journey of this PhD research was started by my desire to reconcile the disagreements of my Christian beliefs and the evolution theory. As a Christian growing up in churches, I had been taught that the evolution theory was non-biblical. Nevertheless, genetics research had confirmed many findings of the evolution theory. My heart was full of confusion.

Around the year 2010, I was blessed to study Information of Communication Technologies at Massey University, and my Lord Jesus Christ also led me to study gene regulation under Professor Justin O'Sullivan for allowing me to explore the evolution theory through genetics. Justin showed me his kindness and gave me the opportunity to participate in his research with my computational skills even though I was lack of biological knowledge. The valuable experience helped me realize that I needed some good statistics understanding in order to apply my computational skills in bioinformatics research. With Justin's blessings and encouragement, I went to study applied statistics. And through the study, I gained a good understanding of randomness, which helped me to resolve the confusion of my heart and to believe that God could use randomness and evolution to create this world advancing his own interests.

After finishing my Master degree in genetics, my Lord once again used Justin to bless me for giving me the opportunity to study under his guidance with a scholarship for my PhD research. In the past few years, Justin gave me his invaluable attention, advice and resources for supporting me to grow and be transformed in different knowledge. I have been enjoying every single step of this PhD research that was the best journey I could ever dream of. Good time always goes too fast. Coming to the end of my PhD study, my heart is full of gratitude. I could not express enough my special thanks to Justin for his support and encouragement. As a student, I seek honour for my teachers. I truly wish that my PhD research could bring some joy to my most respected supervisor Professor Justin O'Sullivan and honour to his name. From now on, my prayers will be continuously for my Lord Jesus Christ to shower his blessings on Justin for his health, family and every step of his future, rewarding his kindness for me.

I was blessed to be supervised by a group of excellent and outstanding scientists. They showed me their kindness, encouragement and support that enabled me to complete my PhD journey. I would like to express my special thanks to my supervisor Dr William Schierding who closely supported me in every step of my research and gave me his helpful feedbacks and precious ideas. I also like to express my special thanks to my supervisor Dr Andreas Kempa-Liehr who enabled me to have a better understanding of machine learning and gave me his invaluable assistance in solving many difficult machine-learning technical issues. Furthermore, I would like to express my special gratitude to my supervisor Prof. Melissa Wake and my advisor Prof. Richard Saffery for their priceless support and kindly encouragement that helped me to develop the required skills at the beginning, and the skills became the cornerstone of my PhD research. My prayers will always be with my supervisors and advisor for blessing them everywhere they go.

Additionally, I like to take this opportunity to express my special thanks to all of my friends and colleagues in Justin's lab for their friendship and support in the past three and half years. I have learned so much from Dr Tayaza Fadason and Dr Denis Nyaga. They always gave me their selfless support whenever I needed their help. I sincerely thank Evgeniia Golovina for her support and for showing me so many good things about Russian culture. I would like to express my special thanks to Sreemol Gokuladhas and Clara Chone for their precious friendship and encouragement. I would also like to express my gratitude to Sophie Farrow, Brook Wilson and all other past and present members of Justin's lab for your endless help and support. It was absolutely my honour and privilege to study along with all the outstanding people in Justin's lab during my PhD research.

Last but not least, I would like to thank all the staff and students at Liggins Institute. None of my research would be possible without your help and support.

To all of you, thank you!

Table of Contents

List of Figures

List of Tables

Glossary

3'UTR	Three prime untranslated region
58C	1958 British Birth cohort
A549	Lung epithelial carcinoma cells
AF	Atrial fibrillation
aFC	The log ratio of the haplotype expression with an alternative allele (SNP) to the haplotype expression with a reference allele
AP4B1-AS1	AP4B1 Antisense RNA 1 gene
AUC	Area under a ROC curve
BGEN	Binary GEN file format
BY	Benjamini Yekutieli procedure
CAMTA1	Calmodulin Binding Transcription Activator 1 gene
cis-eQTL	An eQTL that is <1 Mb from the gene associated with it
CLEC16A	C-Type Lectin Domain Containing 16A gene
CNTN1	Contactin 1 gene
CO_2	Carbon dioxide
CoDeS3D	The Contextualising Developmental SNPs in three-dimensions algorithm
CTLA-4/ CTLA4	Cytotoxic T-Lymphocyte Associated Protein 4 gene
DIS3L2	DIS3 Like 3'-5' Exoribonuclease 2 gene
DNA	Deoxyribonucleic acid
E. coli	Escherichia coli
EAF1	ELL Associated Factor 1 gene
EAF1-AS1	EAF1 Antisense RNA 1 gene
eGenes	Genes modulated by eQTLs
eQTLGen	The eQTLGen Consortium
eQTLs	Expression quantitative trait loci
FDR	False discovery rate
FOXP1	Forkhead Box P1 gene
FOXP3	Forkhead Box P3 gene
GBA	Glucosylceramidase Beta gene
GEN	Oxford text genotype file format
GM12878	B-cell derived lymphoblastoid cell line
gnomAD	The Genome Aggregation Database
GRCh37	Genome Reference Consortium Human Build 37
GRS	Genetic risk scores
GTEx	The Genotype-Tissue Expression project
GWAS	Genome Wide Association Studies
H3K9ac	The acetylation at the 9th lysine residue of the histone H3 protein
HeLa	Human cervix cells
HepG2	Human liver carcinoma cells
Hi-C	Chromosome conformation capture (all-vs-all)
HLA	Human leukocyte antigen complex
HMEC	Human Mammary Epithelial Cells
HUVEC	Human umbilical vein endothelial cells
IBD	Identity By Descent
IDT	Integrated DNA Technologies
IFIH1	Interferon Induced With Helicase C Domain 1 gene
IGF2BP2	Insulin Like Growth Factor 2 MRNA Binding Protein 2 gene
IL2RA	Interleukin 2 Receptor Subunit Alpha gene
IMR90	Human foetal lung cells
INPP5F	Inositol Polyphosphate-5-Phosphatase F gene
INS	Insulin gene
IPDGC	The International Parkinson Disease Genomics Consortium
K562	Human bone cells
KBM7	Human chronic myeloid leukemia cells

KIAA1430	CFAP97 (Cilia And Flagella Associated Protein 97) gene
*Kpn*I	A restriction enzyme
L1	Lasso regularisation
L2	Ridge regularisation
LD	Linkage disequilibrium
LDSC	Linkage disequilibrium score regression
LRRK2	Leucine Rich Repeat Kinase 2 gene
luc2	Luciferase gene
LyP	lymphoid-specific intracellular phosphatase
MICA	MHC Class I Polypeptide-Related Sequence A gene
MR	Mendelian Randomization
NES	Normalized effect size
NHEK	Human Epidermal Keratinocyte cells
NOTCH4	Notch Receptor 4
PCA	Principal component analysis
PD	Parkinson's disease
PINK1	PTEN Induced Kinase 1 gene
PLINK	A free, open-source whole genome association analysis toolset
pMPRA1	Addgene: plasmid #49349
pMPRAdonor2	Addgene: plasmid #49353
PRS	Polygenetic Risk scoring
PSMB9	Proteasome 20S Subunit Beta 9 gene
PSORS1C1	Psoriasis Susceptibility 1 Candidate 1 gene
PTPN2	Protein Tyrosine Phosphatase Non-Receptor Type 2 gene
PTPN22	Protein Tyrosine Phosphatase Non-Receptor Type 22 gene
pymc3	A probabilistic programming package for Python
R software	A free software environment for statistical computing and graphics
r^2	Correlation
R620W	A missense mutation of PTPN22
RBM47	RNA Binding Motif Protein 47 gene
RNA	Ribonucleic acid
RNF5	Ring Finger Protein 5 gene
ROBO2	Roundabout Guidance Receptor 2 gene
ROC	Receiver operating characteristic curves
rsID	Reference SNP cluster ID
Scikit-learn	A machine learning package for Python
SLAMF1	Signaling Lymphocytic Activation Molecule Family Member 1 gene
SNCA	Synuclein Alpha gene
SNP	Single nucleotide polymorphisms
STAT3	Signal Transducer And Activator Of Transcription 3
SVM	Support Vector Machine
T1D	Type 1 diabetes
T2D	Type 2 diabetes
TATA-box	A sequence of DNA found in the core promoter region of genes in eukaryotes
TEDDY	The Environmental Determinants of Diabetes in the Young
TMEM161B-AS1	TMEM161B Antisense RNA 1 gene
trans-eQTL	An eQTL that is > 1 Mb from the gene associated with it
TRIM26	Tripartite Motif Containing 26 gene
tsfresh	A package for Python to extract characteristics from time series data
UKBB	UK Biobank
WTCCC	Wellcome Trust Case and Control Consortium
*Xba*I	A restriction enzyme

Co-Authorship Forms

Chapter 1: General Introduction

1.1 Complex diseases

Complex diseases result from the combined effects of a variety of causes that include genetic, ageing, lifestyle, and environmental factors[1,2]. Examples of complex diseases include obesity, diabetes, auto-immune diseases, and neurodegenerative disorders[1,2]. These diseases impact millions of people globally[3–5], including in NZ, where more than 30% of adults are obese[6]. Genetically, determining the underlying cause of each complex disease can be difficult, as they are typically polygenic and do not follow a strict mendelian inheritance pattern[1,2]. Additionally, the impact of genetics on each complex disease varies widely based on a number of variables, such as age at onset. It is believed that early onset complex disorders (*e.g.* Type 1 Diabetes) have a strong genetic influence, while late onset (*e.g.* Parkinson's disease) have a much weaker genetic basis[7,8].

Type 1 diabetes (T1D) is an early onset complex disease that affects over 6 million children under 15 years of age (as of 2019)[9,10]. The rates of T1D are increasing worldwide[9,10]. T1D is characterized by T cell mediated auto-immune destruction of pancreatic beta cells leading to insufficient insulin secretion[11]. T cells play a major role in adaptive immune defence, directly assaulting and concerting other immune insults to foreign invaders[12]. In T1D, the T cells mistakenly attack and destroy self-pancreatic beta cells[9,10]. This auto-immune attack is preceded by a long developmental period, with damage to the pancreas progressing slowly, gradually reducing the mass of pancreatic beta cells[13]. Eventually, a critical point is reached where the beta cell mass drops below a level from which there is no hope for recovery[13]. While the exact cause of T1D onset and progression are unknown, around 50% of the disorder risk is heritable[14], indicating a strong genetic component. For example, it has been shown that T1D associated genetic variants participate in tissue-specific regulation of genes across multiple tissues[15].

There are a wide number of recognized environmental components influencing T1D onset, consistent with a myriad of possible triggers, including viral exposure[16,17]. From the BABYDIET Study, Beyerlien et al. have found evidence to support the associations of T1D with respiratory infections at early childhood[16]. George and colleagues also found T1D patients with 62% increased risk for having respiratory infections[17]. Therefore, genetic susceptibility to seemingly disconnected phenotypes (*e.g*., viral infections) may also promote T1D risk. As such, the genetic contributions to T1D are complex. Currently, more than 60 loci have been associated with T1D[18,19]. In addition, many T1D associated genes have been identified, using familial studies, including *PTPN22, CTLA-4, PTPN2, INS, IL2RA, IFIH1* and *CLEC16A*[14].

Parkinson's disease (PD) is a late onset complex disease, the second most prevalent neurodegenerative disorder of ageing[20–22]. In 2016, 6.1 million patients were diagnosed with PD globally, with the incidence rate increasing every year[23]. PD development is slow and progressive[20,22,24,25], characterized by the loss of dopamine-producing neurons and the presence of intracellular protein aggregates (Lewy bodies) in brain tissues[20,26]. The disease involves a variety of motor and non-motor symptoms (*e.g*. the signature hallmark tremoring, sleep disorders, constipation, depression, heart rate variation and smell dysfunction)[20,24,25]. Studies show that non-motor features could appear more than 5 years before the motor symptoms[20,24,25]. The motor and non-motor PD features suggest the involvement of multiple-tissue interplays in the disorder pathogenesis[20,24,25].

The genetic contribution to PD risk has been estimated at over 30%[20,27,28]. Mutations of the *SNCA* gene, which encodes the α-synuclein protein, were the first genetic variations that were recognized to play a vital role in PD pathogenesis and the formation of Lewy bodies[26,29]. Subsequently, many other genes (*e.g. PINK1*, *LRRK2*, *GBA* and *INPP5F*) have been associated with the risk of developing PD[20,26,30–33]. Despite the recognized genetic contributions to PD, it is clear that there is also a strong environmental contribution to the aetiology of the disorder[22,34].

1.2 Genome Wide Association Studies

Advances in DNA sequencing technologies following the completion of the human genome project in 2003 have led to the collection of human genomic data at an exponential rate[35–38]. This sequencing has identified millions of single nucleotide polymorphisms (SNPs, also known as genetic variants) in the human genome[39].

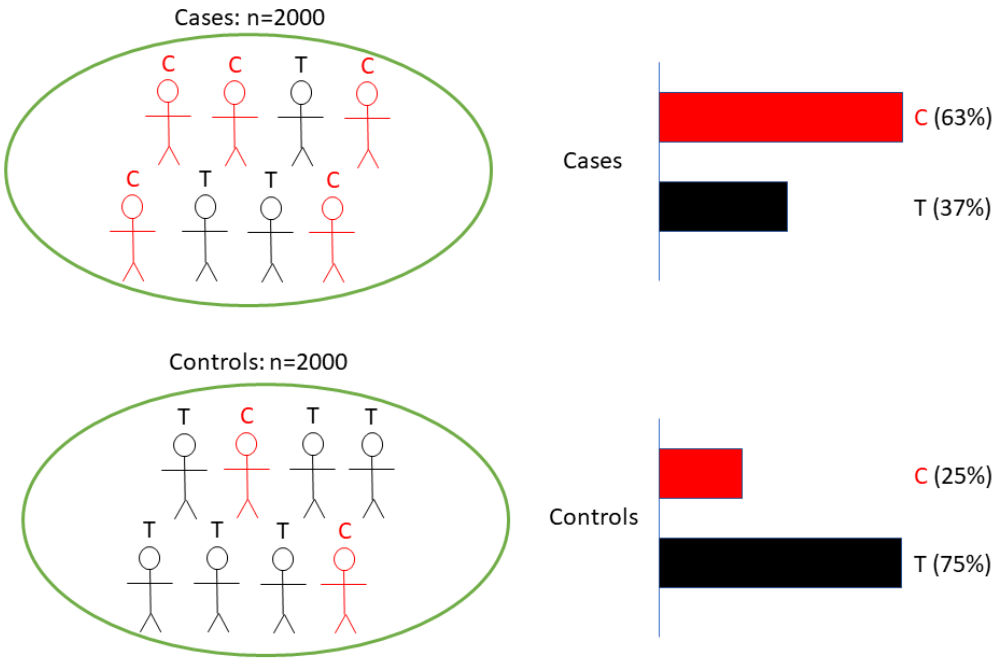


Figure 1-1: Schematic figure of GWAS SNP association

Genome Wide Association Studies (GWAS) compare the genomic information of the case and control individual samples in a population to identify SNPs that are statistically associated with a phenotype[40,41] (Figure 1-1). Genome-wide associations are determined as being significant if the p-value < $5x10^{-8}$ as a typically acceptable threshold. Often the level is dependent on the number of SNP investigated. Nevertheless, this threshold is considered by many to be too conservative[42,43]. This leads to the use of more relaxed p-value thresholds (*e.g.* $1x10^{-5}$)[44–46].

Using large cohorts, GWAS have successfully revealed insights into how polygenic mechanisms affect complex disease development across different complex disorders, including diabetes, Parkinson's disease, auto-immune diseases, and schizophrenia, in the past ten years[40,41,47–51]. As of May 2021, the GWAS catalog (a publicly available database of GWAS SNP information) contains 2,518 published studies[52]. Notably, most complex diseases are influenced by hundreds of SNPs, each imparting a small per-SNP effect size[40]. Of note, the majority of these SNPs are located in non-coding regions and thus must be indirectly involved in their disease association, likely through tissue-specific regulatory activities[40,53]. New methods to understand these regulatory activities include the incorporation of spatial and temporal aspects of gene expression data[15,54–56]. These approaches are providing insights into the impacts of genetic variants that can be incorporated into new approaches to create population-based risk models for predicting individualised risk.

1.3 Predicting Risk Scores and AUC

Population-based risk prediction models serve a robust purpose in disease prediction and prognosis. Specifically, they are especially useful in choosing efficacious treatments without the need for costly and potentially adverse medical screening procedures (*e.g.* invasive biopsies)[57,58]. Thus, the main focus of developing genetic risk models is to achieve accurate predictive power for recognising at-risk individuals in a robust manner[57]. Traditional epidemiological models of disease risk (with limited predictive power) have been primarily informed by lifestyle risk factors such as family history[59,60]. Recently, the inclusion of genetic risk factors, including disease or phenotype associated SNPs, into risk modelling has improved the accuracy of individual disease prediction[59,60]. However, disease and phenotype associations from GWAS inform on genetic contributions to risk at a population level. Therefore, the incorporation of GWAS SNPs into a risk prediction model requires their translation to individual disease risk. The translation of disease related SNPs into individualised risk scores can be achieved by integration of GWAS model weights into risk models to score an individual's genotype and estimate total individual risk.

Genetic risk prediction models are typically constructed by: 1) Regression-based Polygenetic Risk scoring ; or 2) Machine Learning modelling[57,61]. For simplicity, regression-based Polygenetic Risk Scoring is referred as Polygenetic Risk Scoring (PRS) from this point on. PRS sums the impact of a set of risk alleles weighted or unweighted by their odds ratios or effect sizes related to a specific disease[57,61,62]. By contrast, machine learning approaches adapt sophisticated statistical algorithms (*e.g.* Support Vector Machine or Random Forest) to mathematically map the predicted complex associations between a set of risk alleles to disease phenotypes[57,61,63]. The predictive performance of both model types can be evaluated by receiver operating characteristic curves (ROCs)[59,60,64], where the sensitivity and specificity of the predictions are tested at various cut-off values[59,60,64]. The area under a ROC curve (AUC) reflects the probability of the examined model correctly identifying a dichotomous phenotype, *e.g.* the presence or absence of the disease, from a randomly selected sample[59,60,63,64]. AUC results range from 0 to 1. When the AUC = 0.5, the risk model randomly chooses the dichotomous phenotype and is correct 50% of the time. On the other hand, when the AUC = 1, the model recognises the correct phenotype with 100% accuracy[59,60,64].

1.4 Factors for Improving the Predictive Power

Despite initial promise, the predictive performance of PRS for complex diseases has only been moderately successful[57,61,63]. A significant contributor to this relatively poor performance revolves around the finding that experimental GWAS data suggest risk allele contributions to complex diseases have average odds ratios of between 1.1 – 2, as shown by Wray *et al.*[58] However, GWAS analyses are typically underpowered and detect risk SNPs with odds ratios greater than 1.3 [65,66]. Thus, improving the predictive power of disease risk models could be as simple as increasing GWAS sample sizes[57,61,67]. Rapidly decreasing DNA sequencing costs have led to GWAS sample sizes increasing from a few thousand to nearly half a million per meta-analysis[68–70]. These increased sample sizes have increased the frequency of detection of SNPs with small effect sizes, resulting in an increase in accuracy for complex disease predictions[57,61]. Many future disease risk prediction models can be built on big datasets that could not be acquired before.

Big data is a critical element in machine learning modelling. Training datasets need to be sufficiently large to ensure the good accuracy of the prediction models. Wei *et al*. illustrated the impacts of training sample size on the predictive power of a machine learning classification algorithm for Inflammatory Bowel Disease[71]. The dataset used in the study contained 60,828 individual genotype samples from 15 countries in Europe[71]. A machine learning prediction model created from a small subset of the dataset only performed moderately. Nevertheless, the predictive power of the same model consistently improved with increases in the training data sizes until the predictive performance reached the maximum with the full training dataset[71]. Therefore, machine learning model performance is based on the size and quality of the data from which the model is created.

Technological advances are constantly improving the quality and quantity of the complex integrative datasets that are collected on human phenotypes and diseases. Integration of these highly dimensional genomic data within computational models can lead to improvements in genetic risk prediction over that achieved for PRS[67]. PRS predictions are based on a linear parametric regression model which incorporates strict assumptions that include additive and independent predictor effects, normal distribution of residuals, and the data observations being non-correlated[57,61,72].

	Polygenic Risk Scoring	Machine Learning Model
Strengths	• Easy and effective to apply • Easy to interpret the results	• Effective for modelling multi-dimensional data • Account for complex data interactions • No normal distribution assumption for underlying data
Weaknesses	• Additive and independent predictor effects • Normal distribution of underlying data • Not account for complex data interactions	• Difficult to apply • Difficult to interpret the underlying genetic effects from the results • Need a big dataset

Figure 1-2: The strengths and weaknesses of Polygenic Risk Scoring and Machine Learning Model

These PRS regression model assumptions may not hold true for the fundamental genetic structures of complex polygenic diseases, thus leading to greatly reduced predictive efficacy[57,61]. Furthermore, linear additive regression modelling is incapable of accounting for complex interactive effects between associated alleles[57,67], which have been reported to make major contributions to phenotypes[73]. Thus linear additive regression based modelling leads PRSs toward biased and less effective predictions[57,67,74,75]. By contrast, machine learning algorithms employ multivariable, non-parametric methods that robustly recognise patterns from non-normally distributed and strongly correlated data[57,61,67]. The capacity of machine learning algorithms to model highly interactive complex data structures has led to these approaches receiving increasing levels of interest for complex disease prediction[57,61,67]. The strengths and weaknesses of both PRS and machine learning models are shown in Figure 1-2.

1.5 Machine Learning Disease Prediction Models

Machine learning data modelling approaches that describe the associations of genetic information with different complex diseases are either supervised or unsupervised[76]. Although unsupervised machine learning methods and non-genetic data are useful in disease predictions[77,78], we will focus on supervised modelling that is informed by genetic variant (SNP) data.

Supervised machine learning predictions are classification (estimation of a binary phenotype categorical variable, *e.g.*, case/control) or regression (estimation of a continuous variable, *e.g.*, the probability of a case). Supervised disease prediction models are generated by training the pre-set learning algorithms (training phase) to map the relationships between individual sample genotype data and the associated disease[67,76]. The learning algorithms define the structure of the independent data variables (features) in the data mapping and produce models to predict the target disease status. The models then represent a set of features (with their interrelationships) that confer the disease risk. Optimal predictive power for the target disease is achieved by mapping the pattern of the selected features within the training genotype data[57,67]. Some model optimizations use gradient descent procedures with iterative steps to search for optimised predictive power[79,80]. This recursive process continues until the optimal predictive performance is reached[79,80]. At the end of the training stage, the models with the maximum predictive power on the training dataset are selected for validation[57,81]. A generalized workflow for creating a machine learning model from a genotype dataset is illustrated in Figure 1-3.

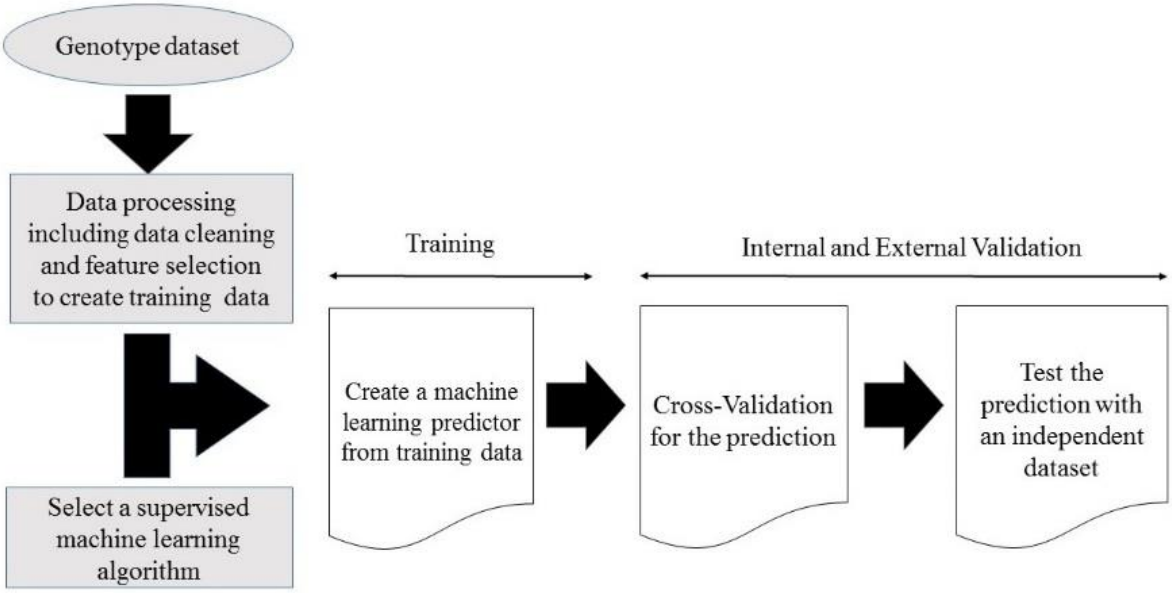


Figure 1-3: Workflow for creating a supervised machine learning model from a genotype dataset

During the validation stage, the performance of the predictive machine learning models is evaluated to determine their power for generalised prediction. As with polygenic risk scoring, the validation stage is accomplished by evaluating the algorithm on an independent dataset. The validation stage is essential for checking whether the prediction models overfitting the training data[57,67,76]. Cross validation is a commonly used procedure for validating the models' performance using the original dataset [63,81–84]. However, external validation (testing) using an independent dataset is required to finally confirm the predictive power of a machine learning model. The utility of the algorithm is finally determined through randomised controlled comparisons to current clinical best practices. Only if the algorithm adds information to more accurately stratify populations, predict disease risk or treatment responses does it ultimately prove its clinical utility.

1.6 Feature Selection and Regularisation

The selection of data features during the training phase is the major factor that impacts on a machine learning model's predictive performance[67,85,86]. Data features can be selected manually, by embedded machine learning modules, or by wrapper methods[63,67,85]. For building models predicting complex polygenic disease, SNPs are currently considered the most informative data features within genetic data[71,87]. The SNPs present in the data enable the model to recognise a genetic signature key to predicting the disease. Notably, it is assumed that the SNPs that are selected for inclusion (*e.g.* GWAS SNPs) are associated with loci that contribute mechanistically to the underlying disease etiology[88]. However, how the SNP relates to the disease may or may not pass through currently understood disease related biological mechanisms.

Regularisation is a technique used for tuning the feature selection process by adding an additional error term. This process maximises the generalised predictive power of machine learning models by keeping them simple and preventing extreme feature weights within the model. This prevents models so closely fitted to a dataset that they cannot generalise to other datasets[67]. The most common types of regression-based regularisation are L1 (Lasso), L2 (Ridge), and Elastic net. L1 and L2 regularisations both use a penalised loss function to minimise weight (model coefficient) values for adjusting data feature effects and keep the regression models from complexity[67].

L1 penalised loss function: $\lambda \sum_{k=1}^{n} |W_k|$ in which λ is the regularization strength, n is the number of model weights, and W is the model weight[67].

L2 penalised loss function: $\lambda \sum_{k=1}^{n} (W_k)^2$ in which λ is the regularization strength, n is the number of model weights, and W is the model weight[67].

L2 regularisation shrinks model weights to non-zero small value sizes[67] for non-essential data features. By contrast, L1 regularisation sets the weights of non-informative data features to zero for eliminating the effects and allowing only essential and valuable data feature effects to be included into the model[67]. Regression-based L1-regularisation is one of the most commonly used machine learning feature selection methods, with Lasso and Elastic Net being the most popular regularisation modules[67]. Elastic net is adopting a regularisation approach in-between L1 and L2 using both their penalised loss functions, which removes non-informative features but retains some redundant valuable information in the modelling[67,89,90].

There are many examples where L1-regularization of machine learning applications has enhanced the algorithm's predictive performance for different diseases[71,87,91,92]. Wei *et al.* implemented a two-step model training process in the development of an L1-regularized algorithm for Crohn's disease prediction[71]. Firstly, the Lasso-logistic regression method identified a set of essential and informative SNPs. Subsequently, the selected SNPs were applied to a Support Vector Machine (SVM) and a logistic predictor for Crohn's disease. Following SNP optimisation by L1-regularization, both the non-parametric and parametric predictors achieved similar outstanding results with an AUC of 0.86 compared to the simple PRS prediction, which had an AUC of 0.73. This example highlights the important role of data feature selection in helping predictive models to achieve enhanced disease predictive power.

A Lasso–SVM integrated model was also reported for celiac disease on multiple European genotype datasets, with an AUC of 0.9 [93]. A disease prediction model with an AUC this high could be considered for clinical use and has already led to the exploration of possible clinical applications for the celiac disease predictive model[57]. The predictive power of the celiac disease predictor is further adjusted and evaluated by various clinical factors (*e.g.* disease prevalence rate) in therapeutic decision making[93]. In addition, the identification of the essential SNPs by the Lasso-SVM model provided a genetic basis for deciphering the etiologic pathways of Celiac disease pathogenesis.

1.7 Regression- and Tree-based algorithms

Supervised learning algorithms can be classified as regression-based or tree-based methods[67,76]. The regression-based supervised learning methods employ polynomial parametric or non-parametric regression methods to map the associations of the multidimensional input data to the outputs[67,76,79]. The popular regression based supervised algorithms are logistic regression, linear regression, neural networks and SVM[63,76]. By contrast, tree-based supervised learning algorithms utilise binary decision splitting rule approaches to model the relationships between the input and output data[67,76,79]. The most popular tree-based learning methods are Decision tree and Random forest[63,76]. Common supervised learning models are listed in Table 1-1.

Table 1-1: Types of machine learning algorithms

Regression Based:	
Logistic regression	• Use parametric regressions to estimate the probabilities of dichotomous outputs[76,94–96]
Neural Network	• Use multi-layers of non-parametric regressions and transformations to model input data to outputs[79,97–99]
Support Vector Machine (SVM)	• Use non-parametric regressions to model input data for creating multi-dimensional hyperspaces to discriminate the outputs[93,100–102]
Regression Based Regularization:	
Lasso	• Apply L1 penalised loss functions in regression[67,71,103,104]
Elastic Net	• Apply L1 and L2 penalised loss functions in regression[67,87,90,105]
Tree-Based and Ensemble-Learning:	
Decision Tree	• Utilise binary decision splitting rule approaches to model the relationships between input data and outputs[79,106–108]
Random Forest	• Utilise an ensemble of randomised decision trees to model input data to outputs[78,79,109,110]
AdaBoost	• Utilise the linear combined predictive power of a group of weak classifiers with weighting to model the relationships between input data and output[111,112]

Regression-based machine learning has been widely employed in the risk prediction of many disease risks, including cancer, Alzheimer's disease, cardiovascular disease, and diabetes[100,113–116]. An SVM regression-based non-parametric machine learning model of the genetics of type 1 diabetes was built and trained from 3,443 individual genotype samples[117] achieving an AUC of 0.84, which is significantly higher than the PRS model AUC of 0.71[59,61,75]. Notably, the non-parametric SVM consistently outperformed the parametric control prediction model (logistic regression) on two independent datasets in predictive power validation[61].

Tree-based machine learning is largely an implementation of the Random Forest algorithm[88,118–120]. The Random Forest algorithm constructs prediction models using an ensemble method with many decision trees, and the algorithm selects for and evaluates SNPs that are informative in the decision-tree building[83,118]. A strength of Random Forest models is their ability to effectively handle missing data and highly dimensional data structures that contain complex interactions[83,118]. For example, in a recent study, the Random Forest algorithm was used to predict Type 2 diabetes risk (AUC = 0.85), outperforming both SVM and logistic regression models[88]. Similar to Lasso-SVM modelling, the Random Forest model identified a set of relevant SNPs that are strongly associated with type 2 diabetes and can be used to interrogate the aetiology of the disease[83,88,118]. The accumulated evidence supports the utility of the Random Forrest algorithm as a useful machine learning method for complex disease risk modelling[88,118,121,122].

Ensemble learning is a method to combine the power of a group of weak classifiers and deliver good prediction[111,123]. AdaBoost first constructs a base prediction model and then recursively creates models to focus on correcting the errors of the previous models[111,123]. In the end, all the models are combined linearly with the weighting of their contribution to form a strong model that delivers outstanding prediction[111,123]. Makarious *et al.*[124] applied the AdaBoost Classifier algorithm to build a predictor model for Parkinson's Disease from 598 case and control samples with their extensive data including genotype, transcriptomic, clinical, and demographic information[124]. AdaBoost classifier was selected for the best performance and delivered PD prediction with AUC = 0.85 on the validation data[124]. Critically, the ability to handle multiple data types and models meant that the AdaBoost classifier outperformed classifiers that were based solely on one or a subgroup of the data types (genotype, transcriptomic, clinical, and demographic).

1.8 Spatial Contacts Reveal the Underlying Mechanism of DNA Variant Modulated Gene Regulation

The expression of different genes is tightly regulated and controlled in order to orchestrate sufficient and appropriate gene products for carrying out a variety of biological and physiological functions[125]. Gene regulation can be executed specifically in different organs and tissues to support local cell activities[126,127]. The deregulation of gene expression will lead to different kinds of diseases[128,129].

Spatial genome organisation represents the impact of all nuclear functions (*e.g.* transcription, repair, and replication) that are occurring within a cell or tissue on the DNA structure. As such, the spatial organisation of genomes represents the key to unlocking genome biology. Chromatin is organized into cell-type specific three-dimensional structures within nuclei. These structures emerge from the totality of functions that are present within the nucleus, including gene regulation and their responses in different physiological environments[130–133] Hi-C experiments are designed to describe 3D genomic organizations in cells with proximal DNA ligation fragments paired with high-throughput parallel sequencing[134]. Chromatin structures form loops and provide platforms for bringing various distance regulatory elements to their target genes[130,132,135] (Figure 1-4). Rao *et al.*[132] employed Hi-C experiments to analyse nine cell types and detected 4.9 billion DNA interactions. They also discovered around 10,000 loop DNA structures that involved enhancer and promoter elements coupled with gene activation[132]. The results support that 3D chromatin conformation is a vital mechanism of gene regulation and Hi-C data can effectively capture the distal and proximal DNA regulatory elements interacting with their target genes.

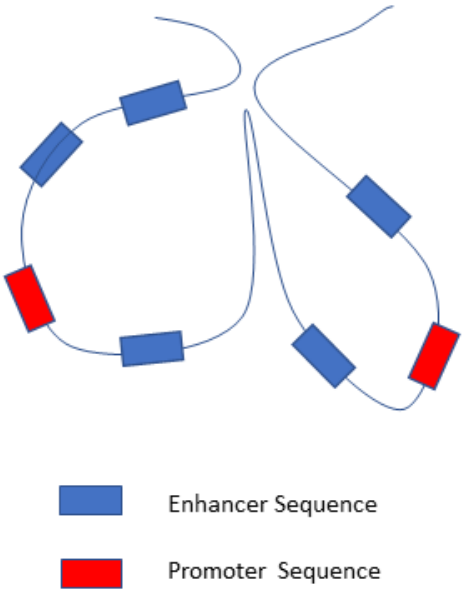


Figure 1-4: Schematic figure of DNA loops

The SNPs associated with gene expression changes are called expression quantitative trait loci (eQTLs) and those genes modulated by the eQTLs are called eGenes. When the SNPs are more than 1Mb apart from their eGenes, they are called *trans*-eQTLs. Otherwise, the SNPs are *cis*-eQTLs (SNPs < 1 Mb from their eGenes). One form of eQTL effect is measured as normalized effect size (NES)[127], where NES is the normalized linear ratio of the tissue-specific gene expression changed by the alternative SNP allele count relative to a reference allele (https://gtexportal.org/home/faq). For example, when NES = 1.26 and the normalised expression = 1, the target gene expression increases 26% when the SNP allele count = 1. The Genotype-Tissue Expression project (GTEx) analysed 15,201 RNA-sequencing samples of 49 different post-mortem tissues collected from 838 individual donors in the USA with a variety of ethnicities including (85.3%) European, (12.3%) African, (1.4%) Asian, and (1.9%) Hispanic and Latino[136]. GTEx found that GWAS SNPs were significantly enriched with *trans*- and *cis*-eQTLs across tissues[136]. Moreover, the *trans*-eQTL gene regulation was highly tissue specific[136]. The results illustrate that GWAS SNPs act through trans and cis gene regulation within various tissues to impact complex diseases.

Using the information on genomic organisations captured by Hi-C experiments[131,132,137], disease-related SNPs can be mapped to tissue-specific physical proximity with genes, modulating expression[133]. Thus, we can postulate that a subset of GWAS SNPs dysregulates gene functions in different tissues and organs via disruption of the physical proximity between the GWAS SNP and a disease gene. The Contextualising Developmental SNPs in three-dimensions (CoDeS3D) algorithm[133] maps DNA variants, associated with disease by GWAS, to their tissue-specific regulated genes. CoDeS3D takes a set of SNPs and uses information on the genomic organisation, captured by Hi-C in various cell lines and tissues, to identify the *trans* and *cis* regulated eGenes using tissue-specific expression data from the GTEx study[133,138]. The output is a dataset with tissue-specific NES information modulated by the disease-related variants. By applying the NES information to individuals in a case and control cohort, machine learning methods can evaluate the associations of the GWAS SNPs to eQTL effects to model complex disease risk.

1.9 Insufficient research for individual disease-associated tissue-specific risks

Although PRS and machine learning approaches have been extensively used in complex disease prediction, little attention has been given to the utility of machine learning applications in calculating tissue-specific disease risk in individuals. This is largely because GWAS studies identify relationships between SNPs and their associated phenotypes without attributing those associations to causative underlying molecular mechanisms[40]. However, GWAS-identified SNPs are likely to be modifying regulatory mechanisms which affect gene expression in a tissue-specific manner[127,139]. Therefore, by expanding GWAS methodology to include expression measures (*i.e.* eQTLs), genetic analyses could help to interrogate the inter-related biological networks between cell and tissue types that propagate the causal effects to complex diseases[127,140].

For example, incorporating eQTL data led to the identification of adipose-specific gene expression patterns that could have an inferred causal role in obesity[141]. Similarly, genes with liver-specific expression are now thought to be a major contributor to Type 2 diabetes (T2D)[142]. By extending eQTL analyses to include chromatin spatial interaction (Hi-C) data, it was shown that T2D and obesity associated SNPs have spatial-eQTLs which implicate dysfunction of specific regulatory actions in various tissue types[133]. These studies strongly suggest that by aggregating biological data types (*e.g.* DNA, RNA, and epigenetic data), the accumulated result becomes a tissue-specific network analysis of associated dysfunctionally regulated genes. Thus, specific disease risk to individuals should be calculated using a tissue-by-tissue approach, concluding with tissue-specific networks and pathways that are particular to the development of a disease.

In so doing, it may be possible to leverage the tissue-effect heterogeneity of patients by identifying the correct genes and tissue loads to provide essential targets for potential therapeutic interventions leading to enhanced therapeutic effectiveness. The tissue-effect heterogeneity could also help to recognise individual subtypes of complex disease, facilitating personalised treatments. By targeting the causal associated SNP tissue-specific effects, predictions of patient specific tissue-effect disease risks could provide informative biomarkers for early disease prevention, bringing about a substantial reduction of later disease burdens and costs. Zhou and Troyanskaya have utilised the machine learning algorithm to predict the functional effects of non-coding variants by modelling the pattern of genomic and chromatin profiling information[82]. They have been able to employ this method to distinguish important eQTLs and disease-related SNPs from various eQTL and SNP databases. Nevertheless, despite the immense promise of machine learning, it is important to recognise that at present, there is insufficient research in their application for the identification of disease-associated tissue-specific risks.

1.10 LDSC and Mendelian Randomization

LDSC (linkage disequilibrium score regression) is a popular method[143,144] developed by Finucane *et al.*[145] to quantify the genetic risk enrichment of tissue types related to complex diseases from GWAS SNP summary statistics data[145]. The basic assumption is that the effect size attributed to each GWAS SNP results from a number of tagged SNPs with various linkage disequilibrium (LD) strengths[145,146], calculated as LD scores. In terms of LDSC for quantifying tissue contributions, the LD scores of GWAS SNPs are calculated by the sum of r^2 (correlation) of the SNPs within the gene region plus 100kb surrounding each set of the tissue specific genes[145,146]. Using gene expression data derived from various tissues, a set of genes expressed significantly in each tissue is identified, and the risk enrichment of each tissue is measured by the regression of the LD scores of GWAS SNPs with their effect sizes[145,146]. LDSC is a useful tool to implicate related tissues and their genetic effects with complex diseases. However, this method cannot reveal the risk contribution of each genetic element in the related tissues.

The GTEx project has shown a disease related SNP can have multiple eQTL effects across tissues[136]. However, identifying the disease causal associations of eQTLs is not straightforward[147]. Mendelian Randomization (MR) can be applied to validate the causal effects[138]. MR is a statistical method to assess the causal impacts of genetic influenced features to diseases[138]. In my research context, the genetic element is a SNP and the influenced feature is a tissue specific eQTL effect. MR can use the SNP and eQTL effect association in addition with the SNP and disease-outcome association to validate the causal impact from a SNP modulated eQTL effect to a disease of interest[138]. However, MR assumes a disease related SNP has no other effects directly influencing the disease except through the eQTL effects and no direct effects to confounders that impact the target gene expression[138]. In other words, the disease related SNP only affects the disease through modulating the target gene expression. Under the MR assumptions, the disease causal association of the eQTL effect (target gene expression) can be validated statistically by the SNP association with the target gene expression and the SNP association with the disease[138]. Nevertheless, MR assumptions are not easy to confirm. MR is a method of choice for recognising a group of the disease causal eQTL effects[138,147,148], but it is not designed to evaluate the risk contributions of eQTL effects and distinguish the major eQTL effect risk contributors.

Although LDSC and MR are two useful methods for understanding disease-associated tissue-specific risks, they both have their own limitations. Better approaches are required to elucidate a clear view of the tissue specific eQTL regulatory networks for mediating complex diseases.

1.11 Summary

GWAS has identified hundreds of SNPs for each complex disease which are playing a crucial role in diseases aetiology. The majority of these disease-associated SNPs are enriched in tissue-specific gene regulation. Applying this information to an individual's SNPs to predict disease risk is an essential element for delivering the fuller promise of precision medicine. PRS is a straightforward model for assigning genetic risk to individual outcomes but has achieved only limited success in complex disease prediction due to its model limitations. The PRS method is ineffective in modelling highly dimensional genotype data with complex interactions. By contrast, the strength of machine learning data modelling in complex disease prediction lies in its handling of interactive high-dimensional data. Coupled with large new population datasets with high-quality phenotyping at different stages in the life course, machine learning models can classify individual disease risks with higher precision. Notably, machine learning predictors that include tissue-specific disease risks for individuals show even greater promise of insights that could and ultimately provide cost-effective and proactive healthcare with great efficacy.

1.12 Hypothesis and the aim of my research

GWAS identified disease-related genetic variants are successfully utilised in predictive models to assess the individual risk of various different disorders[57,98,115,149,150]. Nevertheless, the mechanism of those genetic variants impacting disease development is still unclear. The interpretation of thousands of disease-associated SNPs and their associated tissue-specific gene expression modifications is a complex challenge but necessary to elucidate how genetic variation impacts on complex diseases. My hypothesis is that machine learning models can identify the generalised patterns and associations between disease-associated SNPs and tissue-specific eQTL effects and in so doing implicate the essential tissues and genetic elements that contribute to individual disease risk.

The aim of my research is to develop a novel computational approach that can reveal the tissue-specific eQTL elements and their contributions to the risk of developing T1D and PD.

The aim will be achieved through the following objectives:

1. To apply the CoDeS3D algorithm with Hi-C captured SNP-gene interactions and GTEx NES information to integrate complex disease SNPs and their related tissue-specific gene regulation information.
2. To integrate tissue-specific gene regulation information with genotype data from large case and control cohorts.
3. To apply machine-learning computational methods for building credible predictor models to predict individual disease risk from analyzing the integrated data.
4. To utilize the predictor models to explicate the risk effects of the disease-associated SNP related gene regulation in tissues and organs.

The results of my computational approach will be mainly drawn based on statistical associations. They will require further experiments to validate the causality, which are beyond the scope of this research. On the other hand, the assumption of my approach is that the selected important genetic elements, by my predictor models for contributing to the disease risk prediction, also have essential roles in promoting and contributing to the development of the disease of interest.

Chapter 2: Procedures for Reproducibility

2.1 Introduction

Reproducibility of research is the ability for other independent scientists to obtain the same results by repeating the research experiments and data analyses[151,152] Non-reproducible results have become a significant and widespread concern in different scientific research areas[151,152]. Although many reasons and factors are contributing to the problem[151,152], it is essential and critical for a study design to address this issue from the beginning to protect the credibility of the results.

Data security and code management is the foundation of reproducible and reliable data analyses[153,154]. Without appropriate procedures in place to protect the data and code used in research, valuable research results could be lost or become non-reproducible by simple and accidental errors[153,154]. Hence, data security and code management are essential factors to ensure the reproducibility of my studies in future examinations of the results and hypotheses that are generated from this work. Two key processes were implemented to protect the integrity of data and programming codes in this thesis.

2.2 Data security

All the data used in this research was stored and used in the Nectar Cloud (https://nectar.org.au/cloudpage/) hosted by The National eResearch Collaboration Tools and Resources project (Australia). The Nectar Cloud provides a secure and comprehensive computational infrastructure for effectively managing research data[153]. The Nectar Cloud platform is supported locally by The Centre for eResearch, a cross-faculty research centre at the University of Auckland. Access to the research data was key protected by two-factor authentication[153].

Datasets that were received or downloaded from original sources (e.g. Wellcome Trust Case and Control Consorium[70], and UK Bio bank[155]) were individually maintained in read-only and write-protected directories[153]. The original data were only copied out for data analyses to avoid accidental alterations[153]. Intermediate result data were co-located in directories with the source code to preserve the steps of the data changes.

2.3 Programming code management

Programming codes used in the research were grouped and placed in directories organised according to the sequential steps of the various analyses with clear step description labelling[153,154]. Directories of the individual analysis steps were self-contained with subdirectories to store the required data, code, generated transient data, and results. The self-contained directory arrangement became a form of self-explained documentation of the research analyses and provided a convenient platform to repeat and validate analyses at different stages. Moreover, this structure also enabled the easy detection of errors and redeployment of analyses on other datasets. A Readme.txt file was included in each analysis step directory to describe the requirements, code functions and results of the step module operations[153].

Program code was preserved after producing validated results[153]. The code was named with appropriate functional and step sequential information[153]. This practice created the programming script descriptions of their roles in each analysis (for example, https://github.com/Genome3d/T1D_logistic_lasso_predictor). Each script included clear documentation of data and software requirements, script functionalities, and expected results[153]. Version control (git) was also employed to protect script integrity across the analysis step directories[153,154].

2.4 Conclusions

Reliability and repeatability are essential to protect the validity of scientific studies and were achieved in my research by ensuring the quality of data and code used in each different analysis. The research data were secured with a good IT platform and write-protected directories. The quality of scripts was guaranteed through the procedures of effective programming code management with proper documentation and the self-contained step directories.

Chapter 3: Methods

3.1 Machine learning modelling, feature selection and model validation in my research

Machine learning employs statistical and computational algorithms to create models based on training data information and validate their generalisation power by making good predictions in novel (validation or test) datasets[67,76].

Supervised machine learning disease predictors can be trained by pre-set learning algorithms to describe the relationships between sample genotype data and the disease of interest[67,76,156]. From the machine learning perspective, a feature describes a specific type of information (*e.g.*, specific SNP), which is encoded as a column within a feature matrix. Each row of the feature matrix (the so-called feature vector) summarizes all available information for a specific case/control genotype. The predictors obtain the optimal predictive power by selecting informative features (variables) within the training genotype data[67]. Supervised machine learning distinguishes classification (estimation of a categorical variable, *e.g.*, has T1D / does not have T1D) or regression (estimation of a continuous variable, *e.g.*, time since first symptoms of T1D)[76,79]. Logistic regression is a classification algorithm, which is based on generalised linear regression (basically the output of the linear regression is mapped to a value between 0 and 1 using the logit-function)[157].

3.1.1 Data feature selection

Data feature selection is a major factor affecting model predictive performance as it removes irrelevant data during the model training[67,89] Data feature selection also helps to create unbiased and non-overfitted predictors [67,89]. Overfitting is a phenomenon whereby the predictor is so closely fitted to the training dataset that it lacks generalisable power to other independent data sets[67,89]. Feature selection algorithms, such as Mann Whitley U test[158] in combination with control of false discovery rate (FDR)[159] and regularisation[90], can avoid model overfitting and enhance the generalised predictive power of the optimised model[67,89].

3.1.1.1 Mann Whitley U test

For logistic regression modelling, Mann Whitley U test[158] is a non-parametric method used to evaluate the difference between the distributions of an independent variable with respect to two groups (case and control phenotype). The test does not assume the normality of the associated independent feature. Mann-Whitney U tests the null hypothesis that the data feature distributions with respect to the two groups (cases or controls) are not different and the feature is not relevant for predicting the target (disease). It is a powerful tool to evaluate the association between an independent feature with the dependent phenotype (categorical variable). In combination with a procedure to control the false discovery rate for multiple association testing (e.g. Benjamini Yekutieli)[159], it can effectively remove non-informative data columns (features) from the training dataset.

3.1.1.2 Regression-based machine learning regularisation

The most common types of regression-based machine learning regularisation are L1 (lasso), L2, and Elastic net. Well established regression-based machine learning regularisation approaches use gradient descent procedures with iterative steps of model weight estimation for optimised predictive power [79,80]. L1 and L2 regularisations both use a penalised loss function to assign weights that adjust data feature effects and reduce the complexity of the regression models[67,76,89]. L1 regularisation sets the weights of non-informative data features to zero, thus eliminating effects and allowing only essential and valuable data feature effects to be included in the machine learning regression modelling[67,89,160]. By contrast, L2 regularisation minimises non-essential data features using non-zero weights[67,76,89]. The Elastic net adopts a regularisation approach in-between L1 and L2, which removes non-informative features but retains some redundant valuable information in the modelling[67,89,90].

3.1.2 Predictor model-validation

Cross-validation is commonly used for model performance optimisation during the training process[76,79]. Cross-validation is performed by partitioning the dataset into N equal parts (N fold)[161]. It withholds one part of the data for model testing and uses the other N - 1 parts for training[161]. This process is repeated until every single part of the data is tested with a model. The cross-validation result is the mean predictive performance of the N time testings[161]. Once the optimal model has been selected, the performance of the optimal predictive machine learning model is evaluated to estimate its power for prediction across novel (independent) data[57,67,89]. External validation (testing) using an independent dataset is essential to ensure that the prediction models are not overfitted with the training data. External validation also confirms the generalised predictive power of the optimal machine learning model[57,67,89].

I have developed regularised logistic regression predictors in my research which incorporated: 1) Univariate feature selection for removing irrelevant information; and 2) a multivariable prediction step that considers all features in context and removes redundant information. This approach allowed me to identify the best combination of features for the prediction of complex diseases (Type 1 Diabetes and Parkinson's disease). Regularised logistic regression was incorporated into the models to enable the features that contribute to the final score to be identifiable. The developed predictors were intensively validated by cross-validation and external validation.

3.2 My methodological approach

In this thesis, I established and validated credible independent predictor models for 1) Type 1 Diabetes (T1D) and 2) Parkinson's Disease (PD). For each disease predictor model, I employed supervised machine learning in the form of logistic regression[76] with Mann Whitney U test feature selection[158] and regression-based regularisation[67,89]. These algorithms analysed the tissue-specific gene expression information by mathematically estimating the complex associations of related SNPs and tissue-specific eQTL effects to complex disease phenotype binary labels (e.g. case or control) using individual genotype data[89].

The method I developed follows the basic algorithm:

1. Select disease-related variants, SNPs from GWAS studies
2. Use Hi-C data[132] to map the disease-related SNPs to their tissue-specific eQTL regulated genes and calculate the NES effect sizes[127] using CoDeS3D[162]
3. Integrate the tissue-specific gene eQTL information obtained from CoDeS3D[162] as gene regulation weights (the NES effect sizes[127]) on case and control individual genotype data, *e.g.*, Welcome Trust Case and Control Consortium (WTCCC)[70]
4. Use Mann Whitney U test[158] hypothesis tests in combination with Benjamini Yekutieli (BY)[163] procedure to control the false discovery rate (FDR)[158,159] and select informative data features
5. Create predictor models using logistic regression implemented with Elastic net regularisation[67,89,90] and the integrated eQTL data as the training dataset
6. Use repeated internal cross-validations to evaluate and identify the optimised model parameters (hyperparameters)[161]
7. Create the final predictor model with the optimised hyperparameters from the full set of the integrated tissue-specific gene eQTL data for accurate estimations of data feature weights in the regression model
8. Establish the model performance by evaluating and validating the final predictor model using the tissue-specific gene eQTL information integrated with independent case and control individual genotype (testing) cohort (e.g. UK Biobank)[155]
9. Analyse the final disease predictor model to reveal the tissues and their related elements that significantly contribute to the disorder.

Please refer to chapters 4 and 5 for full descriptions of the methods that were developed specifically for the T1D and PD analyses. In the following sections of chapter 3, I will concentrate on explaining the differences in the predictors and the critical choices I made in their development.

3.3 Considerations for data integration in the T1D and PD studies

3.3.1 Hi-C data

Hi-C data[132] were used by CoDeS3D[162] to detect SNP-gene interactions for the SNPs that were associated with T1D and PD studies. The Hi-C data were obtained from different published cell libraries (see Methods of Chapters 4 and 5) generated from a variety of immortalized cell lines and primary tissues. The primary tissue cells covered a range of tissues and captured an extensive range of potential *trans* and *cis* SNP-gene interactions.

3.3.2 Tissue-specific eQTL data

GTEx tissue-specific eQTL data were used in the two studies. The GTEx database contains the largest variety of tissues with their eQTL effect information[126,136]. GTEx collected post-mortem tissue samples from donors aged 21 to 70 with different ethnicities[126,136]. The GTEx tissue-specific eQTL data enabled me to examine the genetic disease impacts acting through multiple tissues and reveal the putative tissue contributions to complex diseases, which was an important feature of my research.

Because of the timing of my PhD, I could only use GTEx v7 data[126] from 635 donors in the T1D study. GTEx v7 included 44 different tissues in the T1D study. Nevertheless, I utilized the GTEx v8 data[136] from 948 donors (33% increase) in the PD study. This enabled me to extend the number of tissues to 49 in the PD study and thus increased the power of my predictors to identify the disease-related tissue-specific effects.

3.3.3 Choice of SNPs

The T1D and PD associated SNPs were chosen from a range of GWAS studies[19,52,164–171] with a more relaxed threshold[42,43] (association p-values $< 10^{-5}$) that allows incorporation of SNPs associated with the disorders at a suggestive level of significance.

3.3.4 WTCCC case and control genotype data for creating model training datasets

Wellcome Trust Case and Control Consortium (WTCCC) was designated for using GWAS to interrogate the full human genetic impacts on different diseases with considerable cohort sizes in the UK[50,70]. The WTCCC case and control genotype datasets[50,70], for T1D and PD, were acquired and combined with the disease-associated tissue-specific eQTL information to create predictor model training datasets in the T1D and PD studies. These cohorts were obtained because of their cohort sizes (n >= 5000).

Due to the design of the microarray genotyping, the WTCCC datasets only contained a limited number of SNPs (~500k) and were designed to cover the whole genome through the use of LD blocks[172]. The majority of the T1D and PD SNPs of interest in my studies were not present in the un-imputed WTCCC datasets. Genotype imputation is a cost-effective and reliable way to discover a wider range of SNPs from microarray variants[173].Therefore, genotypes were imputed to: T1D, 16,722,059 SNPs (399%); and PD, 27,590,399 SNPs (524%) SNPs using Sanger imputation service (https://imputation.sanger.ac.uk) with EAGLE+PBWT pipeline[174,175] using Haplotype Reference Consortium(r1.1)[176] especially for European genotypes. The imputed T1D genotype data contained 253 of the 313 T1D SNPs with 60 missing SNPs. And, the imputed PD data included 281 of the 290 PD SNPs with 9 missing SNPs.

3.3.5 Choice of human reference genome sequence

Prior to imputation, the WTCCC datasets were cleaned and prepared according to the instructions listed in (https://imputation.sanger.ac.uk/?instructions=1). This cleaning required all the preprocessed genotype data in GRCh37 genomic coordinates. The imputed genotype data from the Sanger service also used GRCh37 genomic coordinates. Hence, GRCh37 coordinates were adopted as the standard in the analyses of both the T1D and PD studies.

3.3.6 UKBiobank and NeuroX-dbGap for independent predictor model validation datasets

The UK Biobank is a database collected from half a million participants for comprehensive biomedical information, including genotype, phenotype, lifestyle and medical history[155]. UK Biobank phenotype and imputed genotype data were obtained for generating test datasets to validate the final predictor models (T1D model-2 and PD model-1) in the T1D and PD studies. Using the medical records, I selected the case and control individual participants for the examined diseases using specific medical criteria (see methods of chapter 4 and 5). I then used their genotype data to create the independent validation datasets.

Due to the low prevalence rate of T1D and PD and the genotype quality issues in the UK Biobank individual samples, limited numbers of cases were found in both of the T1D (415 cases) and PD (928 cases) studies. Thus, I created 30 sets of test cohorts with the same set of cases and 30 different sets of randomly chosen controls for making test datasets to validate the final predictor models in both studies. In addition, the 30 AUC validation results of the UK Biobank derived test data were also utilized to reveal the range of the AUC test performance variations of the predictor models [177].

NeuroX-dbGap was the largest PD single array study with 5,353 cases and 5,551 controls[33,51], and the NeuroX-dbGap genotype data was used to create a second independent test dataset in the PD study. The microarrays for generating the NeuroX-dbGap case and control genotype data were specially designed to assay only the genomic regions highly associated with PD, which were not suitable for genotype imputation[33,51,178]. Hence, the SNPs of interest which were not included in the NeuroX-dbGap data were replaced with proxy SNPs using linkage disequilibrium information ($r^2 > 0.5$)[51,178]

3.3.7 NES eQTL effect sizes

In the PD study, CoDeS3D-v2 used the GTEx v8 data to map the PD associated SNPs to their tissue-specific effects with two options (NES and aFC) for the eQTL effect size calculations[136,162]. NES is the normalized linear ratio of the tissue-specific gene expression changed by the alternative SNP allele count relative to a reference allele (https://gtexportal.org/home/faq)[127]. aFC is the log ratio of the haplotype expression with an alternative allele (SNP) to the haplotype expression with a reference allele. (https://gtexportal.org/home/faq)[179]. aFC effect sizes are specially optimized for SNP allele count[179] = 1 which may not be ideal for my logistic regression predictors models that account for SNP allele count = 0, 1 or 2. Hence, NES values[127] were used in the PD data integration.

3.3.8 SNPs with unknown eQTL effects in the predictor models

In order to enhance the predictive power and account for the SNP effects that modulate disease risk by non-tissue-specific eQTL pathways[180,181] in the regularized predictor approach, the disease associated SNPs presented in the imputed genotype that did not have detectable tissue-specific eQTL effects were retained in the integrated eQTL data. These SNPs represented the genetic disease impacts that occurred through non tissue-specific gene regulatory mechanisms.

3.4 Considerations for the data modelling and differences between T1D and PD analysis

3.4.1 Python programming language

Python is a high-level, easy to use, open-source and powerful computer language with extensive libraries and packages to support advanced machine learning programming development[182]. Moreover, it is one of the most popular computer languages in machine learning[182]. Thus, Python was chosen as the primary computational language to provide robust and concise programming for data processing and machine learning predictor modelling in both the T1D and PD studies.

3.4.2 The Sci-kit learn machine learn package

The Scikit learn machine learning package[183,184] for Python was used as the machine learning programming platform, for the regularized logistic regression predictor modelling, in the T1D and PD studies. Scikit learn is a freely available python package that has many useful modules and libraries to support advanced machine learning developments[183,184].

LogisticRegression implemented with Elastic net regularization[67,89,90] was the primary Scikit learn module[183,184] used in both studies for developing disease status predictors with an option l1_ratio to control the regularization strength. The range of the l1_ratio can be from 0 to 1. When l1_ratio = 0, the regularization is equivalent to L2 regularization[67,76,89]. With l1_ratio = 1, the regularization is in maximum strength that is equal to L1 regularization[67,89,160]. During the predictor model development, I intentionally did not build predictors with l1_ratio = 0 in order to ensure the developed predictor models use regularization that can remove non-informative features. The option C of LogisticRegression is another parameter for adjusting the inverse of regularization strength to minimise model parameter values. The values of C must be positive and have stronger regularization closer to zero[183,184].

Another option, solver of LogisticRegression is used to select an optimization algorithm for supporting regularization[184]. In both studies, solver was set to 'saga' for using the saga algorithm, a kind of gradient descent procedure and the only optimization algorithm supporting Elastic net regularization[185] available for LogisticRegression. The max_iter option is to specific the maximum number of iterations for the solver to coverage[183,184].

GridSearchCV was used for model optimization by identifying the best set of the model hyperparameter values for optimum predictive power. The algorithm tested and evaluated the assigned predictor model performance with different combinations of given hyperparameter values exhaustively using cross-validation[183,184].

3.4.3 AUC model performance measurement

The widely popular predictive measurement AUC[59,60,63,64] was adopted for evaluating the performance of the developed predictors. The use of AUC makes performance comparisons with other genetic disease status predictor models from different studies[19,124] much simpler.

AUC is an effective measurement to evaluate the predictive power of logistic regression predictors[157]. that produce continuous predictive values between 0 and 1. The predictive power of logistic regression predictors can be varied depending on the predictive value cut off points by dichotomous classification (case /control) measurements. AUC predictive measurements are calculated by ROC curves where the sensitivity and specificity of the predictions of a predictor are evaluated at many cut-off predictive values[59,60,64]. Hence, AUC could fairly and accurately reflect the predictive power of the regularized logistic regression predictors developed in both the T1D and PD studies.

3.4.4 Controlling Type 1 errors

In my research, statistical testing was used to assess the associations of elements of interest. Situations arose and multiple statistical comparisons were conducted in analyses that greatly promoted Type 1 errors (falsely discovered as positive results). False discovery rate (FDR) controlling procedures, Benjamini–Hochberg (BH) and Benjamini-Yekutieli, (BY) were adopted and implemented in the software applications used in the predictor modelling analyses for minimizing Type 1 error rates[159,163]. Compared to familywise error rate controlling procedures (*e.g*. Bonferroni), FDR controlling adjustments are less conversative with balanced controlling of Type 1 and Type 2 (falsely discovered as negative results) errors[163]. With the independent test statistics assumption, Benjamini–Hochberg is a procedure using unadjusted statistical p-value ranking to control the expected percentage of the falsely discovered positive results (the number of false positive results/ the total number of positive results) within a given FDR threshold[163]. Benjamini-Yekutieli, is an improved version of BH procedure that controls FDR of multiple statistical testing for correlated data with dependent test statistics by resampling based method[159].

3.4.5 The tsfresh package

tsfresh is a python package that contains modules libraries for time series analysis, and the package works smoothly with the Scikit learn machine learning package. The tsfreah package also includes feature selection modules that have a nice implementation of Mann Whitney U hypothesis tests in combination with the BY procedure for controlling FDR[158,159,163]. Thus, tsfresh was utilized to conduct Mann Whitney U test filtering with the BY control to remove non-informative features in the regularized logistic regression predictor modelling of the T1D and PD studies.

Because of a bug in tsfresh version 0.12.0 in the implementation of the Benjamini-Yekutieli procedure (https://github.com/blue-yonder/tsfresh/pull/570), used in the T1D study leading to overly conservative rejection of features, I adopted a relaxed FDR threshold of 0.2 in the Mann Whitney feature filtering. To avoid the conservative feature rejection problem, the tsfresh package was changed to version 0.16.0 in the PD study and a much stronger FDR filtering threshold = 0.05 was also adopted.

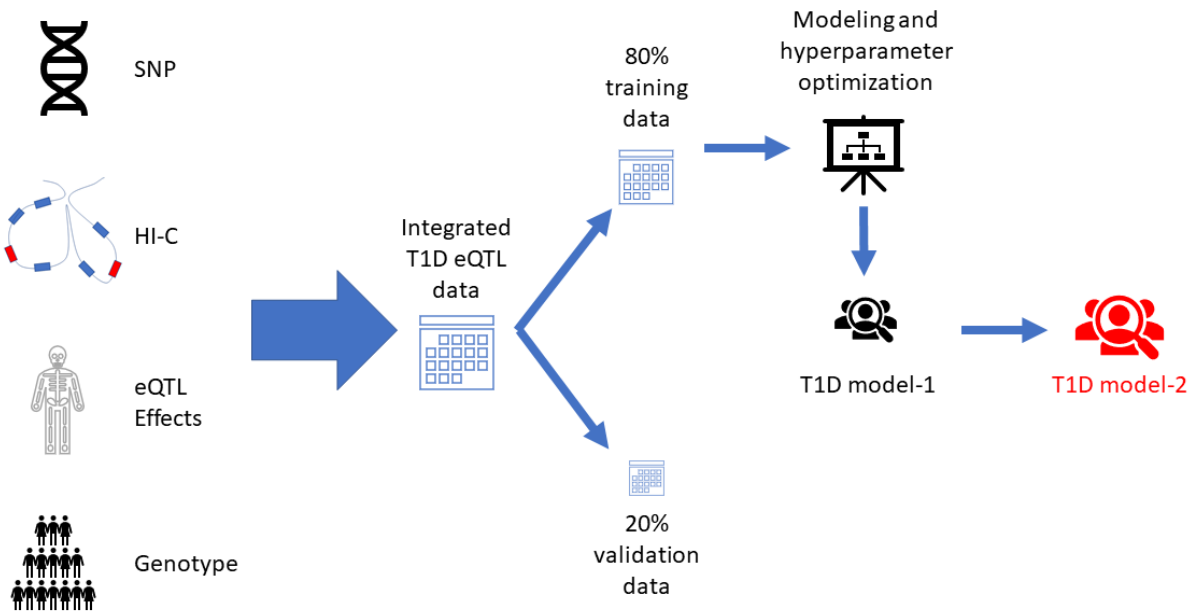


Figure 3-1:T1D model-2 development

3.4.6 The different approaches for creating the final predictor models

In the T1D data modelling (Figure 3-1), I split the integrated T1D eQTL data into two parts with 80:20 proportions to obtain an additional independent dataset for validating the predictive model performance during early model development. First, I created the T1D model-1 from the 80% training data after the identification of the optimized model hyperparameters. Then, I created the final T1D predictor (T1D model-2) with the optimized model hyperparameters from the full set of integrated T1D eQTL data. Following this, I validated model-2 with the test data derived from an independent genotype dataset, UK Biobank[155].

PD is a late-onset disease with a weaker genetic component in pathogenesis when compared to T1D. Therefore, I used the whole integrated PD eQTL data in model training to take advantage of the full embedded genetic information for enhancing the disease risk prediction power. Subsequently, I used two independent genotype datasets, UK Biobank[155] and NeuroX-dbGap[33,51,178] for the final PD predictor model (PD model-1) validation. The PD model-2 was built for comparison from the 90 SNPs of Nalls *et al.*[33] with the same procedures and validations to create PD model-1.

3.5 Summary

To understand the effects of disease-related genetic variants acting on different tissues, I developed a computational approach to integrate disease related SNPs and their tissue-specific gene regulatory effects with case and control individual genotype data. I employed the Mann Whitney U test[158] with the BY[163] control feature filtering to remove the non-informative data from the integrated eQTL data. Then, I utilized machine learning Elastic net regularization to select the essential tissue-specific eQTL effects for building logistic regression predictor models to predict disease status risk. The predictor models were extensively validated by cross-validation and independent case and control genotype data derived test datasets. The best predictor model was chosen to reveal the disease impacts of the related SNPs and their modulated gene regulatory effects acting through different tissues.

The computational approach was applied to study T1D and PD. WTCCC case and control genotype data for T1D and PD cohorts were utilized in the eQTL data integration to generate predictor model training data. The WTCCC datasets required imputation by the Sanger service to obtain the SNP genotype data of interest. UK Biobank and NeuroX-dbGap data were acquired for creating the independent test data for validating the final predictor models. The Scikit learn python package was chosen to support machine learning regularisation and logistic regression predictor building in both of the studies. The tsfreah package was used to implement the Mann Whitney U test with the BY control feature filtering. The predictive power of the developed predictors in both of the studies was measured by AUC for the measurement can fairly and precisely evaluate the prediction of logistic regression predictor models.

In the T1D predictor modelling, 80% of the integrated T1D eQTL data were used to create T1D model-1 after searching for the best optimized hyperparameters. The final predictor T1D model-2 was built with the best optimized hyperparameters from the full integrated T1D eQTL data. By contrast, in the PD study, the entire integrated PD eQTL data was utilized to identify the best optimized hyperparameters and then to create the final predictor model PD model-1 with the identified best optimized hypermeters.

Chapter 4: Machine learning identifies the lung as a susceptible site for allele specific regulatory changes associated with risk for type 1 diabetes

Computational methods are capable of providing unparalleled insights into potential mechanisms by which genetic variants regulate disease risk. As the aetiology of type 1 diabetes (T1D), an early onset complex disease, remains complex and poorly defined, I have applied my machine learning approach that integrated T1D case and control genotypes from the Wellcome Trust Case Control Consortium (WTCCC) and UK Biobank (UKBB) with tissue-specific expression quantitative trait loci (eQTL) data on T1D SNPs. The integrated data were selected and analyzed by regularized logistic regression predictors models. My results identify key genetic factors influencing the conversion of genetic factors into T1D risk and may help explain the reported association between respiratory infections and risk of islet autoantibody seroconversion reported in young children.

4.1 Introduction

Type 1 diabetes (T1D) is characterized by immune-mediated destruction of insulin-producing pancreatic beta cells leading to loss of insulin production and hyperglycemia. Population level data have enabled genome-wide association studies (GWAS) that have identified ~60 genetic loci that are associated with the risk of developing T1D[186]. In addition to the GWAS studies, a number of highly phenotyped prospective birth cohort studies have investigated potential early determinants of T1D risk[187–189]. Notably, the transition from genetic risk to T1D onset is hypothesized to require an environmental trigger event, such as infection, in those individuals who go on to develop the disorder[13]. However, the mechanisms responsible for this transition remain poorly characterized, limiting strategies for optimizing treatment and furthering therapeutic development.

One hindrance to characterizing the genetic mechanisms responsible for T1D development is the finding that the majority of SNPs are within intergenic regions of the genome. Previously, our lab used information on the spatial organization of the genome (captured by Hi-C) to identify the tissue-specific gene regulatory impacts (*i.e.* eQTLs) of SNPs associated with T1D[15]. Consistent with our understanding of T1D pathology, we reported that the differentially expressed genes were enriched for immune activation and response pathways[15]. However, we did not investigate which specific eQTLs were responsible for the conversion of risk to pathology.

In the present study, we assigned SNPs associated with T1D to the genes they modulate through Hi-C chromatin interactions captured from primary tissues (*i.e.* pancreas and spleen) and immortalized cells. We integrated a regularized logistic regression model on European ancestry genotypes of T1D case and control to identify transcriptional changes in the lung involving *AP4B1-AS1* and *CTLA4* (associated with rs6679677) as the greatest individual contributors to the conversion of the genetic risk for the development of T1D. Finally, a plasmid-based luciferase reporter expression assay was performed to validate the allele specific enhancer activity of the locus marked by rs6679677 in lung cells.

4.2 Methods

4.2.1 Identification of genetic variants associated with the development of T1D

In total, 313 genotyped and imputed SNPs associated with T1D, and those associated with time-to-event development of islet autoimmunity and T1D, were retrieved from the GWAS catalog (www.ebi.ac.uk/gwas, downloaded 8th February 2019; p-value < 1.0 x 10^{-5}), prospective studies[164–167], TrialNet PTP cohort[168], adult-onset[169], and GRS prediction studies[19,170,171] (Appendices: Supplementary Table 1). All genomic positions for SNPs and genes are annotated according to reference human genome hg19/GRChr37.

4.2.2 Identification of SNP-gene pairs and expression QTL associations in human tissues

The CoDeSS3D analysis was performed by Dr Denis Nayaga, in Justin O'Sullivan's lab group in late 2018

The Contextualizing Developmental SNPs in three-dimensions algorithm (CoDeS3D[162]) was used to identify genes that physically interact with loci marked by the T1D associated SNPs. Briefly, the CoDeS3D modular python scripts integrate Hi-C contact libraries from published sources (Appendices: Supplementary Table 2) to identify spatial co-localization of two DNA fragments, with one fragment marking the queried SNP. Gene-containing restricted fragments that are in physical contact with fragments containing the queried SNPs are identified as spatial pairs to the SNPs. Finally, the resultant spatial SNP-gene pairs are queried in the Genotype-Tissue Expression database (GTEx) to identify SNPs that are associated with transcript levels of genes through physical interaction at FDR < 0.05[162].

In the present study, spatial interactions were identified in Hi-C chromatin contact libraries captured from: 1) immortalized cell lines that represent the embryonic germ layers (*i.e*. HUVEC, NHEK, HeLa, HMEC, IMR90, KBM7, and K562); 2) B-cell derived lymphoblastoid cell line (GM12878); and 3) primary human tissues (*i.e*. spleen and pancreas) (Appendices: Supplementary Table 2). The identified putative spatial interactions aggregated from different Hi-C libraries were utilized to select tissue-specific eQTL effects. Of note, the summary statistics for the CoDeS3D run include the Hi-C cell line/tissue in which the interaction was observed (Appendices: Supplementary Table 3; column=cell lines). The regulatory potential of the spatial connections was identified by incorporating eQTL information from 44 human tissues (Genotype-Tissue Expression database [GTEx] v7; www.gtexportal.org)[126]. Spatial eQTLs were deemed significant and recorded if the $q < 0.05$ after correcting for multiple testing using the BH procedure[163]. Finally, genes whose transcript levels were associated with a spatial-eQTL were denoted as eGenes. The eQTL-eGene interactions were defined as either: *cis*, the eQTL and eGene are separated by a linear distance of <1Mb on the same chromosome; or *trans*, eQTLs and their eGenes were separated by >1Mb on the same chromosome or located on different chromosomes.

4.2.3 Genotype imputation for T1D cases and controls

Genotypes from T1D cases (n=2000) and controls (n=3000) were obtained from the Wellcome Trust Case Control Consortium **(**WTCCC)[70]. SNPs within individual genotypes were converted to rsIDs and genomic positions mapped (GRCh37, hg19). PLINK (v1.90b6.2, 64-bit) was used for quality control. Genotypes were cleaned using the Method-of-moments F coefficient estimate to remove homozygosity outliers (F values < -0.04 or 0.025 < F values). Related individuals were identified and removed using proportion IBD (PI_HAT > 0.08). Ancestry outliers (identified by principal component analysis [PCA] plotting), individuals with sex genotype errors (identified by PLINK), or individuals with missing genotype data (missing rate > 5%) were also removed. Finally, SNPs that were not in Hardy-Weinberg Equilibrium ($p < 10^{-6}$) or had a minor allele frequency < 1% were removed before SNP data imputation (Sanger imputation server; https://imputation.sanger.ac.uk)[176]. Following imputation, the T1D genotype data was cleaned to remove SNPs with an: impute2 score < 0.3; missing data rate > 5%; or a minor allele frequency < 1%.

4.2.4 Creation of a WTCCC genotype T1D-eQTL matrix

The machine learning only used SNPs (quantified or imputed) that did not have any missing data across the cohort and thus were present within each of the genotypes that were used. From the total 313 T1D SNPs included in the study, 253 SNPs were present within each of the WTCCC genotypes (Appendices: Supplementary Table 1). Of these 253 T1D SNPs, 224 had detectable eQTLs, connecting to 758 eGenes (6307 tissue-specific eQTL effects). The tissue-specific eQTL normalized effect size for each T1D-associated SNP within the imputed WTCCC genotypes was extracted from the GTEx eQTL summary table of significant eQTLs (Appendices: Supplementary Table 3). The normalized effect size for each tissue-specific eQTL was weighted by the number of alternative alleles at the eQTL SNP position in each individual's genome. The 30 T1D-associated SNPs that were not eQTLs were unweighted, using solely SNP allele count from the imputed genotype.

4.2.5 Generation, training and validation of the regularized logistic regression models

In order to identify the optimized predictor model parameters, the weighted WTCCC genotype T1D eQTL matrix was randomly split (80:20) into two groups that contained case and control genotype data for model training and validation. The Mann-Whitney U test[158] with the BY control[163] (tsfresh version 0.12.0 [190]) was used to select the individual feature columns within the 80% training dataset that were the most relevant attributes for predicting the T1D status (*i.e.* the relevant subset; FDR = 0.2) [191]. Note the large FDR of 0.2 was probably caused by a bug in tsfresh 0.12.0 in the implementation of the Benjamini-Yekutieli procedure https://github.com/blue-yonder/tsfresh/pull/570, which caused a much too conservative rejection of features. The relevant subset was then used to train a multiple logistic regression algorithm (Scikit-learn version 0.21.3; [183,184]) implemented with elastic net regularization using the SAGA solver to predict T1D disease status. The training was optimized using a Grid Search algorithm with 10-fold cross-validation to identify the best predictor with the optimised parameters and created T1D model-1. Hence, T1D model-1 was created from 80% of the data with the optimised parameters (hyperparameters: C=1, l1_ratio=1, max_iter=500, penalty='elasticnet', random_state=1, solver='saga') from the search of following:

- 'C':[1e-4,1e-3,1e-2,1e-1,0.2,0.3,0.4,0.5,0.6,0.7,0.8,1e0,3,5,7,8,10, 15,20,25,30,40,50, 100]'
- 'max_iter':[1,5,70,100,130,150,170,180,200, 300,500,1000,1200,1400,1600,1800,2000,2200,2400,2600,3000]
- 'l1_ratio':[1,0.9,0.8,0.7,0.6]

'C' is the inverse of regularization strength that must be positive and specify stronger strength with smaller values[183,184]. 'solver' is the algorithm used in the optimization. 'max_iter' is the maximum number of iterations for the solvers to converge[183,184]. 'penalty' is the penalty function[183,184]. 'l1_ratio' is the elastic net regularization strength[183,184]. When l1_ratio = 0, the regularization is equivalent to L2 regularization[183,184]. When l1_ratio = 1, the regularization is equivalent to L1 regularization[183,184].

Prediction performance (measured by area under the curve [AUC]) for T1D model-1 was tested using the relevant subset from within the 20% validation dataset. The optimal hyperparameter l1_ratio=1 effectively reduces the elastic net regularization to a lasso regularization. We used Elastic net regularization and found via hyperparameter optimization that the limit case of lasso regularization was the most performant.

To calculate a measure of the variation in AUCs of the modelling with the optimised parameters from the above model optimization; we undertook 10 repeats of 5-fold cross-validation of model generation and validation using the Scikit-learn RepeatedKFold algorithm[183,184] with the full weighted WTCCC genotype T1D-eQTL matrix, starting with the random generation of the 80:20 training:validation data sets and without Grid Search optimisation. This resulted in 50 T1D logistic regression predictors derived using the same general parameters as T1D model-1.

4.2.6 Calculation of tissue-specific contributions to T1D risk

The 50 T1D regularized logistic regression predictors created from the 10 repeats of 5 fold cross-validation were used to test the predictive power of tissue-specific eQTL effects on individual genotype risk scores. Tissue-specific contributions to the T1D risk were extracted from each predictor as the sum of the absolute values of the weights associated with each tissue.

4.2.7 Validation of the importance of the lung eQTLs in UK Biobank data (T1D model-2)

A second model T1D model-2 (Figure4-1) was created and trained using the full WTCCC training dataset with the optimised parameters. This model did not use the 80:20 split that was used in T1D model-1.

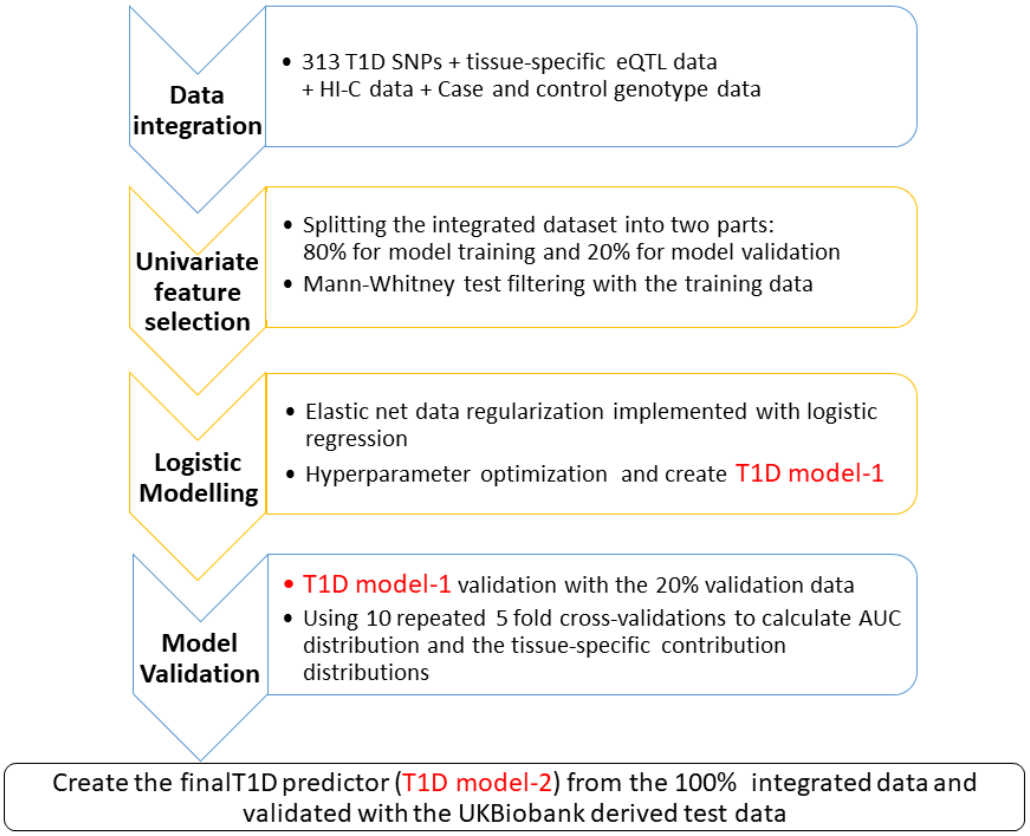


Figure 4-1: T1D model-2 development workflow

T1D model-1 was then validated using 30 cohorts of 993 individual samples (415 cases and 578 controls) derived from the UK Biobank. The 415 cases were selected, using a modification of Sharp *et al.*[19], from the UK Biobank imputed (487,411 individual samples) BGEN format dataset using the following criteria:

1. European Caucasian by genetic clustering methods
2. Clinical diagnosis of diabetes at ≤ 20 years of age
3. On insulin within one year from the time of diagnosis
4. Still on insulin at the time of recruitment
5. Never self-report as having type 2 diabetes (T2D)
6. All SNPs included in the model 2 predictor were present within each individual imputed genotype data

The 578 control individual samples, without missing data for any of the SNPs included in the model 2 predictor, were randomly selected from the healthy controls within the UK Biobank data for each of the 30 test datasets. The genotype data for the 993 case and control UK Biobank samples in each test dataset was used to build a weighted eQTL-genotype matrix as outlined for model 1.

Table 4-1: Primer sequences used for plasmid DNA amplification and Sanger sequencing

Primer	Sequence
RVprimer3 (forward primer)	CTAGCAAAATAGGCTGTCCC
EBV-rev (reverse primer)	GTGGTTTGTCCAAACTCATC
Luciferase (forward primer)	GAGATCGTGGACTATGTGGC
MPRA_SfiI_F (forward primer)	GCTAAGGGCCTAACTGGCCGCTTCACTG
MPRA_SfiI_R (reverse primer)	GTTTAAGGCCTCCGTGGCCGACGCTCTTC

4.2.8 Reporter assay for validating the regulatory effects of genetic sequences

The Reporter assay analysis was performed by Dr Denis Nayaga, in Justin O'Sullivan's lab group in early 2020

Luciferase reporter assays (Figure 4-3) were performed using a modification of [192]. Briefly, DNA sequences flanking rs6679677 (*i.e*. 74bp 5' – ref/alt allele – 75bp 3' [chr1:114303734-114303884; GRCh37]) containing the reference and alternative sequences) were synthesized by Integrated DNA Technologies (IDT). To ensure compatibility with the pMPRA vectors (pMPRA1 [Addgene: plasmid #49349] and pMPRAdonor2 [Addgene: plasmid #49353]), each sequence was designed using the following template: 5'-ACTGGCCGCTTCACTG-var-GGTACCTCTAGAAGATCGGAAGAGCGTCG-3' (*i.e*. var denotes the 150bp sequence to be assayed) (Figure 4-2). The variable region (var) was separated by a pair of *Kpn*I (GGTACC) and *Xba*I (TCTAGA) restriction sites to enable directional insertion of a reporter gene. PCR amplification was performed using primer sequences (MPRA_SfiI_F [forward], MPRA_SfiI_R [reverse]; Table 4-1) to add two distinct *Sfi*I (GGCCNNNNNGGCC) tails to enable directional ligation of the oligonucleotide into pMPRA1 (Figure 4-2). An aliquot of the amplification product was electrophoresed (2% agarose, 100V, 45 min) to visualize and verify that the product was the correct size (~200bp) and that there were no non-specific amplification products. PCR amplicons were digested with *Sfi*I (50°C, 2 hours), purified (QIAquick PCR Purification Kit; Qiagen), and quantified by Nanodrop.

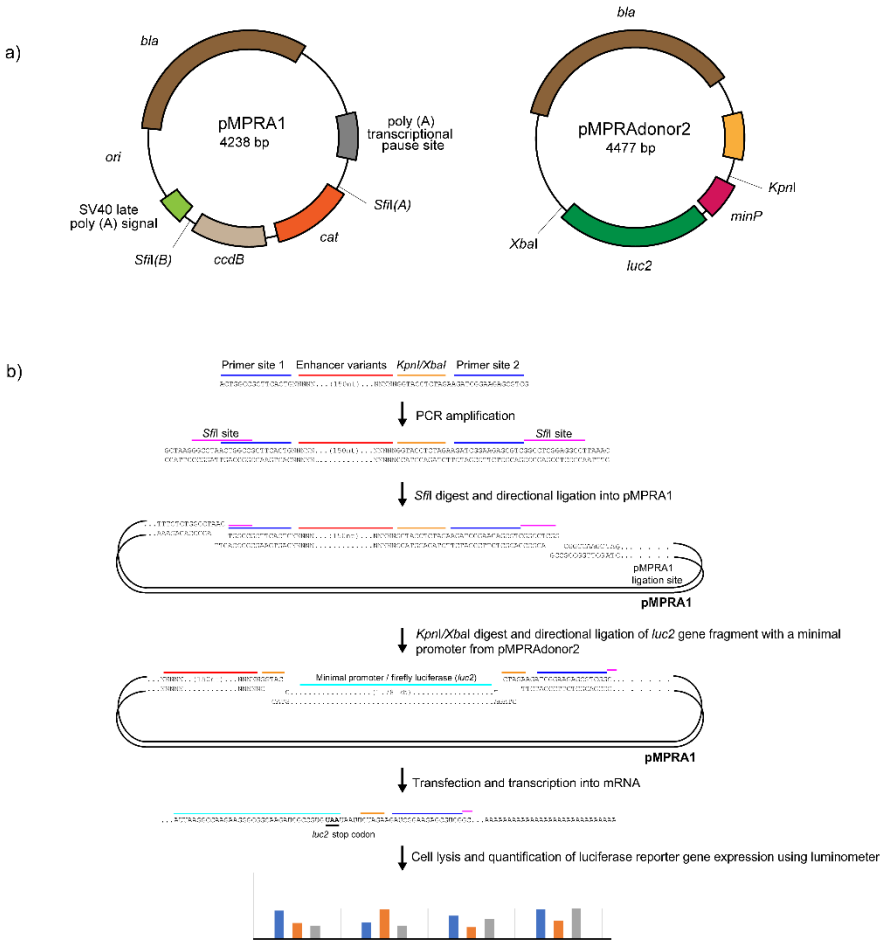


Figure 4-2: A flow chart of the plasmid-based reporter assay methodology

(a) Simplified maps of pMPRA1 and pMPRAdonor2 plasmid vectors. The pMPRA1 plasmid contains two *Sfi*I restriction sites for directional ligation of the oligonucleotide sequences. The pMPRAdonor2 contains *Kpn*I and *Xba*I restriction sites to facilitate directional ligation of the luciferase (*luc2*) reporter gene fragment, together with the minimal TATA-box promoter (minP) within the oligonucleotide sequences. **(b)** Oligonucleotide sequences used in the reporter assay are synthesised DNA sequences flanking the T1D SNPs (*i.e.* reference and alternative sequences) and containing a pair of *Kpn*I (GGTACC) and *Xba*I (TCTAGA) restriction sites for directional ligation of the reporter gene fragment. PCR amplification is performed to add two distinct *Sfi*I tails for directional ligation of the oligonucleotide sequences into the pMPRA1 backbone. Transformation of the plasmid containing the oligonucleotide sequences is performed in competent *E. coli* cells. The plasmid DNA is extracted from *E. coli* following successful transformation (*i.e.* presence of colonies). The extracted plasmid DNA is digested with *Kpn*I and *Xba*I, and the luciferase gene (*luc2*) fragment is ligated within the sequence. Transformation of the plasmid is again performed using competent *E. coli* cells. Following a successful transformation, plasmid DNA is extracted and sequenced to confirm the absence of indels. The plasmid DNA is transiently transfected into A549 and HepG2 cells, and luciferase activity is assessed after 48 hrs in a luminometer. This assay is a modification of [192]. Figure part **(a)** is modified from https://www.addgene.org/; and part **(b)** is redrawn from[193]. pMPRA1 – Addgene: plasmid #49349. pMPRAdonor2 – Addgene: plasmid #49353.

pMPRA1 was linearized by digesting with *Sfi*I (50°C, 2 hours), electrophoresed (0.8% agarose, 60V, 1 hour), the 2.5kb linearized vector backbone excised, gel purified (Zymoclean™ Gel DNA Recovery Kit; Zymo Research), and quantified by Nanodrop.

*Sfi*I-digested oligonucleotides (100 ng) were mixed with linearized pMPRA1 vector backbone (50 ng) and ligated by T4 DNA ligase (1U, 16°C, min) to create pMPRA1: rs6679677_ref and pMPRA1: rs6679677_alt. The ligation reaction was stopped by heating (65°C, 20 min).

pMPRA1:rs6679677_ref and pMPRA1:rs6679677_alt were amplified and selected by the transformation in competent *E. coli* DH5-alpha cells (Mix & Go competent cells) according to the manufacturer's instructions (www.zymoresearch.com/). Briefly, competent *E. coli* DH5-alpha cells (100 µL) were thawed on ice before the addition of ligation products (1-5 µL), gentle mixing (by flicking), and incubation on ice (5 min). Immediately following incubation on ice, the transformed competent cells (100 µL) were spread onto pre-warmed LB agar plates supplemented with ampicillin (100 µg/mL) and incubated (37°C, overnight). Single colonies were picked and inoculated into LB:Ampicillin media (5 ml containing 100 µg/mL Ampicillin) and incubated (37°C, overnight) with shaking (~ 200 rpm). Plasmid DNA was extracted using a QIAprep Spin Miniprep Kit, according to the manufacturer's instructions. Plasmids were Sanger sequenced (Massey Genome Service; Massey University) using RVprimer3 (forward) and EBV-rev (reverse) primers in (Table 4-1) to confirm the sequences of the inserts.

pMPRA1:rs6679677_ref, pMPRA1:rs6679677_alt were linearized with *Kpn*I (10 U, 37°C, 1 hour) and purified using the QIAquick PCR Purification Kit (Qiagen). Samples were subsequently digested with *Xba*I (10 U) in the presence of Shrimp Alkaline Phosphatase (1 U, 37°C, 2 hours) prior to heat-inactivation (65°C, 5 min) and purification using the QIAquick PCR Purification Kit (according to the manufacturer's instructions).

A luc2 open reading frame (ORF) was prepared from pMPRAdonor2 (1 µg) by *Kpn*I (20 U) *Xba*I (20 U) double digestion (37 °C, 1 hour), electrophoresis (0.8% agarose, 60V, 1 hour), and gel purification of the 1.7 kb band using the Zymoclean™ Gel DNA Recovery Kit (Zymo Research).

The *luc2* open reading frame was cloned into the *Kpn*I and *Xba*I digested pMPRA1:rs6679677_ref, pMPRA1:rs6679677_alt plasmids, transformed and selected as described earlier. The resultant plasmids (pMPRA1:luc_rs6679677_ref, pMPRA1:luc_rs6679677_alt) were Sanger sequenced (Massey Genome Service; Massey University) using the luciferase primer in (Table 4-1) to confirm the *luc2* gene insertion.

A549 (lung epithelial carcinoma; ATCC) and HepG2 (human liver carcinoma; ATCC) cells were maintained in DMEM and RPMI 1640 (ThermoFisher), respectively, supplemented with 10% fetal bovine serum, 1% GlutaMAX, and 1% penicillin/streptomycin at 37°C in a humid incubator purged with 5% CO_2. Cells were routinely tested for mycoplasma contamination. For transfection, ~1.0×10^5 cells were seeded in a single well of a 24-well plate, followed by the addition of 500 µL of the appropriate complete media. On the day of transfection (24 hours following cell plating), ~75% confluent wells were co-transfected with luciferase plasmid DNA (i.e. 800 ng of pMPRA1:luc_rs6679677_ref, pMPRA1:luc_rs6679677_alt, or pMPRAdonor2 luciferase control) and a beta-galactosidase control plasmid (200 ng) using lipofectamine 3000 (ThermoFisher; according to the manufacturer's instructions).

At 48 hours following transfection, cells were lysed using the Glo lysis buffer (Promega) and luciferase activity assessed using the ONE-GloTM Luciferase Assay System (Promega) in a VarioskanTM LUX multimode microplate reader (according to the manufacturer's instructions). For the beta-galactosidase assay, 20 µl beta-galactosidase reagent (*i.e.* 0.2M phosphate buffer (pH 7.4), 2 mM MgCl2, 100 mM β-mercaptoethanol, and 1.3 mg/ml ortho-Nitrophenyl-β-galactoside) was added to 20 µl of transfection cell lysate (prepared using the Glo lysis buffer) in a 96 well plate and incubated at 37°C for 30 minutes. The absorbance was then read at 420 nM in a VarioskanTM LUX multimode microplate reader following the manufacturer's instructions.

4.2.9 Data analysis

All statistical testing was performed using R software (version v3.6.3) [194], Scikit-learn (version 0.21.3; [183,184]), tsfresh (version 0.12.0 [190]), and pymc3 (version 3.8; [195]). Visualization for the luciferase luminescence was performed using GraphPad Prism (v8.4.3).

4.2.10 Code Availability

CoDeS3D pipeline is available at: https://github.com/Genome3d/codes3d-v1.

Python scripts used for machine learning are available at:
https://github.com/Genome3d/T1D_logistic_lasso_predictor.git/

Python version 3.7.3 was used for all the python scripts.

4.3 Results

4.3.1 T1D SNPs impact an extensive gene regulatory network

The methodology used for the characterization of the regulatory networks for T1D-associated SNPs is summarized in Figure 4-3. Briefly, we used the CoDeS3D[162] algorithm to analyze 313 T1D-associated SNPs (Methods; Appendices: Supplementary Table 1) using Hi-C chromatin contact libraries (Appendices: Supplementary Table 2) and GTEx (v7) RNAseq data (Methods). The Hi-C libraries that were used in this study included immortalized cell lines and primary human tissues (Appendices: Supplementary Table 2) and were chosen to ensure a range of known and possible interactions were included in the analysis. We define a spatial eQTL as SNP that tag a locus that: 1) physically interacts with a gene; and 2) explains a fraction of the genetic variance of the interacting gene transcription level. According to our definition, the eQTL variant can sit anywhere within the genome. This includes within the boundaries of the gene, as long as the gene is covered by ≥3 restriction fragments in the Hi-C library. Of the 313 SNPs, 57 SNPs had no identifiable eQTLs, resulting in 256 T1D-associated SNPs connecting to 822 genes (1479 spatial eQTL-eGene associations; FDR $q < 0.05$; Appendices: Supplementary Table 3). As expected from our previous study[15], the 822 genes were enriched for immune activation and response pathways (Appendices: Supplementary Table 4).

The eQTL-eGene interactions were categorized as either: *cis*, the eQTL and eGene are separated by a linear distance of ≤1Mb on the same chromosome; or *trans*, eQTLs and their eGenes were separated by >1Mb on the same chromosome or located on different chromosomes. Notably, of the 256 T1D-associated SNPs with spatial-eQTLs, 190 affected the *trans*-regulation of 361 genes, while 201 affected the *cis*-regulation of 493 genes. Some genes (n= 32) were regulated by different eQTLs in both *cis* and *trans* (*e.g. TRIM26, RNF5, PSMB9,* and *NOTCH4*; Appendices: Supplementary Table 2). Notably, the 112 *trans*-regulated genes (*e.g. FOXP1*, *CAMTA1*, and *ROBO2*) we identified are enriched for being less tolerant of inactivating (*i.e.* Loss-of-Function) mutations (Figure 4-4; Appendices: Supplementary Table 5).

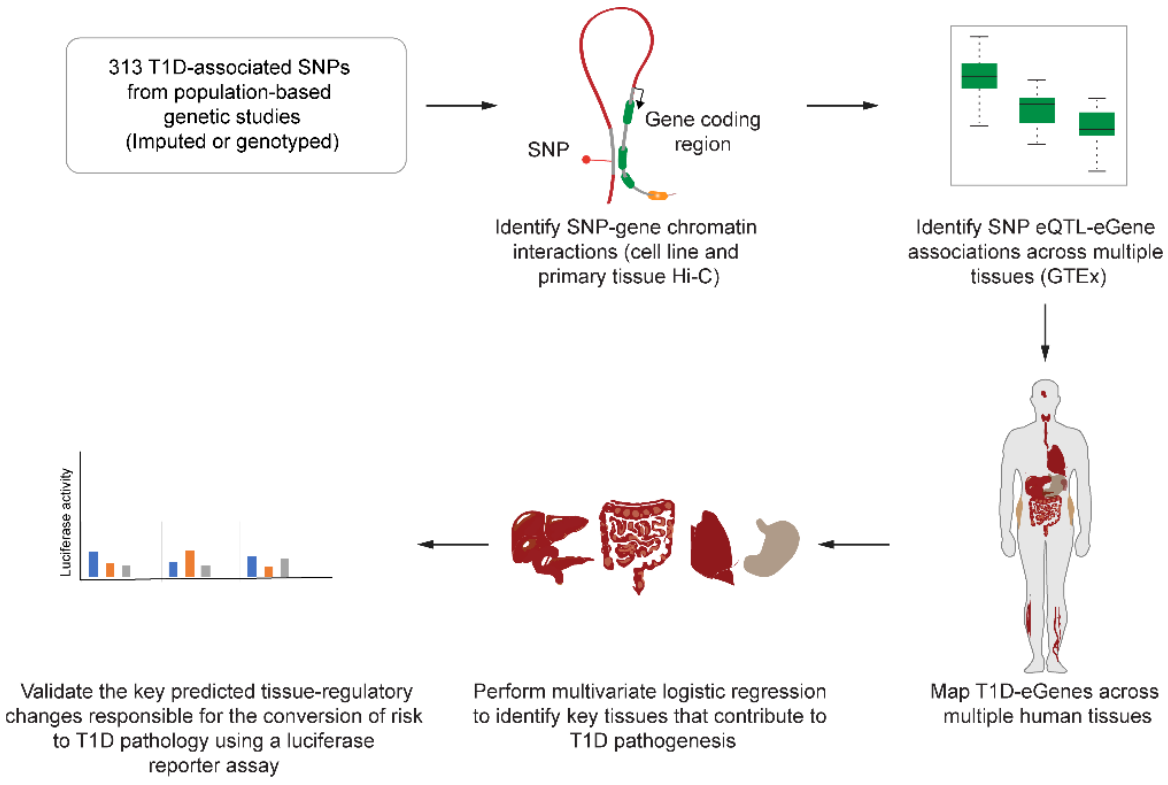


Figure 4-3: Overview of the methods used to predict the regulatory effects of genetic variants associated with the development of T1D

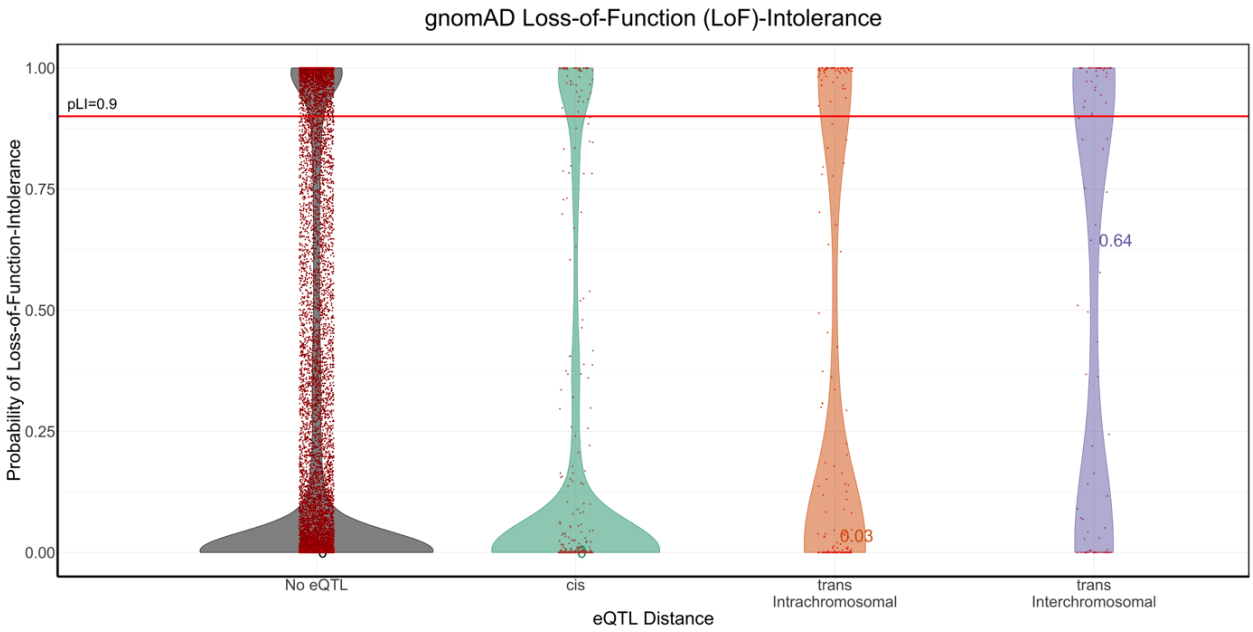


Figure 4-4: Loss of function analysis for spatially regulated genes

Analyses of the loss of function intolerance of the genes that are spatially regulated by T1D SNPs using the Genome Aggregation Database (gnomAD)[196]. *Trans*-regulated genes are less tolerant of inactivating mutations. No eQTLs – gnomAD genes with no significant eQTL associations with T1D SNPs; *cis* – genes with significant eQTL associations with T1D SNPs < 1 Mb window; *trans*-intrachromosomal – genes with significant eQTL associations with T1D SNPs > 1Mb but within the same chromosomes; *trans*-interchromosomal – genes with significant eQTL associations with T1D SNPs > 1Mb but on different chromosomes. Probability of Loss of Function-Intolerance is the probability for genes to loss their biological functions with a single genetic mutation.

4.3.2 Machine learning identifies transcriptional changes in the lung as key for conversion of risk to T1D risk

Based on our eQTL analysis, we determined that T1D SNPs form an integrated gene regulatory network across tissues and immune cell types. We reasoned that we could use a machine learning approach to integrate the tissue-specific spatial eQTL-eGene associations with thousands of individual genotypes from large T1D cohorts to convert population level risk (*i.e.* GWAS SNPs) to individualized risk (*i.e.* the burden for an individual's genotype).

We trained and validated the predictive accuracy of a regularised logistic regression predictor (Methods; Figure 4-1), which predicts the T1D disease status. This model was used to estimate the additive tissue-specific contribution of the spatial eQTLs within genotypes from individuals who developed T1D (WTCCC; 1960 T1D cases and 2933 controls). Individual genotypes were weighted using the tissue-specific spatial eQTL effect sizes from the CoDeS3D analysis (Appendices: Supplementary Table 3). Of the 313 T1D-associated SNPs, 253 were present within each of the WTCCC genotypes (Appendices: Supplementary Table 1 and 6). Of the 253 SNPs, 224 had identifiable eQTLs, connecting to 758 eGenes (6307 tissue-specific eQTL effects).

Essential feature selection was performed using the Mann-Whitney U test[158] with the BY control[163] (selected 2048 data features from 6307 eQTL features and 29 SNP features with unknown eQTL effects) and lasso regularization of the logistic regression. T1D regularised logistic regression predictor training (80% of the WTCCC derived eQTL dataset) was implemented with Grid search and 10-fold validation to identify the optimized hyperparameters (Methods). In this study, the AUC was used to identify the top performing predictor developed using the additive tissue-specific contributions of the spatial eQTLs within genotypes from individuals who developed T1D. As such, our predictor AUC is not directly comparable to AUCs generated by polygenic risk scores on the same datasets. The T1D regularised logistic regression predictor with the optimized hyperparameters (T1D model-1 with 135 selected data features) delivered an AUC 0.76 prediction performance with the validation data (20% of the WTCCC derived eQTL dataset) (Methods). T1D model 1 was trained by 80% and validated by 20% of the WTCCC derived eQTL dataset. Thus, T1D model-1 was further validated by 10 repeats of 5-fold cross validation. We estimated the variation in predictor performance using the AUCs for 50 T1D regularised logistic regression predictors that were created with the optimised hyperparameters of T1D model 1 on randomized subsets (10 repeats of 5-fold cross-validation [80% training and 20% validation]) of the WTCCC derived eQTL dataset (Figure 4-5 (Methods). The mean AUC prediction, calculated from the 50 T1D regularised logistic regression predictors, was only 1.7% different (i.e. [0.76-0.747]/0.76) and the distribution encompassed the original AUC (min = 0.712, max = 0.771, standard deviation of 0.14; Figure 4-5, [95% confident interval (0.743,0.751)]).Therefore, we concluded that the predictors created with the optimized hyperparameters performed well across different data sets generated from the 10 repeats of 5-fold cross-validation.

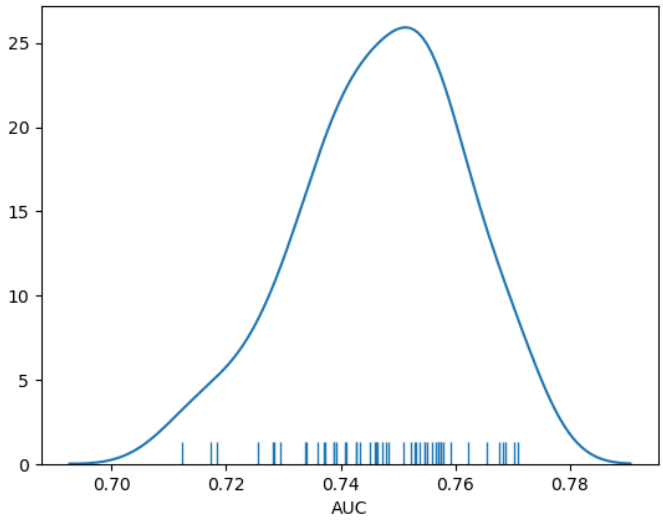


Figure 4-5: AUC distribution of the 50 predictors

Kernel Density Estimate plot of AUC distribution created from the 10 repeated 5-fold cross-validations (50 predictors) of the T1D regularized logistic regression predictors with the optimised hyperparameters of T1D model-1. The AUC mean is 0.747 with a standard deviation of 0.14. [X axis: AUC (percentage/100), Y axis: frequency of AUC (%)] [95% confident interval (0.743,0.751)]

In the regularised logistic regression predictors, the absolute values of the model weights (coefficients) of the eQTL effects or SNPs were used as proxies of their contributions to the predicted disease risk. Tissue-specific contributions to the T1D risk were extracted from each of the 50 T1D regularised logistic regression predictor models as the sum of the absolute values of the model weights (coefficients) of the selected eQTL or SNP features associated with each tissue group (Figure 4-6; Appendices: Supplementary Table 7) (Methods). We then ranked the tissue-specific contributions to the 50 regularised logistic regression predictors. This ranking identified the lung as the top average contributor to the relative risk (case:control) of developing T1D. Across all 50 regularised logistic regression predictors, the lung explained a mean of 13.6% (standard deviation of 2.51%) of the relative risk of developing T1D (Figure 4-6; Appendices: Supplementary Table 8).

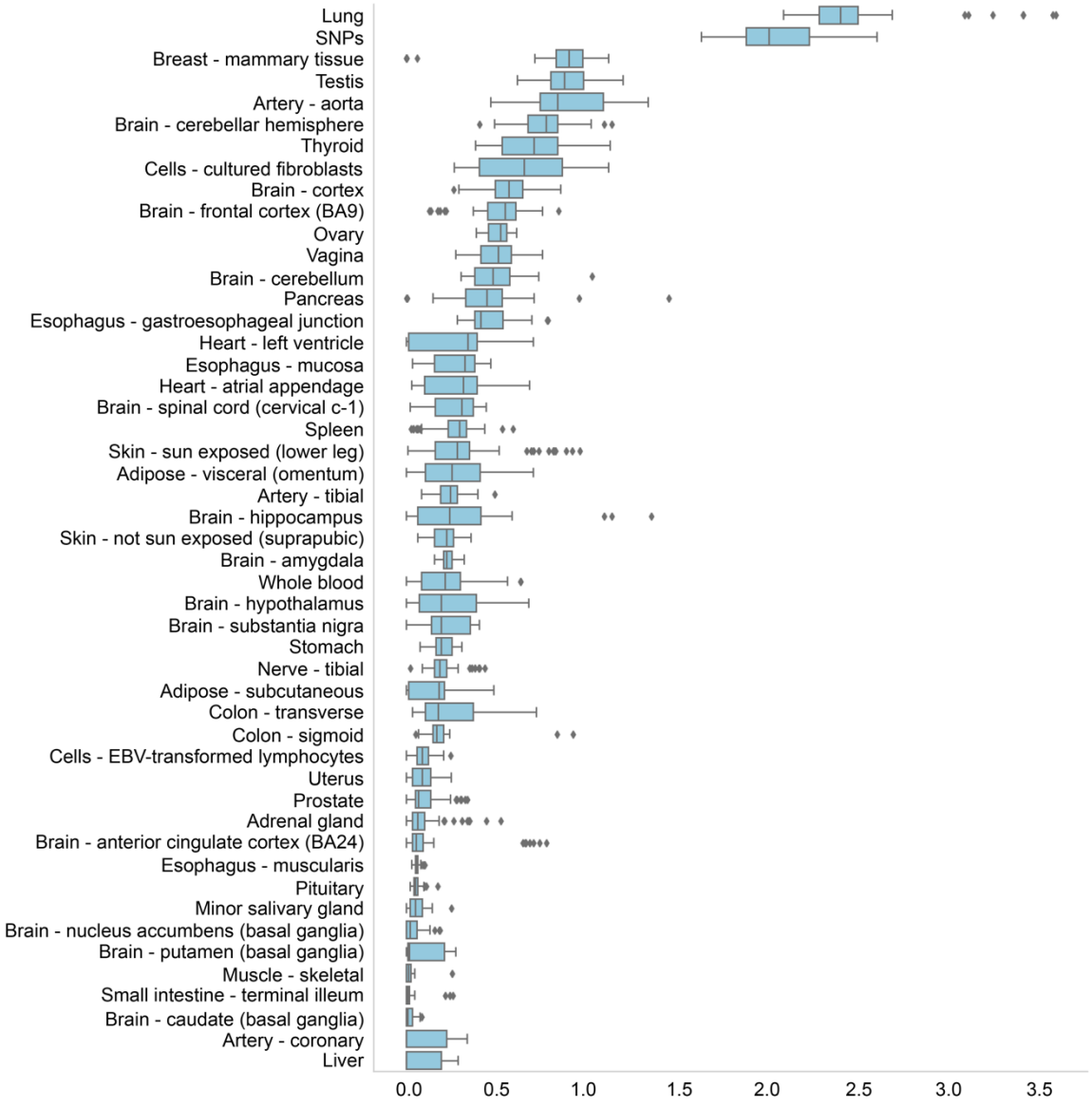


Figure 4-6: Tissue specific contributions of 50 regularised logistic regression predictors created with T1D model-1's hyperparameters

Kernel Density Estimate plot of AUC distribution created from the 10 repeated 5-fold cross-validations (50 predictors) of the T1D regularized logistic regression predictors with the optimised hyperparameters of T1D model-1. The AUC mean is 0.747 with a standard deviation of 0.14. [X axis: AUC (percentage/100), Y axis: frequency of AUC (%)] [95% confident interval (0.743,0.751)] The distributions were created from the 50 T1D regularised logistic regression predictors that were created using the optimised hyper-parameters of model 1. These regularised logistic regression predictors integrated four different forms of biological information: 1) GWAS or fine mapping; 2) Hi-C; 3) eQTL; and 4) genotype data. SNPs denote T1D-associated SNPs for which no eQTLs were identified. X axis is the total value of the model weights (no units).

4.3.3 *CTLA4* contributes to the risk associated with the lung and testes

When training our T1D regularised logistic regression predictors, we identified a split distribution for the lung that was dependent upon the lasso regularization inclusion or exclusion of the rs3087243-*CTLA4* spatial eQTL within the two tissues where it was identified (*i.e.* lung or testes; Figure 4-7). Since lasso regularization keeps only one of a group of highly correlated features, we sought to validate the possibility that rs3087243-*CTLA4* has significant effects on the contribution of the lung and testis to T1D risk.

To test the effects of specifically removing the rs3087243-*CTLA4* feature, we created two alternative weighted WTCCC genotype T1D eQTL models in which this eQTL was removed from either the lung or the testis within the weighted WTCCC genotype T1D-eQTL matrix. The AUCs and tissue-specific contributions from 50 T1D predictors with the optimized hyperparameters from each of the alterative matrices were evaluated by Mann-Whitney U test[158] controlled by Benjamini Yekutieli FDR[191] and Bayesian estimation (Figure 4-8 and 4-9;[197]). The Mann Whitney U Test method was used to assess the differences of the risk contribution distributions of tissue groups and the AUC results from the two sets of the 50 predictors of the two alternative models (Figure 4-8). On the other hand, the Bayesian estimation method was only used to evaluate the differences of the AUC results from the two sets of the 50 predictors (Figure 4-9)[197]. The distribution of the AUC result differences was prior descripted as Student-T. with the mean ~ N(sample mean, sample std * 2), stand deviation ~ U(0,1) and degree of freedom ~ (Exp(1/29) +1). With 2000 iterations of MCMC stimulation with the prior parameter distributions, the Bayesian estimation method pulled out all the creditable combinations of the parameters and created the posterior distributions of each model parameter given the AUC result difference data[197]. There was a 94% chance that the mean was between -6.7×10^{-5} and 3.6×10^{-5} including zero, and the stand deviation was between 1.5×10^{-3} and 2.3×10^{-3}.

The results indicated that no other tissues were affected, consistent with the rs3087243-*CTLA4* spatial eQTL only being detected in the lung and testes. The lung rs3087243-*CTLA4* eQTL contributed an average of 4% to T1D risk (Appendices: Supplementary Table 9). The lung eQTL involving rs3087243 and *CTLA4* is also notable as: 1) rs3087243 has also been linked with progression from single to multiple autoantibodies in the TrialNet PTP cohort[168]; 2) *CTLA4* encodes an immunoglobulin protein crucial for modulating T cell function and mediating autoimmunity; and 3) immune intervention trials targeting *CTLA4* have reported significant but short-term positive metabolic outcomes[198].

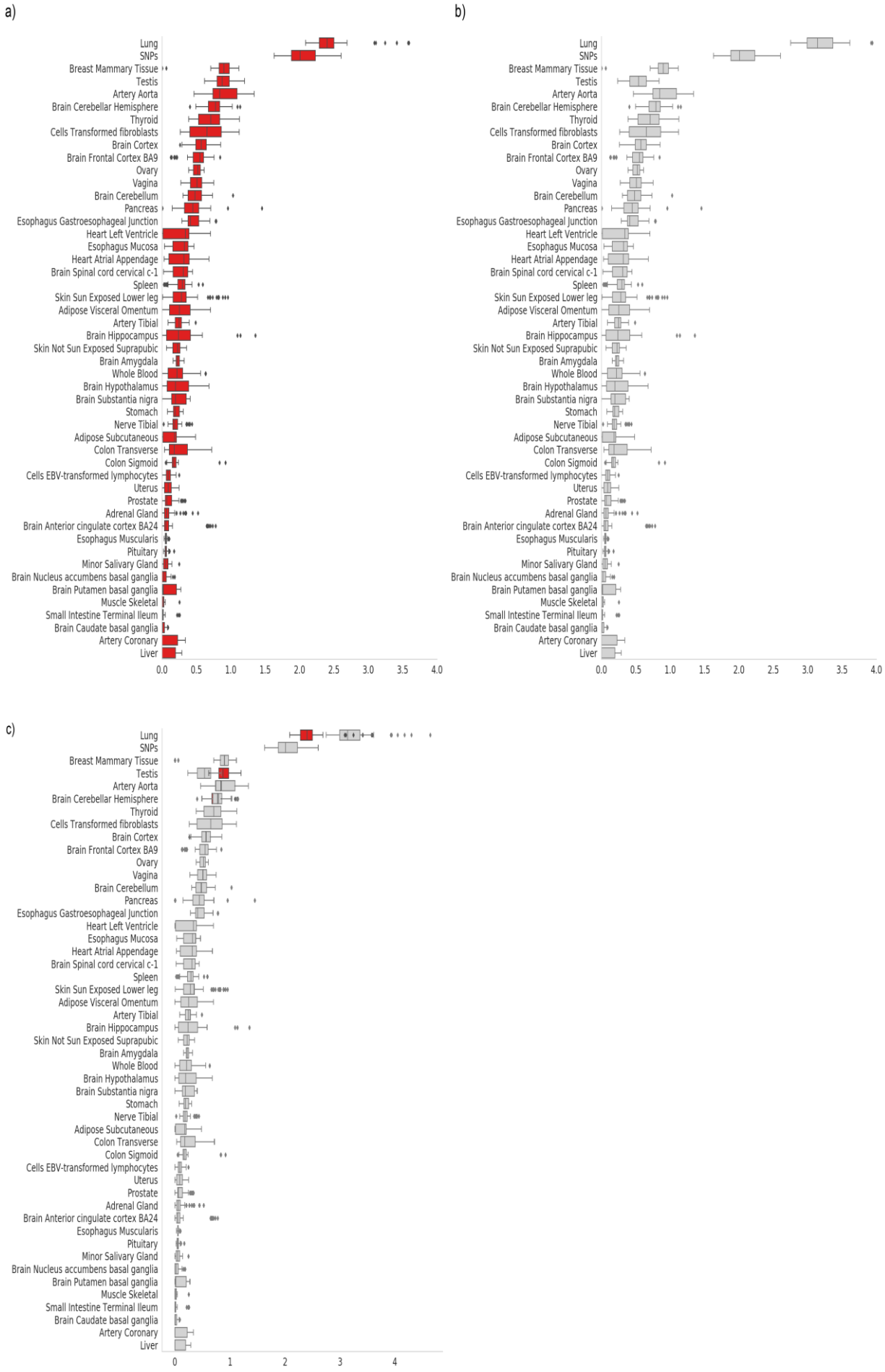

a)
b)
c)
Lung
SNPs
Breast Mammary Tissue
Testis
Artery Aorta
Brain Cerebellar Hemisphere
Thyroid
Cells Transformed fibroblasts
Brain Cortex
Brain Frontal Cortex BA9
Ovary
Vagina
Brain Cerebellum
Pancreas
Esophagus Gastroesophageal Junction
Heart Left Ventricle
Esophagus Mucosa
Heart Atrial Appendage
Brain Spinal cord cervical c-1
Spleen
Skin Sun Exposed Lower leg
Adipose Visceral Omentum
Artery Tibial
Brain Hippocampus
Skin Not Sun Exposed Suprapubic
Brain Amygdala
Whole Blood
Brain Hypothalamus
Brain Substantia nigra
Stomach
Nerve Tibial
Adipose Subcutaneous
Colon Transverse
Colon Sigmoid
Cells EBV-transformed lymphocytes
Uterus
Prostate
Adrenal Gland
Brain Anterior cingulate cortex BA24
Esophagus Muscularis
Pituitary
Minor Salivary Gland
Brain Nucleus accumbens basal ganglia
Brain Putamen basal ganglia
Muscle Skeletal
Small Intestine Terminal Ileum
Brain Caudate basal ganglia
Artery Coronary
Liver
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
0
1
2
3
4

Figure 4-7: Tissue contributions of the T1D logistic lasso regression models

The tissue contributions with eQTL rs3087243 either at testis **(a)** or at lung **(b)** created from 10 repeated 5-fold cross-validations (50 predictors). **(c)** Overlap of **(a)** and **(b)** highlighting the lung and testes differences. The SNPs category denotes T1D-associated SNPs that are not eQTLs. The X axes are the total value of the model weights (no units).

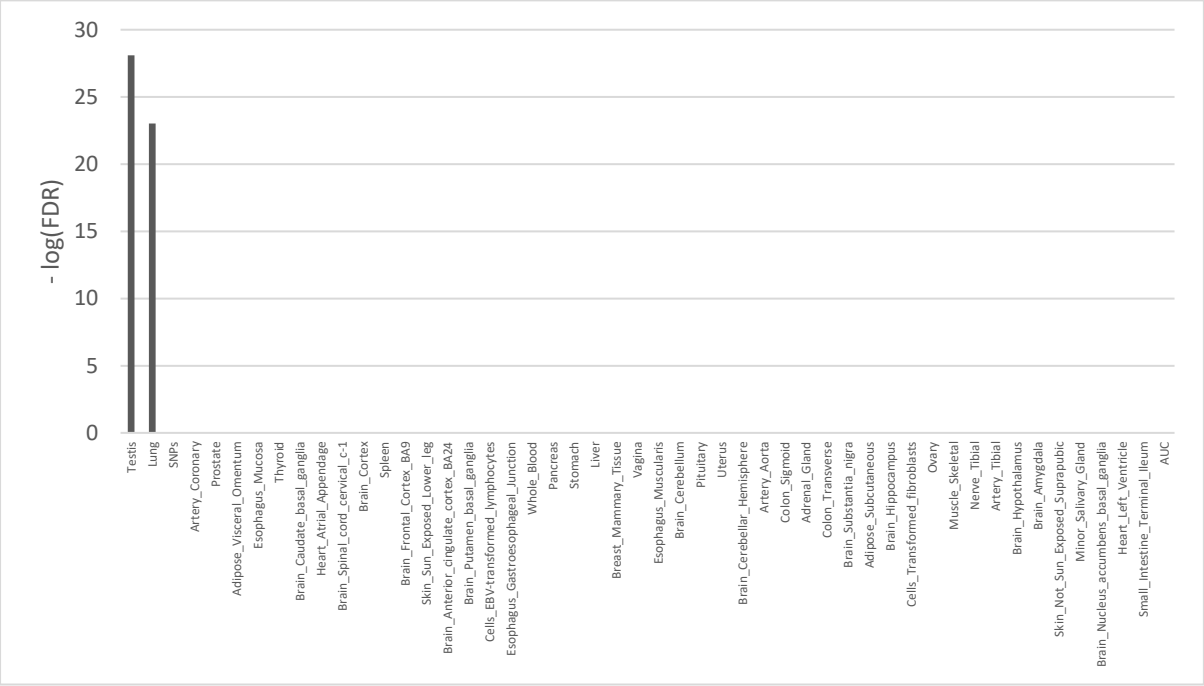


Figure 4-8: P-values (-log10) of the tissue-specific contribution and AUC differences of 50 T1D predictor pairs with eQTL rs3087243 at testis or at lung

Most of the GTEx tissue-specific contributions and AUCs of the predictor pairs are not significantly different. However, the contributions at lung and testis have significant differences (Mann Whitney U Text controlled by BY FDR <0.01).

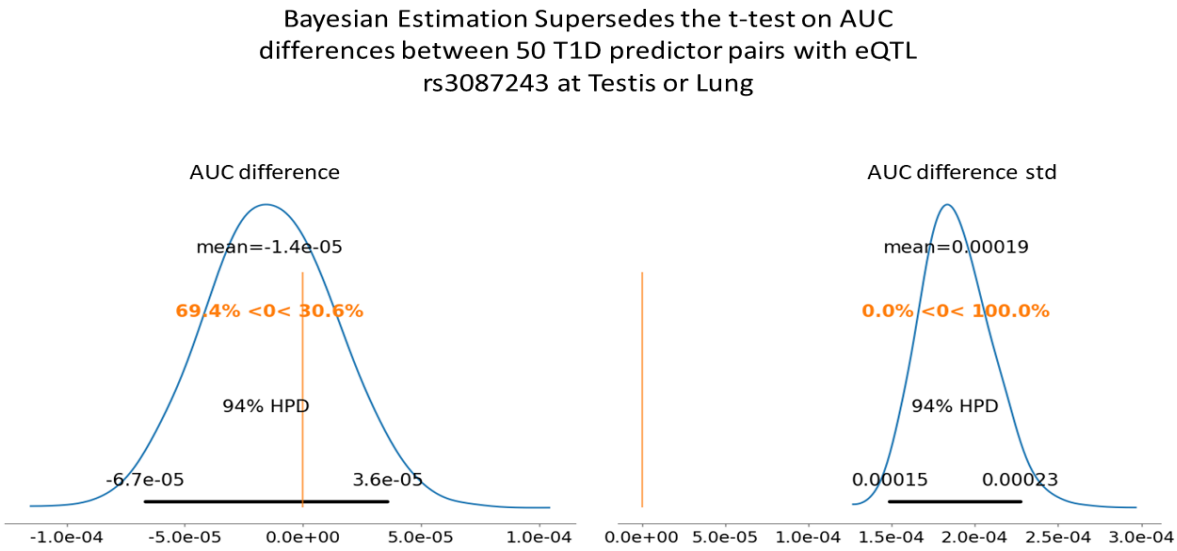


Figure 4-9: AUC difference distribution between 50 T1D regularised logistic regression predictor pairs with the eQTL rs3087243 either at testis or at lung evaluated by Bayesian estimation supersedes the t-test (2000 iterations of model simulation)

The simulated AUC difference data mean is 1.4e-05 with the standard deviation of 0.00019. [Y axis: posterior distribution; X axis: Mean of AUC difference and Standard deviation of AUC differences; AUC; Area under the Curve. HPD; highest posterior density interval]. Informed prior was modelled as Student-T with mean ~ N(sample mean, sample std * 2), stand deviation ~ U(0,1) and degree of freedom ~ (Exp(1/29) +1).

4.3.4 Predictions from T1D model-1 were confirmed using a second model T1D model-2 trained with more data

To enable more precise data-feature risk estimations, a second T1D regularised logistic regression predictor (T1D model-2) (134 tissue-specific eQTL effects across GTEx tissues and six SNPs with unknown eQTLs) was created (Methods) and trained with the optimised hyperparameters using the full WTCCC cohort (In-sample AUC = 0.774). (T1D model-2 differed from T1D model-1, which used an 80:20 split for internal WTCCC validation.) T1D model-2 was validated against the UK Biobank (UKBB) (30 subsets of T1D datasets of 415 cases and 578 controls). T1D model 2 achieved a mean AUC = 0.754 (Table 4-2; Figure 4-10; Table 4-3) and was used to rank the eQTLs that impacted on the lung contribution to T1D risk (Table 4-4). It should be noted that this T1D model-2 excluded the rs3087243-*CTLA4* eQTL, which contributed an average of 4% to T1D risk in the 50 T1D regularised logistic regression predictors (calculated with the optimised hyperparameters of T1D model-1; see above).

Table 4-2: A summary from the Bayesian analysis of the validation AUCs from the model 2 predictor on the 30 UK Biobank dataset

	Mean	SD	HPD_2.5%	HPD_97.5%
μ	0.754	0.002	0.749	0.758
σ	0.011	0.002	0.009	0.014

μ - mean of the simulations (1000) from the model created using 30 AUC prediction results; σ - standard deviation from the simulations; SD - standard deviation; HPD - highest posterior density interval; AUC - Area Under the Curve. Uninformed prior distribution was normal with mean ~ U(0,1) and standard deviation ~ HalfNormal(std=0.01). [X axis: mean AUC and standard deviation of AUC, Y axis: posterior distribution]. From the 1000 simulation iterations with the uninformed prior distributions, the posterior distributions of parameters showed there was a 95% chance that the mean (μ) AUC was between 0.749 and 0.758, and its SD (σ) was between 0.009 and 0.014.

The major contributor eQTL (rs6679677) down-regulated *AP4B1-AS1* transcript levels in the lungs and conferred a 13.3% contribution to the T1D risk prediction model. This was comparable to the mean predictor weighting (13.6%) of the 50 models created with T1D model-1's hyperparameters. Notably, rs6679677 also down-regulates *AP4B1-AS1* expression in whole blood samples (eQTLGen; Table 4-5) as well as modulating the expression of genes associated with immune regulation, including *FOXP3*, *CTLA4*, *IL2RA*, and *SLAMF1* in whole blood (eQTLGen; http://www.eqtlgen.org [199]; Table 4-5).

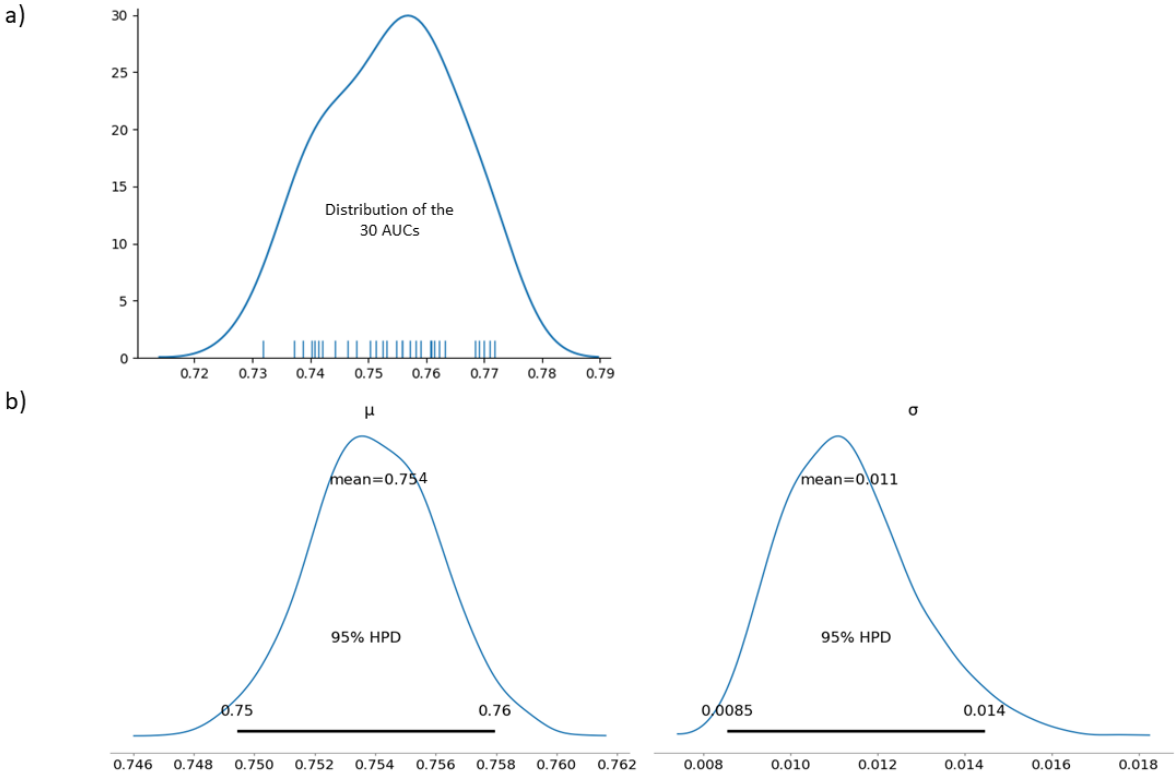


Figure 4-10: Validation of AUC results from T1D model-2 on the 30 UK Biobank test dataset

a) Kernel density estimate (KDE) plot of the 30 AUC results (mean=0.754; SD=0.0011. [X axis: AUC, Y axis: frequency of AUC values (%)] **b)** The posterior estimation results from the Bayesian analysis (1000 iterations of model simulation) of the 30 AUCs. From the Bayesian estimation, there was a 95% chance that the mean (μ) AUC was between 0.749 and 0.758, and its SD (σ) was between 0.009 and 0.014. SD; standard deviation, AUC; Area under the Curve. HPD; highest posterior density interval. Uninformed prior distribution was normal with mean ~ U(0,1) and standard deviation ~ HalfNormal(std=0.01). [X axis: mean AUC and standard deviation of AUC, Y axis: posterior distribution]

Table 4-3: AUC results of the validating final T1D predictor on the 30 UK Biobank test dataset

UKBIO T1D AUC	
0.7318	
0.7373	
0.7387	
0.7402	
0.7408	
0.7414	
0.7421	
0.7443	
0.7464	
0.7480	
0.7503	
0.7513	range from 0.732 to 0.772
0.7526	mean: 0.754
0.7531	std: 0.011
0.7548	
0.7558	
0.7558	
0.7572	
0.7583	
0.7590	
0.7607	
0.7609	
0.7614	
0.7622	
0.7633	
0.7685	
0.7692	
0.7700	
0.7709	
0.7719	

Table 4-4: Ranking of eQTLs on tissue-specific contribution to T1D risk using the final T1D classification model

The major contributor eQTL (rs6679677) modulates AP4B1-AS1 transcript levels in the lung and contributes 13.3% to T1D risk. (Data features are the features selected in the T1D classification model. Weights are the model weight of the data features in the T1D regularised logistic predictor model. The Lung eQTL risk contribution is calculated by ([abs(-2.3626)/sum(abs(weights)] *100).

Data feature	weight
Lung--rs6679677_A--AP4B1-AS1	-2.3626
rs9272346_G	-1.2298
Breast_Mammary_Tissue--rs947474_G--ZNF365	-0.9106
Ovary--rs11711054_G--CCRL2	-0.4915
Testis--rs3087243_A--CTLA4	-0.3969
Pancreas--rs10795791_G--CD300A	-0.3949
Vagina--rs12927355_T--RP11-848P1.9	-0.3843
Brain_Frontal_Cortex_BA9--rs743777_G--MYO18B	-0.3742
Brain_Cortex--rs17388568_A--KIAA1109	-0.3291
Cells_Transformed_fibroblasts--rs2269241_C--RN7SL130P	-0.3149
Artery_Aorta--rs2269241_C--RN7SL130P	-0.2727
Esophagus_Mucosa--rs16956936_T--CHRNB1	-0.2653
Brain_Hypothalamus--rs1250563_C--CCAR1	-0.2641
Brain_Amygdala--rs17388568_A--KIAA1109	-0.2362
Artery_Aorta--rs13415583_G--AFF3	-0.1911
Skin_Sun_Exposed_Lower_leg--rs16956936_T--EFNB3	-0.1878
Adipose_Visceral_Omentum--rs2069762_C--36951	-0.1865
Adipose_Subcutaneous--rs16956936_T--EFNB3	-0.1842
Spleen--rs151234_C--SULT1A2	-0.1798
Thyroid--rs12927355_T--RMI2	-0.1613
Brain_Cerebellar_Hemisphere--rs7221109_T--KRT24	-0.1521
Thyroid--rs16956936_T--KCNAB3	-0.1362
Stomach--rs11711054_G--OSBPL10	-0.1344
Colon_Sigmoid--rs193778_G--RMI2	-0.1157
Testis--rs231775_G--CTLA4	-0.1092
Artery_Tibial--rs16956936_T--CNTROB	-0.0964
Brain_Spinal_cord_cervical_c-1--rs11052552_T--CLECL1	-0.0860
Artery_Tibial--rs2542151_G--RP11-973H7.1	-0.0857
Stomach--rs9924471_A--SULT1A2	-0.0793
Esophagus_Mucosa--rs7221109_T--KRT24	-0.0784
Adipose_Visceral_Omentum--rs151234_C--SULT1A2	-0.0779
Brain_Frontal_Cortex_BA9--rs10947332_A--HLA-DQB2	-0.0777
Whole_Blood--rs151234_C--SULT1A2	-0.0750
Brain_Frontal_Cortex_BA9--rs13415583_G--AFF3	-0.0701
Minor_Salivary_Gland--rs1052553_G--KANSL1-AS1	-0.0685
Brain_Cerebellum--rs2292239_T--RPS26	-0.0672
Brain_Cerebellar_Hemisphere--rs757411_C--KRT24	-0.0650
Nerve_Tibial--rs11052552_T--CLECL1	-0.0641
Brain_Cortex--rs13415583_G--AFF3	-0.0572
Cells_EBV-transformed_lymphocytes--rs3135002_A--HLA-DRB6	-0.0562
Brain_Hippocampus--rs9924471_A--SULT1A2	-0.0545
Skin_Not_Sun_Exposed_Suprapubic--rs2292239_T--RPS26	-0.0500
Spleen--rs151234_C--ZNF423	-0.0483
Brain_Hypothalamus--rs917911_C--CLECL1	-0.0439
Heart_Atrial_Appendage--rs9924471_A--SULT1A2	-0.0430

Nerve_Tibial--rs917911_C--CLECL1	-0.0417
Brain_Cerebellar_Hemisphere--rs231775_G--FMNL2	-0.0406
Artery_Aorta--rs2542151_G--RP11-973H7.1	-0.0392
Brain_Anterior_cingulate_cortex_BA24--rs10947332_A--HLA-DQB2	-0.0374
Colon_Sigmoid--rs2542151_G--RP11-973H7.1	-0.0363
Brain_Hypothalamus--rs4505848_G--ELF2	-0.0329
Colon_Transverse--rs2542151_G--RP11-973H7.1	-0.0326
Esophagus_Muscularis--rs2542151_G--RP11-973H7.1	-0.0314
Nerve_Tibial--rs151234_C--SULT1A2	-0.0309
Heart_Atrial_Appendage--rs2292239_T--RPS26	-0.0299
Brain_Spinal_cord_cervical_c-1--rs2292239_T--RPS26	-0.0292
Brain_Cortex--rs416603_A--RMI2	-0.0288
Pituitary--rs2292239_T--RPS26	-0.0266
Esophagus_Mucosa--rs72727394_T--RP11-275I4.2	-0.0262
Nerve_Tibial--rs10947332_A--HLA-DQB2	-0.0229
Artery_Tibial--rs1052553_G--CRHR1-IT1	-0.0225
Minor_Salivary_Gland--rs10947332_A--HLA-DQB2	-0.0216
Cells_Transformed_fibroblasts--rs2476601_A--AP4B1-AS1	-0.0215
Adrenal_Gland--rs917911_C--CLECL1	-0.0198
Brain_Caudate_basal_ganglia--rs17388568_A--KIAA1109	-0.0175
Lung--rs1052553_G--KANSL1-AS1	-0.0162
Colon_Transverse--rs2251396_A--PSORS1C1	-0.0150
Artery_Aorta--rs2292239_T--RPS26	-0.0147
Esophagus_Muscularis--rs2292239_T--RPS26	-0.0123
Thyroid--rs2292239_T--RPS26	-0.0112
Brain_Putamen_basal_ganglia--rs2292239_T--RPS26	-0.0082
Cells_EBV-transformed_lymphocytes--rs11203203_A--CDC42BPA	-0.0069
Cells_EBV-transformed_lymphocytes--rs9268645_G--HLA-DQB2	-0.0066
Ovary--rs2292239_T--RPS26	-0.0062
Brain_Cerebellar_Hemisphere--rs917911_C--AC091814.2	-0.0058
Brain_Cerebellum--rs3135002_A--HLA-DRB6	-0.0058
Colon_Sigmoid--rs10947332_A--HLA-DQB2	-0.0037
Esophagus_Gastroesophageal_Junction--rs2542151_G--RP11-973H7.1	-0.0032
Adrenal_Gland--rs2251396_A--PSORS1C1	-0.0028
Cells_EBV-transformed_lymphocytes--rs10947332_A--HLA-DQB2	-0.0020
Uterus--rs11171739_T--RPS26	0.0014
Esophagus_Muscularis--rs45450798_G--RP11-973H7.1	0.0021
Brain_Spinal_cord_cervical_c-1--rs3135002_A--HLA-DQB1	0.0045
Esophagus_Gastroesophageal_Junction--rs45450798_G--RP11-973H7.1	0.0059
Heart_Left_Ventricle--rs3129889_G--HLA-DRB5	0.0062
Pituitary--rs3135002_A--HLA-DQB1	0.0073
Brain_Cerebellum--rs1689510_C--SUOX	0.0074
Artery_Tibial--rs9981624_C--XRRA1	0.0101
Small_Intestine_Terminal_Ileum--rs2251396_A--MICA	0.0103
Skin_Sun_Exposed_Lower_leg--rs3129889_G--HLA-DRB5	0.0108
Spleen--rs62447205_G--FIGNL1	0.0124
Spleen--rs3129889_G--HLA-DRB5	0.0125
Pituitary--rs62447205_G--FIGNL1	0.0150
Muscle_Skeletal--rs10947332_A--HLA-DQB1	0.0162
rs2647044_A	0.0188
Thyroid--rs3129889_G--HLA-DRB5	0.0190
Thyroid--rs2903692_A--HNRNPCP4	0.0192

Spleen--rs478222_T--CACNA2D3	0.0202
Nerve_Tibial--rs1689510_C--SUOX	0.0225
Colon_Sigmoid--rs45450798_G--RP11-973H7.1	0.0274
Brain_Cerebellar_Hemisphere--rs4763879_A--AC091814.2	0.0287
Spleen--rs2251396_A--MICA	0.0307
Brain_Cerebellum--rs9653442_C--AFF3	0.0368
Testis--rs2269241_C--PGM1	0.0459
Brain_Cerebellum--rs1270942_G--C4A	0.0464
Brain_Cerebellar_Hemisphere--rs9653442_C--AFF3	0.0468
Artery_Tibial--rs45450798_G--RP11-973H7.1	0.0483
Prostate--rs62447205_G--FIGNL1	0.0592
Brain_Cerebellar_Hemisphere--rs2153977_C--TSPAN2	0.0596
Brain_Cerebellar_Hemisphere--rs62447205_G--FIGNL1	0.0608
Uterus--rs2903692_A--COTL1	0.0664
Testis--rs3129889_G--HLA-DRB5	0.0713
Nerve_Tibial--rs4763879_A--CLECL1	0.0737
Esophagus_Gastroesophageal_Junction--rs12722495_C--ITIH5	0.0759
Adrenal_Gland--rs4763879_A--CLECL1	0.0780
Skin_Sun_Exposed_Lower_leg--rs72727394_T--RASGRP1	0.0841
rs2187668_T	0.0988
Brain_Hypothalamus--rs4763879_A--CLECL1	0.0988
Brain_Cerebellum--rs2857595_A--MICB	0.1029
Brain_Cerebellar_Hemisphere--rs2857595_A--MICB	0.1097
Brain_Cerebellum--rs10758593_A--PDCD1LG2	0.1101
rs9469200_C	0.1147
Testis--rs193778_G--RP11-485G7.5	0.1172
Brain_Cerebellum--rs62447205_G--FIGNL1	0.1310
Skin_Not_Sun_Exposed_Suprapubic--rs72727394_T--RASGRP1	0.1460
rs2327832_G	0.1568
Artery_Aorta--rs9653442_C--AFF3	0.1655
Brain_Cortex--rs4505848_G--KIAA1109	0.1740
Thyroid--rs12927355_T--HNRNPCP4	0.1763
Brain_Spinal_cord_cervical_c-1--rs11203203_A--XKR6	0.1769
Heart_Atrial_Appendage--rs9981624_C--XRRA1	0.1782
Whole_Blood--rs478222_T--ADCY3	0.1845
Vagina--rs8056814_A--CFDP1	0.1854
rs9269173_A	0.2252
Colon_Transverse--rs6534347_A--KIAA1109	0.2547
Cells_Transformed_fibroblasts--rs7020673_C--EIF4ENIF1	0.3010
Brain_Hippocampus--rs12444268_T--RP11-525K10.3	0.3070
Brain_Cerebellar_Hemisphere--rs61839660_T--RIN2	0.3162
Esophagus_Gastroesophageal_Junction--rs61839660_T--ITIH5	0.3463
Heart_Left_Ventricle--rs8056814_A--RP11-252K23.2	0.3544

Table 4-5: Regulatory effects of rs6679677 from the blood eQTL database (http://www.eqtlgen.org)

SNP	Chr	Pos (hg19)	Gene	Gene_Symbol	Chr	Pos (hg19)	Z-score	Assessed	Other	Nr Cohorts	Nr Samples	P-value	FDR	Interaction type
rs6679677	1	114303808	ENSG00000163599	CTLA4	2	204735596	6.1673	A	C	34	29781	6.95E-10	2.041E-05	Trans
rs6679677	1	114303808	ENSG00000111728	ST8SIA1	12	22403341	6.0496	A	C	33	29741	1.45E-09	5.764E-05	Trans
rs6679677	1	114303808	ENSG00000013725	CD6	11	60763482	5.8645	A	C	34	29781	4.51E-09	0.0001048	Trans
rs6679677	1	114303808	ENSG00000134460	IL2RA	10	6078470	5.6158	A	C	33	29741	1.96E-08	0.0002454	Trans
rs6679677	1	114303808	ENSG00000049768	FOXP3	X	49114092	5.5337	A	C	15	9188	3.14E-08	0.0004083	Trans
rs6679677	1	114303808	ENSG00000198821	CD247	1	167443862	-5.4182	A	C	34	29781	6.02E-08	0.0008063	Trans
rs6679677	1	114303808	ENSG00000117090	SLAMF1	1	160597487	5.3696	A	C	34	29781	7.90E-08	0.0009991	Trans
rs6679677	1	114303808	ENSG00000110324	IL10RA	11	117864629	5.0145	A	C	34	29781	5.32E-07	0.0046849	Trans
rs6679677	1	114303808	ENSG00000226167	AP4B1-AS1	1	1.14E+08	-6.0837	A	C	12	4988	1.17E-09	1.313E-05	Cis
rs6679677	1	114303808	ENSG00000134262	AP4B1	1	1.14E+08	-5.5585	A	C	32	28851	2.72E-08	0.0001148	Cis

Table 4-6: HLA SNPs used in the study

HLA region SNPs that are present in the WTCCC cohort data	HLA region SNPs that passed Mann Whitney feature selection	Tissue interactions that are represented in T1D models 1 and 2
rs12153924	rs2187668*	Adrenal_Gland--rs2251396_A--PSORS1C1
rs17211699	rs2251396	Colon_Transverse--rs2251396_A--PSORS1C1
rs1980493	rs3129889	Small_Intestine_Terminal_Ileum--rs2251396_A--MICA
rs2187668	rs9268645	Spleen--rs2251396_A--MICA
rs2251396	rs9272346*	Heart_Left_Ventricle--rs3129889_G--HLA-DRB5
rs3129889	rs9469200*	Skin_Sun_Exposed_Lower_leg--rs3129889_G--HLA-DRB5
rs660895	* SNPs with no eQTL effects	Spleen--rs3129889_G--HLA-DRB5
rs9268645		Testis--rs3129889_G--HLA-DRB5
rs9272346		Thyroid--rs3129889_G--HLA-DRB5
rs9378176		Cells_EBV-transformed_lymphocytes--rs9268645_G--HLA-DQB2
rs9469200		

4.3.5 Regulatory changes in the HLA locus associate to the risk of developing T1D in both T1D model-1 and model-2

The HLA-region is strongly associated with the development of T1D, accounting for 40-50% of the familial aggregation[200]. We observed spatial eQTLs involving HLA (Table 4-6) within T1D models 1 and 2. The significant HLA spatial eQTLs involved SNPs rs2251396-*PSORS1C1*, rs2251396-*MICA*, rs3129889-*HLA-DRB5*, and rs9268645-*HLA-DQB2* and were observed as contributing to the risk of developing T1D in multiple tissues (e.g. adrenal gland, transverse colon, small intestine, spleen, heart left ventricle, sun exposed skin, testis, and thyroid).

4.3.6 Validation of lung cell allele-specific enhancer activity of locus marked by eQTL rs6679677

Our results indicate that eQTL (rs6679677) down-regulates *AP4B1-AS1* transcript levels in the lung. Therefore, we performed a luciferase enhancer assay to experimentally validate that the top ranked eQTL (rs6679677) marks an allele and tissue specific enhancer. DNA sequences flanking rs6679677 (*i.e*. 74bp 5' – ref/alt allele – 75bp 3' [chr1:114303734-114303884; GRCh37]) were cloned into the 3'UTR of a minimal TATA-box promoter and luciferase gene construct to test whether the cloned sequence contain enhancer elements for gene expression (Methods; Figure 4-2)[192]. Transient transfection of the plasmid vector containing the reference locus (*i.e*. the major allele for rs6679677) resulted in a fold increase in luciferase activity when compared to the control vector in A549 (lung) and HepG2 (liver) cells (~11 and ~5 fold increase, respectively). This is consistent with the existence of H3K9ac histone modifications at the locus tagged by rs6679677 in both the lung and liver tissues (see HaploReg; https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php). Notably, a significant allele specific reduction in enhancer activity (*i.e*. nucleotide change from C>A) was observed only in the A549 cells (p = 0.005; Figure 4-11), consistent with the identification of an eQTL involving this locus in the lung but not the liver. Collectively, these results support the allele specific enhancer activity for the locus marked by rs6679677 in the lung.

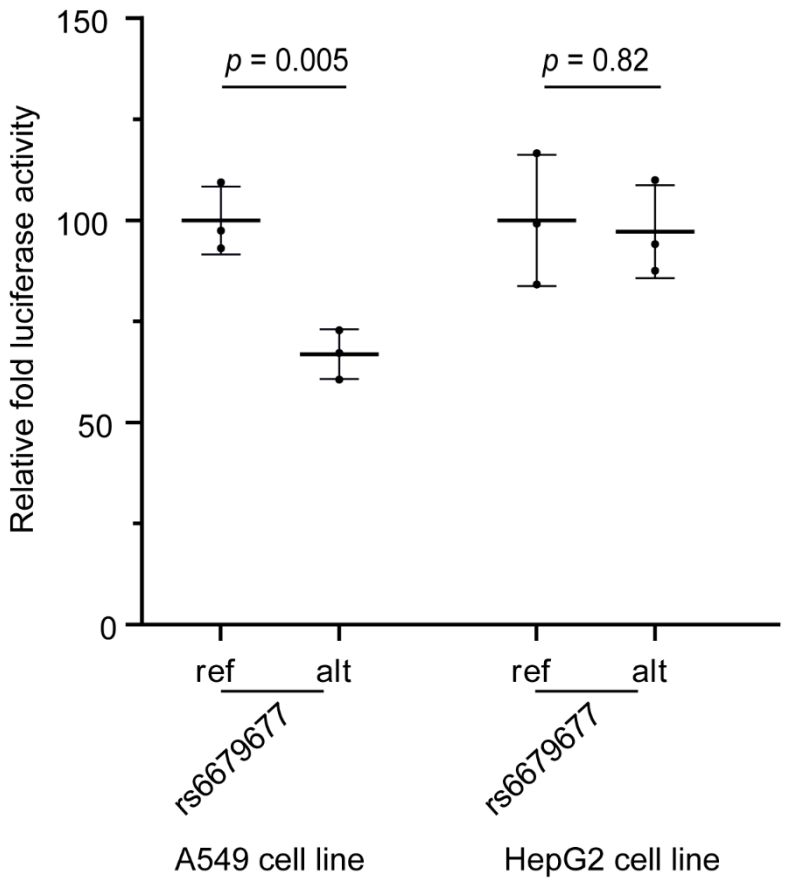


Figure 4-11: rs6679677 is an allele specific enhancer (i.e. nucleotide change from C>A) in lung (A549) but not liver (HepG2) epithelial cells

The locus tagged by rs6679677 was cloned within the 3' UTR of a luciferase gene driven by a minimal promoter and transiently transfected into A549 and HepG2 cells. The relative enhancer activity for the ref and alt versions of the rs6679677 locus was calculated compared to the empty control vector (pMRAdonor2). Relative luminescence units (RLU) for the luciferase assay were normalized using the absorbance values for the beta-galactosidase assay. Results were plotted as the percentage of the ref alleles enhancer activity. Transfection experiments were repeated three times across 3 or 9 technical replicates in A549 (lung) and HepG2 (liver) cells, respectively. Representative results are shown for one transient transfection of A549 and HepG2 cells.

4.4 Discussion

In the present study, we have used a logistic lasso regression model to integrate T1D case and control genotypes with spatial eQTL data to predict the relative tissue-specific contributions for the conversion of genetic risk to T1D. Notably, the predictor models validated across two independent case-control cohorts and identified the lungs and an eQTL involving rs6679677 as the largest single contributor of the risk for developing T1D. These observations are consistent with an environmental event impacting on the largest single interface between the body and environment to precipitate the development of T1D[16,201,202]. Moreover, our results are consistent with the association of respiratory infections (*i.e.* influenza-like illness) with an increased risk of islet autoantibody seroconversion in young-onset T1D study cohorts[16,203]. Collectively, our results demonstrate that tissue-specific approaches can improve our understanding of disease aetiology, potentially aiding therapeutic development in preventing T1D onset.

It is widely recognized that the vast majority of measurable genetic risk for T1D is associated with the HLA region and polymorphisms of the class II HLA DQ, DR, and DP genes[200]. Typically these effects are ascribed to polymorphisms within the HLA genes that affect the shape of the peptide binding grove and the scope of the peptides that can bind to the allele and thus be presented on the cell surface[200]. Our analysis focused on the identification and ranking of the tissue-specific regulatory changes, and thus it does not capture variants that change the function or structural properties of a gene encoded protein. Therefore, it is notable that the HLA locus variants that were retained within our models (rs3129889-*HLA-DRB5,* and rs9268645-*HLA-DQB2)* regulated the expression of DRB and DQB alleles. In addition rs2251396 was associated with expression of *PSORS1C1,* which has been associated with T1D[204]. Similarly, rs2251396 was associated with the expression of *MICA,* which has previously been identified as part of an extended HLA haplotype that associates with TID risk[205], Collectively these observations support the hypothesis that changes in HLA gene expression contribute to T1D risk, in addition to the recognized role for HLA polymorphisms.

In the lung, our eQTL analysis revealed a significant association between rs6679677 and the expression of *AP4B1-AS1*. The allele specific enhancer activity of the rs6679677 tagged region in lung cells was confirmed using a plasmid-based luciferase assay. However, rs6679677 also: 1) has eQTLs with *AP4B1-AS1* and immune regulation genes (i.e. *FOXP3*, *CTLA4*, *IL2RA*, and *SLAMF1*[206–209]) in whole blood; 2) has been associated with the development of multiple persistent autoantibodies (including islet autoantibody), but not progression to T1D development in the TEDDY prospective cohort[167]; 3) has been associated with the development of other autoimmune disorders (i.e. juvenile idiopathic arthritis and rheumatoid arthritis)[210,211]; and 4) was reported as the top non-HLA SNP associated with T1D from WTCCC studies[70]. Collectively, these results support an important molecular role for the locus tagged by rs6679677 in a lung-specific increase in the risk of the development of T1D. However, our results do not prove the effect is exclusively due to the impact on lung cells. As such, future work should dissect if the lung-specific regulatory impact of rs6679677 actually contributes to the mechanism of T1D risk.

AP4B1-AS1 is located in a genomic region that is recognized as being strongly associated with autoimmune disorders[186]. *PTPN22*, which is on the antiparallel DNA strand to *AP4B1-AS1*, encodes a lymphoid-specific intracellular phosphatase (LyP), from the non-receptor class 4 subfamily of the protein-tyrosine phosphatases, that acts as a critical negative regulator of T cell activation and T cell receptor signalling pathways[212]. Notably, rs6679677 down-regulates *AP4B1-AS1* expression in whole blood samples. Similarly, C1858T-rs2476601 (in complete linkage with rs6679677) has been linked with reduced protection against the influenza virus[213]. Therefore, we propose that future studies should ascertain the regulatory roles of rs6679677 on *AP4B1-AS1* in T cells, particularly in response to viral infections in the lungs. We contend that this will help untangle the genetic mechanisms that connect respiratory infections and the induction of islet autoantibodies that has been observed in young children[16,203].

This study has limitations. Firstly, the genetic data used in the analyses are predominantly from people of European ancestry, which limits the immediate translation of our findings to populations with different genetic structures (*e.g.,* variable haplotypes in this region). Secondly, thirty SNPs identified as being associated with T1D did not have identifiable eQTLs in any of the GTEx tissues studied, consistent with the presence of alternative methods or developmental stages through which SNPs can mediate their effects on phenotypes. Thirdly, not all of the eQTLs we identified were represented in the individual genotypes we analysed (*i.e.,* SNPs were unable to be imputed), meaning we could have missed effects. Fourthly, the reporter assay does not take into account the genomic context through which chromatin looping influences the enhancer-promoter interactions that mediate transcriptional activity[214]. Fifthly, the use of a lung cancer cell line may limit the interpretation of the transcriptional control of genes in normal lung tissue. Lastly, most of the spatial chromatin interactions were identified from immortalized cancer cell lines or primary tissues. By contrast, the eQTL associations were obtained mostly from post-mortem samples taken from a cross-sectional cohort (20- 70 years of age). Therefore, it is possible that the Hi-C interactions and eQTL sets were not representative of the tissues in which they were tested. Nevertheless, the reproducibility of the prediction model, across independent cohorts, supports the utility of the approach and its use with expanded datasets for T1D as well as other immune and non-immune diseases.

The novelty of the approach we undertook lies in: 1) the integration of T1D-associated SNPs with their tissue-specific eQTLs (in both cis and trans); 2) interpreting individual case and control genotypes in terms of these tissue-specific eQTL effects; 3) including effects from variants that do not have detectable eQTLs in the reference library that is used in the assay; and 4) the application of machine learning to select and rank the tissue-specific eQTL effects that confer disease risk. This approach moves T1D research away from a candidate gene approach to include gene regulatory changes, including within the HLA locus, as possible contributors to the risk of developing T1D. However, all results are putative until they are followed up by integrative empirical methods that prove the link between gene expression in the lung and other tissues, and the conversion of T1D risk.

In conclusion, our work provides novel insights into the role of variation in gene regulation in the risk of developing T1D. The transcriptional changes (including the changes to *AP4B1-AS1* and *CTLA4*) we identified in the lung may help explain the reported association between respiratory infections and risk of islet autoantibody seroconversion reported in young children.

Chapter 5: Machine learning identifies six genetic variants and alterations in the Heart Atrial Appendage that are important for PD risk predictivity

Parkinson's disease (PD) is a globally prevalent complex disorder of ageing influenced by a range of genetic and environmental causes. Genetic variants have been found to predispose individual risk to develop Parkinson's disease, yet the impact of these variants has not been fully elucidated**.** I hypothesized that genetic variants associated with PD modulate disease risk through tissue specific eQTL effects. In this study, machine learning was used to understand the genetic architecture of PD risk by identifying and ranking the pivotal variants and tissue-specific eQTL effects that contribute to such risk. A regularized logistic regression predictor model for predicting individual PD risk was created from PD related SNPs integrated with the Wellcome Trust Case and Control (WTCCC) genotype and GTEx tissue specific eQTL effect data. The findings reveal the genetic impacts acting through heart tissues to modulate PD disease and provide new insights into the involvement of cardiovascular function in the risk of PD.

5.1 Introduction

Parkinson's disease (PD) is a complex neurodegenerative disease with a range of causes and clinical presentations. The diagnosis of PD is based on the presence of the cardinal motor symptoms, (bradykinesia; muscular rigidity; 4-6 Hz resting tremor; postural instability)[215] which typically manifest many years after initial disease onset. This long period between disease onset and the presentation of noticeable symptoms, known as prodromal PD, provide optimal years for therapeutic intervention[216]. However, diagnosing PD in the prodromal phase remains problematic, with no PD-specific manifestations, and no accurate biomarkers for diagnosis[28]. Genetic/genomic datasets offer an option to help facilitate early detection and stratification of patients, as has been shown for other complex diseases[217,218].

Despite initial conceptions that PD is predominantly a sporadic disease, genome wide association studies (GWAS) have identified human genetic variants that are associated with many complex diseases[40,51,219,220], including PD[33,50]. Chang et al. utilized PD related variants to predict individual PD risk with a PRS model (AUC = 0.652)[221]. In the most recent PD GWAS meta-analysis, Nalls *et al*. identified 90 independent single nucleotide polymorphisms (SNPs) that are significantly associated with PD risk and used for creating a PRS model to predict individual PD risk (AUC = 0.651)[33]. However, given that the majority of the GWAS identified PD SNPs are located in non-coding regions of the genome, understanding the mechanisms by which these SNPs contribute to PD remains challenging[40,41,222]. In order to understand the potential impact of the PD SNPs, they need to be considered in the context of the three-dimensional genome.

Chromatin is organised into cell-type specific, hierarchical, three-dimensional structures within nuclei. These structures emerge from the essential functions that are present within the nucleus, including gene regulation and their responses in different physiological environments[130–133]. Chromosome conformation capture techniques (*e.g.* Hi-C) provide a method to capture interacting loci within the three-dimensional chromatin structure, including interactions that are associated with gene regulation. The physical interactions occurring between the loci and gene provide one method by which regulation of gene expression can occur. Such regulatory loci are known as expression quantitative trait loci (eQTLs), and their target gene, eGene[133,135,162,223,224]. eQTLs typically act in a tissue specific manner, and thus the target eGenes can be used to identify biological pathways that putatively contribute to disease etiology[126,225].

In the past decade, machine learning has been successfully used to develop predictor models for estimating an individual's disease risk[98,100,226]. If PD associated SNPs contribute to disease development through gene regulatory effects, then the tissue-specificity of these eQTLs is important for the aetiology of PD[126,140]. As such, developing a machine-learning predictor model for PD disease status that utilises and selects the informative tissue-specific eQTL data may reveal the most essential SNPs and their tissue-specific regulatory effects that are associated with PD risk.

In this study, we used a matrix of: 1) PD associated SNPs that act as eQTLs, 2) their regulated eGenes and 3) the tissues in which these effects were observed, to build a logistic predictor that was validated using genotype data from three independent studies[50,51,155]. The best PD risk status predictor, with the highest predictive ability, was selected using the WTCCC cohort[50]. The predictor model was then validated using two datasets derived from UK Biobank[155] and NeuroX-dbGap[51]. The top contributors to the PD risk predictor model included six PD SNPs without related known eQTL information and SNP modulated gene regulation specific to the heart atrial appendage.

5.2 Methods

5.2.1 Workflow for developing the PD predictor

We developed a machine learning approach, which incorporates feature selection with cross validation, to calculate the additive tissue-specific contribution of spatial eQTLs within genotypes from individuals who developed PD (Figure 5-1).

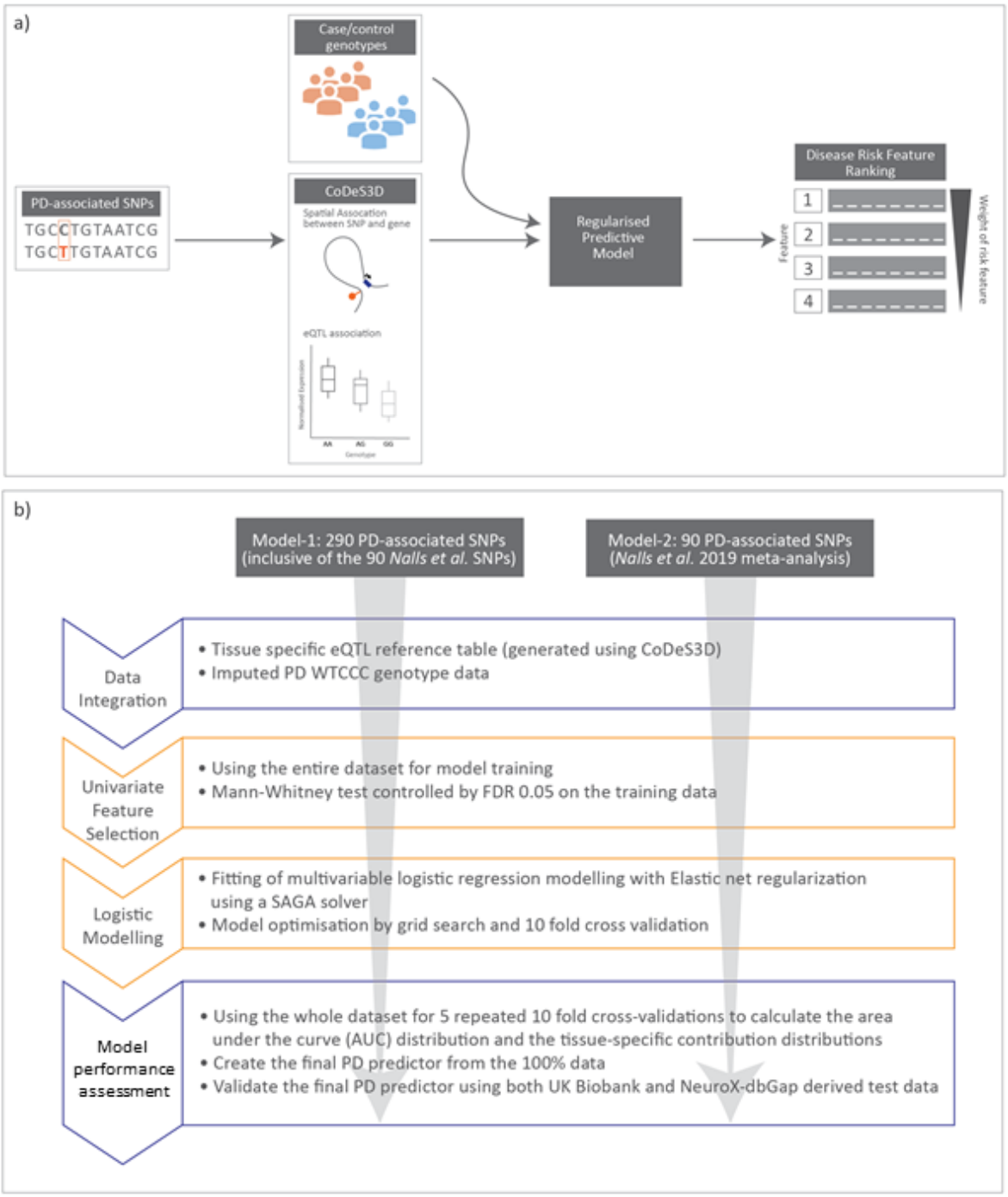


Figure 5-1: Data integration and workflow the regularised logistic regression modelling

a) Schematic diagram for data integration to become disease risk feature ranking. b) Workflow for creating the PD disease status predictor using a regularised logistic regression model.

5.2.2 Generation of tissue specific PD eQTL reference table

SNPs associated with PD (n=290; Appendices: Supplementary Table 10) were obtained from the GWAS catalog (www.ebi.ac.uk/gwas, downloaded 27th August 2020; p-value <1.0 x 10^{-5}). This SNP set included young adult-onset Parkinsonism SNPs[227] and the 90 SNPs identified by Nalls *et al.*[33]

The PD associated SNPs (Table 5-1) were analysed using the Contextualize Developmental SNPs in 3-Dimensions (CoDeS3D) algorithm[222], with the beta effect calculation option (no-afc), to identify: a) the genes that physically interact with the PD associated SNPs; and b) which of these SNP-gene interactions are eQTLs. Physical interactions between PD associated SNPs and genes were identified using Hi-C chromatin contact libraries (Appendices: Supplementary Table 11) captured from:

1. Cell lines from primary human tissues (*e.g.* brain, skin and spinal cord)
2. Immortalised cell lines that represent the embryonic germ layers (*i.e.* HUVEC, NHEK, HeLa, HMEC, IMR90, KBM7, K562, and GM12878)

Table 5-1: 290 PD SNPs used in the study

rs112485576 has been merged with rs504594, and they both in the PD SNP set.

rs504594	rs12637471	rs2042477	rs4130047	rs73656147
rs10121009	rs12726330	rs2086641	rs4140646	rs74335301
rs10221156	rs12817488	rs2102808	rs415430	rs7479949
rs10256359	rs12921479	rs2209440	rs4266290	rs7577851
rs10463554	rs1293298	rs2230288	rs429358	rs75859381
rs10464059	rs12951632	rs2242330	rs4538475	rs76116224
rs10501570	rs12959200	rs2248244	rs4653767	rs7617877
rs10513789	rs1296028	rs2251086	rs4698412	rs76763715
rs10519131	rs13016703	rs2269906	rs4713118	rs76904798
rs10746953	rs13117519	rs2270968	rs4771268	rs76949143
rs10748818	rs13153459	rs2275336	rs4778720	rs7702187
rs10756907	rs13294100	rs2280104	rs4784227	rs77351827
rs10767971	rs1362858	rs2292056	rs4837628	rs78736162
rs10788972	rs138017112	rs2295545	rs4954162	rs78738012
rs10797576	rs1384236	rs2296887	rs4964469	rs7938782
rs10847864	rs141128804	rs2338971	rs5019538	rs79503702
rs10877840	rs141863958	rs2395163	rs55818311	rs7984966
rs10906923	rs14235	rs2414739	rs55961674	rs8005172
rs10918270	rs143918452	rs246814	rs57891859	rs8017172

rs10929159	rs144074972	rs26431	rs5910	rs8070723
rs10958605	rs1442190	rs2694528	rs591323	rs8087969
rs11012	rs144755950	rs2736990	rs601999	rs8118008
rs11026412	rs144847051	rs2740594	rs61169879	rs816535
rs11060180	rs1450522	rs2823357	rs620513	rs8180209
rs11150601	rs1474055	rs2839398	rs62053943	rs823114
rs11158026	rs1480597	rs28903073	rs62120679	rs823118
rs11186	rs148294058	rs2904880	rs62333164	rs823128
rs11248051	rs1491942	rs2921073	rs6416935	rs823156
rs11248060	rs1536076	rs2942168	rs6430538	rs849898
rs112485576	rs1555399	rs3027247	rs6449168	rs850738
rs11343	rs1564282	rs3104783	rs6465122	rs873786
rs113434679	rs162227	rs3129882	rs6476434	rs896435
rs114138760	rs1630500	rs316619	rs6482992	rs9261484
rs11557080	rs16846351	rs329648	rs6497339	rs9267659
rs11578699	rs17000647	rs34025766	rs6500328	rs9275152
rs11610045	rs17115100	rs34043159	rs6532194	rs9275326
rs11658976	rs17425622	rs34311866	rs6532197	rs9323124
rs11683001	rs17497526	rs34372695	rs6599388	rs9356013
rs117073808	rs17565841	rs344650	rs6599389	rs943437
rs11707416	rs17577094	rs34637584	rs6658353	rs9468199
rs11711441	rs17649553	rs34656641	rs666463	rs947211
rs11724635	rs17686238	rs34778348	rs6679073	rs9516970
rs117267308	rs17767294	rs353116	rs6710823	rs9568188
rs117615688	rs17833740	rs35541465	rs67383717	rs959573
rs117896735	rs181609621	rs356182	rs67460515	rs979812
rs11865038	rs183211	rs356203	rs6783485	rs983361
rs11868035	rs1867598	rs356219	rs6808178	rs9858038
rs11931074	rs1879553	rs356220	rs6812193	rs9912362
rs11950533	rs1887316	rs356228	rs6825004	rs9917256
rs12063142	rs188789342	rs35643925	rs6826785	rs997277
rs12147950	rs190807041	rs35749011	rs6854006	rs997368
rs12185268	rs1941184	rs365825	rs687432	
rs1223271	rs1941685	rs3742785	rs6875262	
rs12278023	rs199347	rs3773384	rs7077361	
rs12283611	rs199351	rs3793947	rs7118648	
rs12431733	rs1994090	rs3802920	rs7134559	
rs12456492	rs199453	rs393152	rs7221167	
rs12497850	rs199498	rs3935740	rs7225002	
rs12528068	rs199515	rs4073221	rs72840788	
rs12600861	rs199533	rs4101061	rs73038319	

The potential regulatory effects (beta or NES calculated by no-afc option in CoDeS3D) of the spatial connections were mapped by leveraging the eQTL information from 49 human tissues (Genotype-Tissue Expression database [GTEx] v8; www.gtexportal.org). GTEx derived eQTL significance levels were adjusted for multiple testing [Benjamini–Hochberg FDR][163] and considered significant if $q<0.05$.

5.2.3 PD genotype imputation

The PD genotype dataset was acquired from the Wellcome Trust Case Control Consortium (WTCCC; Request ID 10584). The WTCCC PD genotype dataset, generated using the Illumina microarrays, contained one case cohort (2197 individual samples) and two control cohorts (58C: 2930 individual samples and NBS: 2737 individual samples). Thus, the total number of control samples was 5667, more than double the number of cases. It is likely that the use of imbalanced training data would create biased disease status predictors. Therefore, only the 58C 1958 (British Birth cohort) of control samples was used in this study.

SNPs and individual samples that were of poor quality and were recommended for study exclusion by the WTCCC were removed (Appendices: Supplementary Table 12). SNPs within individual genotypes were converted to dbSNP rsIDs and genomic positions mapped (GRCh37, hg19) by Python scripts. PLINK (v1.90b6.2, 64-bit)[228] was used for quality control. Genotypes were cleaned using the Method-of-moments F coefficient estimate to remove case homozygosity outliers (F values < -0.02 or 0.02 < F values) and the control outliers (F values < -0.016 or 0.19 < F values). Related individuals were identified and removed using proportion IBD (PI_HAT > 0.08). Ancestry outliers (identified by principal component analysis [PCA] plotting), individuals with sex genotype errors (identified by PLINK), or individuals with missing genotype data (missing rate > 5%) were also removed. Finally, SNPs that were not in Hardy-Weinberg Equilibrium ($p < 10^{-6}$) or had a minor allele frequency < 1% were also removed.

The WTCCC PD case and control genotype data were obtained using two different Illumina microarrays (Human670-QuadCustom and Human1-2M-DuoCustom_v1_A)[50]. Therefore, we only used the 526,576 SNPs that were present in both microarrays for imputation. SNP data imputation was performed to recover a total of 27,590,399 SNPs using the Sanger imputation service (https://imputation.sanger.ac.uk), EAGLE+PBWT pipeline[174,175] and Haplotype Reference Consortium(r1.1)[176]. Imputation was performed according to the default instructions (https://imputation.sanger.ac.uk/?instructions=1). Following imputation, PLINK was used to update the genotype data with rsIDs and remove SNPs with an: impute2 score < 0.3; missing data rate > 5%; or those that were not in Hardy-Weinberg Equilibrium ($p < 10^{-6}$). The genotypes for 281 of the 290 PD SNPs used in this study (Table 5-1) were extracted from the imputed PD genotype data.

5.2.4 Creation of a weighted WTCCC PD genotype eQTL matrix

We created a matrix that combined each individual's genotype with the eQTL effects for the PD associated SNPs. There were three groups of data fields in the PD genotype eQTL table:

1. Individual sample information (sample id, sex, and disease status)
2. Individual sample PD associated SNP genotype (SNP minor allele count) weighted by GTEx tissue-specific eQTL normalised effect sizes
3. Individual PD associated SNP genotype for the SNPs with no eQTL effect information

The tissue-specific eQTL normalised effect size (NES) for the PD associated SNPs were extracted from the GTEx eQTL summary table of significant eQTLs (Appendices: Supplementary Table 13). The NES for each tissue-specific eQTL was weighted by the number of alternative alleles (0, 1 or 2) at the eQTL SNP position in each individual's genome. 54 of the 290 PD associated SNPs had no identifiable eQTL effects (Appendices: Supplementary Table 14) and were input into the model unweighted, using solely SNP allele count from the imputed genotype.

We created two regularised logistic regression models (see below): for PD model-1, we created a weighted WTCCC PD genotype eQTL matrix for all 290 SNPs that were imputed and represented in the PD eQTL reference table. By contrast, for PD model-2, we created a weighted WTCCC PD genotype eQTL matrix for the subset of 90 SNPs from the PRS analysis in Nalls *et al*.[33]

5.2.5 Generation, training, and validation of the regularised logistic regression models (PD model-1 and PD model-2)

We developed a regularised logistic regression predictor that incorporated: 1) a Mann-Whitney U tests in combination with Benjamini-Yekutieli procedure for controlling false discovery rate (FDR) to identify relevant information[158,163,190]; and 2) a multivariate prediction step that considers all features in context, and removes redundant information, to identify the best combination of features for prediction of PD. Regularised logistic regression was incorporated into the models to enable the features that contribute to the final score to be identifiable[67].

The weighted WTCCC PD genotype eQTL matrix that contained all the case and control genotypes that passed quality control (4366 individual samples: 1698 cases and 2668 controls) was used to train PD model-1 (using all 290 SNPs), or PD model-2 (using the 90 Nalls *et al.* SNPs).

The Mann-Whitney U test with the BY control[163] (tsfresh version 0.16.0)[190] was used to select the individual feature columns within the full training dataset that were the most relevant attributes for predicting PD status (*i.e*. the relevant attribute subset; FDR = 0.05)[191]. The relevant attribute subset was then used to train a multivariate logistic regression model (Scikit-learn version 0.23.2)[183,184] implemented with elastic net regularisation using the SAGA solver to predict PD disease status. The machine learning elastic net regularisation prevented overfitting the predictor model by further sub-selecting the essential features for delivering the best prediction.

The training was optimised (measured by area under the receiver operating characteristic curve [AUC])[81] using a Scikit-learn Grid Search algorithm[93,184] with 10-fold cross-validation setting to select the optimised model hyperparameters from the training stage (90% of the cohort used for training, 10% used for cross-validation). The optimised hyperparameters for PD model-1 were: C=0.5, l1_ratio=0.6, max_iter=800, penalty='elasticnet‘, random_state=1, solver='saga‘ from the search space of following:

- 'C': 0.01, 0.05, 0.1, 0.5,1, 10, 20, 30,
- 'max_iter': 200, 500, 800, 1000, 1200, 1400, 1500, 1600,
- 'l1_ratio': 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1.

The optimised hyperparameters for PD model-2 were: C=0.6, l1_ratio=0.1, max_iter=130, penalty='elasticnet‘, random_state=1, solver='saga‘ from the search of following:

- 'C': 0.0001, 0.001, 0.01, 0.1,0.2,0.3,0.4,0.5,0.6,0.7,0.8, 1, 3,
- 'max_iter': 1, 5, 70, 100, 130, 150, 170, 180, 200, 300, 500, 1000, 1200, 1400, 1600, 1800, 2000, 2200, 2400, 2600, 3000,
- 'l1_ratio': 1, 0.9, 0.8, 0.7, 0.6, 0.5, 0.4, 0.3, 0.2, 0.1.

‘C’ is the inverse of regularization strength that must be positive and specify stronger strength with smaller values[183,184]. ‘solver’ is the algorithm used in the optimization. ‘max_iter’ is the maximum number of iterations for the solvers to converge[183,184]. ‘penalty’ is the penalty function[183]. ‘l1_ratio’ is the elastic net regularization strength[183,184]. When l1_ratio = 0, the regularization is equivalent to L2 regularization[183,184]. When l1_ratio = 1, the regularization is equivalent to L1 regularization[183,184].

The search space of model-1 and 2 did not include l1_ratio = 0 for excluding L2 regularization and implementing feature selection. To calculate the variation in AUCs of the models with the optimised parameters; we undertook 5 repeats of 10-fold cross-validation of model generation and validation by Scikit-learn RepeatedKFold algorithm[183,184]. The 10-fold cross-validation started with the random generation of 10 equal parts from the full dataset. Nine parts of the data were used for training, and the remaining data were for validation. Mann Whitney U test[158] filtering controlled by FDR = 0.05 with the BY procedure[163] was applied to the training set. Subsequently, the filtered training data were modelled by the multiple regularised logistic regression algorithm with the optimised hyperparameters of (PD model-1 or PD model-2). This process was repeated until all parts of the data were used for validation.

5.2.6 Calculation of tissue-specific contributions to PD risk

The 50 PD regularised logistic regression predictors created from the 5 repeats of 10-fold cross-validation above were used to test the predictive power of the model created with the optimised hyperparameters from the tissue-specific eQTL effects. Tissue-specific contributions to the PD risk were extracted from each of the 50 PD regularised logistic regression predictors as the sum of the absolute values of the model weights associated with each tissue.

5.2.7 Validation of PD model-1 and PD model-2

PD models-1 and -2 were validated by two independent test datasets derived from the genotype data of UK Biobank and NeuroX-dbGap for testing the generalising PD predictive power. Since the UK Biobank only has a small number of PD case samples, we created 30 different test cohorts of individual samples (without missing data for those SNPs included in PD model-1 or 2) using the same PD cases with 30 independently and randomly chosen non-PD diagnosed controls. Each cohort was used to create a weighted eQTL-genotype matrix for testing the predictive power with AUC. The mean AUC of the 30 predictive tests was used as the validation result of the models. NeuroX-dbGap was the largest PD single array study[33,51,178], and we selected all the PD case and control samples of NeuroX-dbGap (without missing data for those SNPs included in PD model-1 or 2) to build a weighted eQTL-genotype matrix for validating the PD predictive power of each model.

PD model-1 was validated using the following two independent datasets: 1) 30 cohorts of 2384 individual samples (928 cases and 1456 controls) derived from the UK Biobank; and 2) the 5,224 cases and 5,563 controls from NeuroX-dbGap (the largest PD single array study)[33,51,178].

PD model-2 was also validated by the same two independent datasets: 1) the 30 cohorts of 3812 individual samples (1484 cases and 2319 controls) derived from the UK Biobank; and 2) the cases and control from the NeuroX-dbGap dataset (5,224 cases and 5,563 controls) [33,51,178]. Note the number of cases in the UK Biobank differed as there were fewer cases excluded due to missing data (see below).

Genotypes that were used from the UK Biobank (BGEN format dataset) were selected as follows. European Caucasian samples identified by genetic clustering methods were selected from the UK Biobank and imputed (487,411 individual samples). The genomic analysis and relatedness analysis excluded SNPs that the UK Biobank recommended were removed from the selected case and control data.

The cases (model-1: 928 cases or model-2:1484 cases) were selected using the following criteria:

1. PD patients identified by the UK Biobank developed algorithm (field 42033)
2. PD patients identified by hospital records G20
3. PD patients had no missing data for any SNPs within the predictor model (model-1 or model-2). The greater number of SNPs used in PD model-1 meant that more cases were excluded due to missing data.

Control genotypes, not having records of Parkinsonism and without missing data for any of the SNPs included in the final predictor, were randomly selected from the healthy controls within the UK Biobank data for each of the 30 test cohorts.

Genotype data of the UK Biobank case and control samples in each test cohort were used to build a weighted eQTL-genotype matrix for testing PD model-1 or 2 to recognise the disease statuses of individual samples correctly.

Genotypes were also obtained from the NeuroX-dbGap dataset. Genotypes were cleaned by removing all insertion and deletion variants. SNP IDs were converted to dbSNP rsIDs. Variants in chromosome 24(Y), 25(XY) and 26(MT) that are not included in the study due to the inconsistency with the Sanger imputed SNP data were also removed. Ancestry outliers (identified by principal component analysis [PCA] plotting), individuals with sex genotype errors (identified by PLINK), or individuals with missing genotype data (missing rate > 5%) were also removed. Finally, SNPs that were not in Hardy-Weinberg Equilibrium ($p < 10^{-5}$) or had a minor allele frequency < 1% were removed. All the variants in the final model (model-1or model-2) which were not present in the NeuroX-dbGap data were replaced with proxy SNPs using *linkage disequilibrium* information ($r^2 > 0.5$)[178] calculated by PLINK from European 1000 genome genotype data (https://www.internationalgenome.org/about)[39]. The European 1000 genome genotype data were downloaded from (https://ctg.cncr.nl/software/magma)[229] on 10th August 2020.

5.2.8 Mann-Whitney U test filtering on 290 PD and 313 T1D SNPs derived eQTL matrix

We have previously developed regularised logistic regression models for Type 1 Diabetes (T1D; see chapter 4). We used the 313 T1D SNPs (Appendices: Supplementary Table 1) and their GTEx eQTL summary table of significant eQTLs (Appendices: Supplementary Table 3). We mixed the 290 PD and 313 T1D SNPs to create a weighted WTCCC PD and T1D genotype eQTL matrix, as outlined above. Mann-Whitney U test filtering (controlled by FDR = 0.05 with the BY procedure[163]) was applied to the weighted WTCCC PD and T1D genotype eQTL matrix to determine the filtering power for removing non-related PD features.

5.2.9 Data analysis

All statistical tests were performed with Scikit-learn (version 0.23.2)[183,184], and tsfresh (version 0.16.0)[190]

5.2.10 Code Availability

The CoDeS3D pipeline is available at: https://github.com/Genome3d/codes3d-v2.
The Python scripts and machine learning code used in this analysis are available at: https://github.com/Genome3d/PD_lg_predictor_analysis
Python version 3.7.3 was used for all the python scripts.

5.3 Results

5.3.1 PD associated SNPs act as tissue specific eQTLs for 1,334 eGenes

We hypothesised that PD SNPs modulate disease risk through tissue-specific eQTL effects[126,140]. We therefore analysed 290 PD GWAS associated SNPs (Table 5-1) for spatial eQTL interactions[131,133,137] across 49 GTEx tissues[126]. 231 of the 290 (79.7%) PD SNPs tested were involved in 18,041 tissue-specific eQTL associations (Benjamini–Hochberg FDR < 0.05 [163]; Appendices: Supplementary Table 15), regulating 1,334 eGenes across the 49 GTEx tissues. Gene ontology analysis (David Functional Annotation)[230] identified that the regulated genes were enriched for intracellular signal transduction and antigen processing & presentation of peptides (Appendices: Supplementary Table 16).

5.3.2 Modelling genotype data to identify the genetic risk associated with tissue-specific eQTL effects for PD disease status

Understanding the impacts and complex networks associated with eQTLs is challenging. We hypothesised that regularised logistic regression models could be used to identify and rank the eQTLs that were significant contributors to PD risk.

We integrated the CoDeS3D eQTL analysis of the 290 PD SNPs with the genotype data for individuals within the Wellcome Trust Case and Control Consortium (WTCCC)[70] PD cohort (4366 individual samples: 1698 cases and 2668 controls; methods)[50]. Of the 290 PD SNPs retrieved from the GWAS catalog, 281 SNPs were present in the WTCCC data. This resulted in the generation of a PD SNP derived eQTL effect matrix containing 17,829 tissue-specific eQTL-eGene pairs (227 SNPs, 1310 eGenes, 49 tissues) and 54 SNPs that had no known eQTL effects following our CoDeS3D analysis. Uninformative features for PD prediction were removed using a Mann-Whitney U test[158] (FDR < 0.05 with the BY procedure[163]) (Methods). After filtering, 11,288 PD SNP derived features (53 SNPs, 245 eGenes, 49 tissues) remained within the PD variant derived eQTL effect matrix.

To test the effectiveness of the Mann-Whitney U test[158] with the BY control[163] filter, we generated a PD and type 1 diabetes (T1D) SNP derived eQTL effect matrix(Ho et al.) using a mixed set of 290 PD and 313 T1D associated SNPs and integrating with the WTCCC PD cohort genotypes (Appendices: Supplementary Table 1; Table 5-1). The PD + T1D SNP derived tissue-specific eQTL effect matrix included 25,052 SNP related data fields (556 SNPs, 1927 eGenes, 49 tissues). After the Mann-Whitney U test filtering (FDR < 0.05 with the BY procedure[163]), 11,147 of the data fields (45 SNPs, 209 eGenes, 49 tissues) were selected using PD as the phenotypic outcome. Only one T1D-associated SNP, rs1052553, remained following the Mann-Whitney U test with the BY control filtering. Although rs1052553 has not previously been associated with PD in GWA studies, it has been implicated in PD as part of a PD risk haplotype[231,232]. Therefore, these results confirm that the Mann-Whitney U test with the BY control filters uninformative data while preserving valuable PD information for our modelling.

We created regularised logistic regression models for PD risk using the Mann-Whitney U test with the BY control filtered PD variant derived eQTL effect matrix (11,288 variant derived features). The AUCs of 50 logistic regression predictor models had a mean of 0.565 (distributed from 0.516 to 0.637) and a standard deviation of 0.024 (generated with the optimised predictor model hyperparameters by 5 repeats of 10-fold cross validation). The final PD predictor model (for PD model-1) was created with the optimised predictor model hyperparameters using 100% of the WTCCC PD cohort for training. After the Mann-Whitney U test with the BY control filtered WTCCC PD variant derived eQTL effect matrix contained 17,829 variant derived features. PD model-1 selected for 827 tissue-specific eQTLs and 6 SNPs with no eQTL effect (Appendices: Supplementary Table 17). PD model-1 had an enhanced diagnostic ability as represented by an in-sample (training data) AUC of 0.627.

We validated the predictive power of PD model-1 using two independent PD cohorts (UK Biobank[155] (30 datasets of 923 cases and 1456 controls) and NeuroX-dbGap[51,178]). PD model-1 was validated in both cohorts, producing a mean AUC of 0.572 (distributed from 0.555 to 0.587), and an AUC of 0.571 in the UK BioBank and NeuroX-dbGap cohorts respectively. These two validation results are highly consistent and within the range of the model AUCs estimated by the 50 optimised predictor models created from the randomised WTCCC derived training data using the same hyperparameters as PD model-1.

5.3.3 eQTLs specific to the heart atrial appendage contribute to genetic risk in PD

PD model-1 was used to rank the genetic elements that associate with PD disease risk. In this analysis, we used the magnitude of the model weights (coefficients) for the genetic features, grouped by tissue-specificity of the effects, in the logistic regression PD model-1 as proxies for the contribution of the features to PD risk. Six SNPs that had no known eQTL effects (from CoDeS3D analysis of GTEx) made the most significant group contribution (18% of the total model weight) to the risk of PD development (Table 5-2). The non eQTL SNP set contained rs117896735, rs144210190, rs35749011, rs12726330, rs356220 and rs5019538 (Table 5-2). Note that rs356220 and rs5019538 were removed from the tissue-specific eQTL data of the GTEx study[126] due to QC processing, and therefore we were unable to test if these SNPs were eQTLs. rs117896735 also has no eQTL effect information found in the GTEx database. The other three SNPs rs144210190, rs35749011 and rs12726330 were not detected by CoDeS3D to have spatial eQTL and eGene interactions. eQTLs that affected the Heart Atrial Appendage (9%) and Brain Cerebellum (4%) made the next biggest contributions to the risk of PD development (Figure 5-2). Conversely, eQTL gene regulation specific to the Substantia Nigra contributed ~1.5% of the risk of PD development. The tissue-specific contribution ranking obtained from the 50 optimised predictor models, generated with model-1's hyperparameters by 5 repeats of 10-fold cross validation (randomizing the full Mann-Whitney U test with the BY control filtered PD variant derived eQTL effect matrix), were consistent with these findings, identifying the SNPs without eQTL effects, Heart Atrial Appendage, and Brain Cerebellum as the top three genetic contributors to the risk of PD development (Figure 5-3).

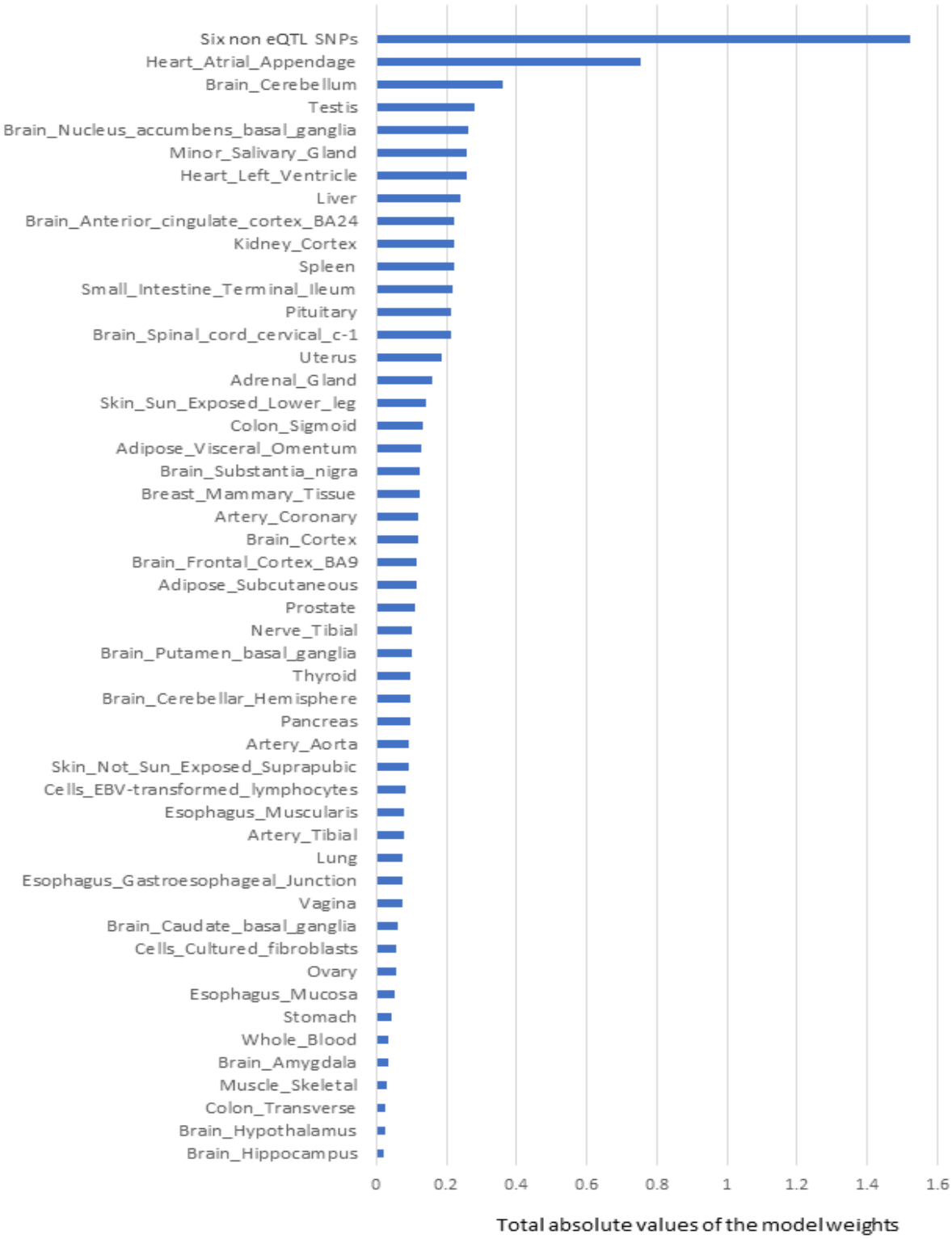


Figure 5-2: The rank order of tissue-specific risk contributions to risk of developing PD calculated using PD model-1

Tissue PD risk contributions were the sum of the absolute values of the model weights (coefficients) of the features used in the logistic regression predictor (PD model-1) according to their tissues. The SNPs/eQTLs that contributed to each category are listed (Table 5-2).

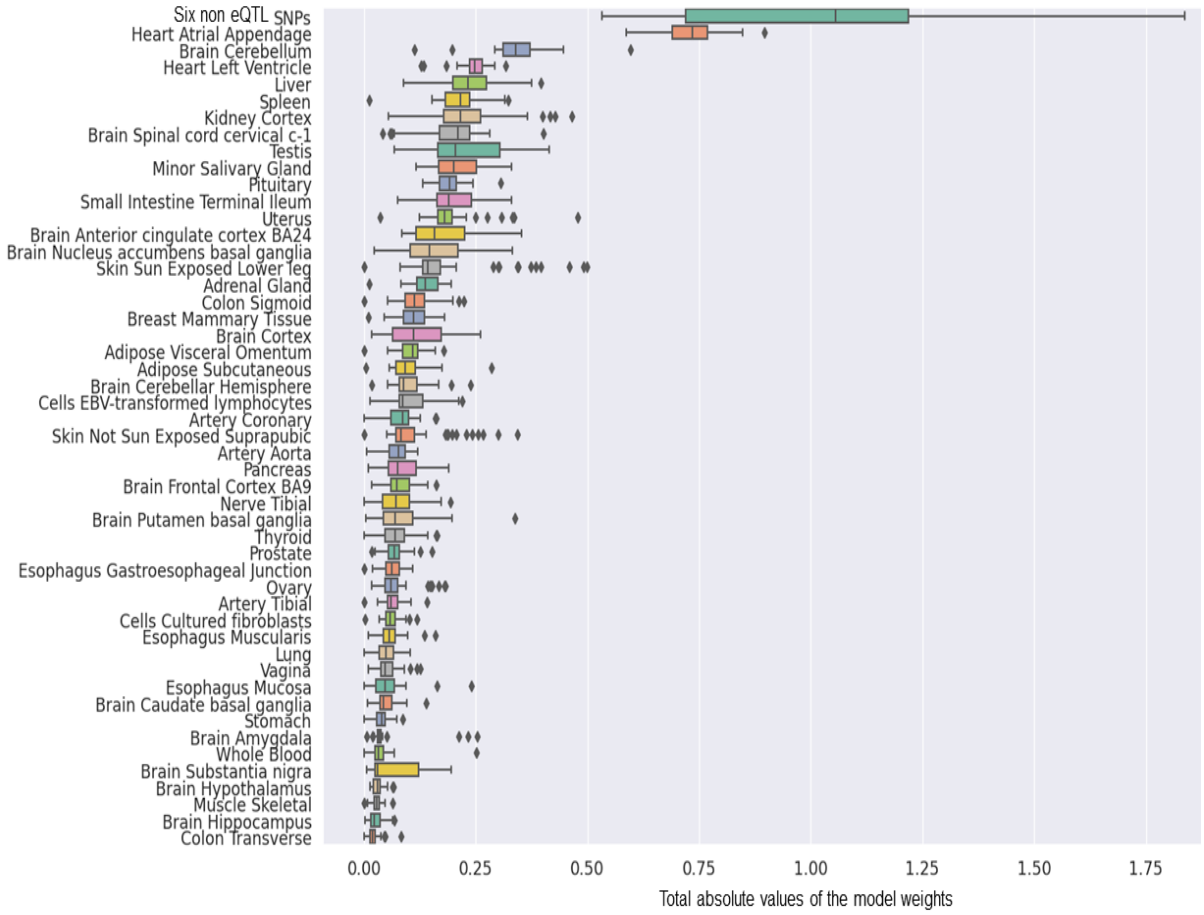


Figure 5-3: The rank order of tissue-specific risk contributions calculated across 50 predictor models created from randomised modelling and PD model-1's hyperparameters

The tissue ranking was consistent with that observed for PD model-1.

Fifteen eQTLs impacted on the Heart Atrial Appendages contribution to the risk of developing PD measured in model-1 (Table 5-2). However, SNPs rs7617877 and rs6808178 each contributed approximately 3% to the total weights calculated by model-1. rs7617877 and rs6808178 are in high linkage disequilibrium (R^2 = 0.86)[233] within European populations. rs7617877 and rs6808178 do not regulate their nearest genes and instead act as eQTLs for a gene > 13Mb downstream, *EAF1-AS1*, in the Heart Atrial Appendage. *EAF1-AS1* is a long antisense non-coding RNA gene that undergoes an isoform switch and has a significantly different transcript usage in the brains of patients with Parkinson's disease[234]. rs6808178 also acts as an eQTL for *TMEM161B-AS1* transcript levels in the Heart Atrial Appendage. Notably, there is evidence indicating a possible role for *TMEM161B-AS1* in neurodegeneration[235].

Table 5-2: SNP and eQTL-gene contributors to the impact of the SNP set and Heart atrial appendage on PD model-1

SNP and eQTL-gene contributors to the impact of the SNP set and Heart atrial appendage, respectively, on the final PD logistic regression predictor (PD model-1). The model weight is the coefficient assigned to each variant or eQTL in the logistic regression predictor model. Abs (model weight) indicates the absolute value of the model weight. '*' indicates the non eQTL SNP is in the 90 SNPs of Nalls *et al.*

SNPs without detected eQTLs	model weight	abs(model weight)
*rs117896735_A	0.42436	0.42436
rs1442190_A	0.40106	0.40106
*rs35749011_A	0.24949	0.24949
rs12726330_A	0.18701	0.18701
rs356220_T	0.17507	0.17507
*rs5019538_G	-0.08418	0.08418

Heart_Atrial_Appendage	SNP (rsID_major allele)	eGene	model weight	abs(model weight)
Heart_Atrial_Appendage	rs7617877_A	EAF1-AS1	0.28996	0.28996
Heart_Atrial_Appendage	rs6808178_T	EAF1-AS1	0.25261	0.25261
Heart_Atrial_Appendage	rs6808178_T	TMEM161B-AS1	0.16339	0.16339
Heart_Atrial_Appendage	rs11707416_A	P2RY12	0.01163	0.01163
Heart_Atrial_Appendage	rs26431_G	EIF3KP1	0.0089	0.0089
Heart_Atrial_Appendage	rs17577094_G	RP11-259G18.3	0.00703	0.00703
Heart_Atrial_Appendage	rs365825_G	RP11-259G18.3	-0.00467	0.00467
Heart_Atrial_Appendage	rs17577094_G	LRRC37A4P	-0.00434	0.00434
Heart_Atrial_Appendage	rs17577094_G	KANSL1-AS1	0.0036	0.0036
Heart_Atrial_Appendage	rs8070723_G	RP11-259G18.3	-0.00294	0.00294
Heart_Atrial_Appendage	rs365825_G	LRRC37A4P	0.00237	0.00237
Heart_Atrial_Appendage	rs365825_G	KANSL1-AS1	-0.00128	0.00128
Heart_Atrial_Appendage	rs17577094_G	DND1P1	0.00116	0.00116
Heart_Atrial_Appendage	rs17577094_G	MAPK8IP1P2	0.00058	0.00058
Heart_Atrial_Appendage	rs199515_G	RP11-259G18.3	-0.00011	0.00011

5.3.4 Creating a PD logistic regression predictor model using the 90 SNPs of Nalls *et al.*

Nalls *et al.* identified 90 SNPs that contribute to a PRS model for PD risk[33]. We, therefore, sought to understand the PD risk contribution in our model that was specific to these 90 SNPs and thus created a separate logistic regression predictor model using only this subset. 88 of the 90 variants passed quality control (post-imputation data cleaning and quality checking). The 88 SNPs were integrated with the WTCCC PD genotype data to create a PD SNP derived eQTL effect matrix of WTCCC individual samples (4,366 individual samples: 1,698 cases and 2,668 controls). The PD SNP derived eQTL effect matrix contained 3,206 features consisting of related tissue-specific eQTL-eGene pairs (76 SNPs, 518 genes, 49 tissue types) and 12 SNPs that lacked CoDeS3D detectable eQTL effects. Mann-Whitney U test filtering (FDR < 0.05 with the BY procedure) left 920 features (12 SNPs, 95 genes, 49 tissue types) that were used in the subsequent logistic regression modelling[158,183,190]. Model training was repeated using the optimised hyperparameters and the eQTL effect matrix for the full WTCCC cohort to create predictor PD model-2. PD model-2 achieved in-sample PD prediction with an AUC = 0.604 using 311 features (12 SNPs, 46 genes, 49 tissue types) (Appendices: Supplementary Table 18) that included 308 tissue-specific eQTLs and 3 SNPs without known eQTL effects. PD model-2 was validated using the UK Biobank[155] (AUC = 0.554) and NeuroX-dbGap[51,178] (AUC = 0.568) genotype data.

We determined the tissue-specific distribution for the 50 predictors that were created with PD model-2's hyperparameters. The results we observed were consistent with what we observed using PD model-1 (Figure 5-4). Specifically, three SNPs (rs117896735, rs35749011 and rs5019538) with non-identifiable eQTL effects (Table 5-3) and the eQTLs within the Heart Atrial Appendage were the top contributors to the risk of developing PD (Figure 5-4 and Table 5-3). The three non-eQTL SNPs in the 90 SNPs were observed to have similar effect sizes (both magnitude and direction) as in the six non-eQTL SNPs in PD model-1. Also consistent with PD model-1, PD model-2 identified rs6808178 as the top eQTL contributing to the Heart Atrial Appendage signal.

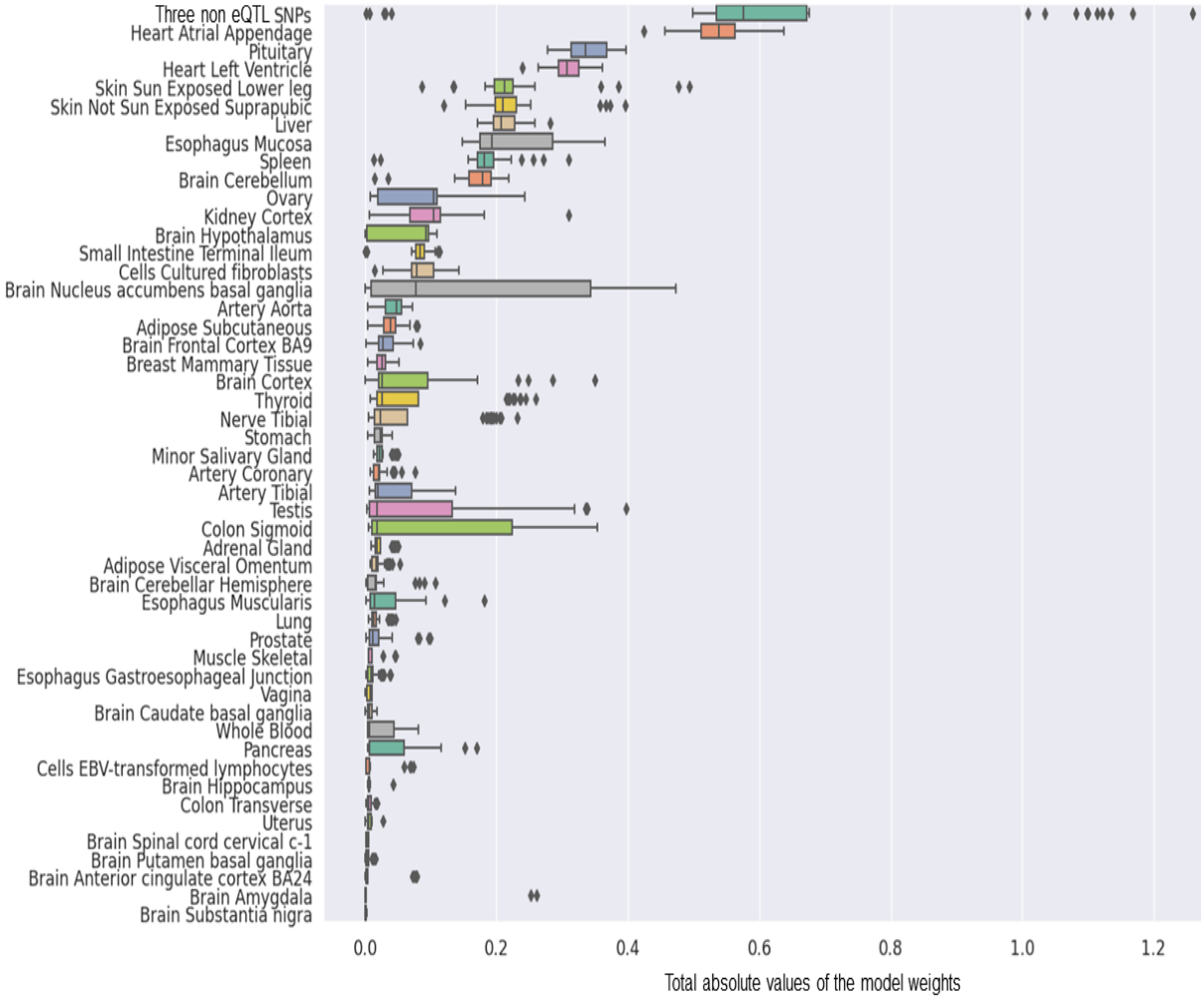


Figure 5-4: The group contributions of 50 predictors created with PD model 2 hyperparameters by 5 repeats of 10 fold cross-validation

Table 5-3: The variants (with known eQTL effects) and eQTLs (Heart Atrial Appendage) of the final PD logistic regression predictor (PD model 2)

SNPs without detected eQTLs	model weight	abs(model weight)
rs117896735_A	0.54172	0.54172
rs35749011_A	0.47224	0.47224
rs5019538_G	-0.04028	0.04028

Heart_Atrial_Appendage	SNP (rsID_major allele)	eGene	model weight	abs(model weight)
Heart_Atrial_Appendage	rs6808178_T	EAF1-AS1	0.53154	0.53154
Heart_Atrial_Appendage	rs26431_G	EIF3KP1	0.0079	0.0079
Heart_Atrial_Appendage	rs504594_A	HLA-DQA2	-0.00333	0.00333
Heart_Atrial_Appendage	rs62053943_T	RP11-259G18.3	-0.0014	0.0014
Heart_Atrial_Appendage	rs62053943_T	DND1P1	-0.00092	0.00092
Heart_Atrial_Appendage	rs62053943_T	KANSL1-AS1	-0.00064	0.00064
Heart_Atrial_Appendage	rs62053943_T	LRRC37A4P	0.00061	0.00061

5.4 Discussion

The mechanisms by which PD-associated genetic variants[40,51,236,237] contribute to disease risk and development have not been fully elucidated. Yet, it is critical that we identify the mechanisms by which they impact on PD because this will allow patient stratification and the development of therapeutics that target disease progression and not just pathology. We used machine learning to understand the genetic architecture of PD risk, by identifying and ranking the pivotal risk variants and tissue-specific eQTL effects that contribute to such risk.

Curated PD-associated SNPs from the GWAS catalog[52] were analysed to identify their tissue-specific eQTL effects. Regularised logistic regression predictor models that evaluated PD risk were built and validated across three independent case and control cohorts[50,51,155]. PD model-1 achieved superior predictivity in comparison to PD model-2, and delivered an in-sample predictive AUC = 0.627, and was subsequently validated in two independent test datasets derived from the UK Biobank[155] (AUC = 0.572) and NeuroX-dbGap[51] (AUC = 0.571). Although greater PRS predictivity has been achieved for PD by other groups, our main aim was to determine the SNPs-genes-tissue combinations that have the greatest contribution. PD model-1 (generated from 290 SNPs) identified 6 SNPs without known eQTL effects and the SNP modulated gene regulation within the Heart Atrial Appendage as being the major contributors to the predicted risk of developing PD. A second model (PD model-2) that was generated using only 90 SNPs[33] (which were previously identified to have the greatest predictive power with a PRS analysis) confirmed a subset of the top predictors we observed with PD model-1. Collectively, our results confirm roles for SNPs that are significantly connected with *INPP5P*, *CNTN1*, *GBA* and *SNCA* in PD and separately suggest a key role for transcriptional changes within the heart atrial appendage in the risk of developing PD. Effects associated with eQTLs located within the Brain Cerebellum were also recognized to confer major PD risk in the more extensive model (PD model-1), consistent with current hypotheses suggesting the Brain Cerebellum plays a role in PD development[238–240].

For the top six contributing SNPs to the model, our analyses did not identify any spatial eQTL interactions. However, previous research has shown connections between these SNPs and three well-known PD-associated genes (*INPP5F, GBA, SNCA*)[30,31,241,242], and an additional gene (*CNTN1*). rs117896735, the top contributor to PD model-1, is an intronic variant of *INPP5F* and has previously been identified as eQTL for *INPP5F* transcript levels (the IPDGC locus browser[243]). *INPP5F* is a known risk gene for PD[31] that regulates STAT3 intracellular signalling pathways[244] and has functional roles in cardiac myocytes and axons[245,246]. rs1442190 is an intronic variant within *CNTN1*, a known risk gene for dementia with Lewy bodies[247,248] that encodes a cell adhesion protein, which is important for axon connections and nervous system development[249]. rs35749011 and rs12726330 are linked to the well-known PD-associated gene *GBA*[241] through strong linkage disequilibrium connections ($R^2 = 0.77$ [233]) with rs2230288[241,250], a missense coding variant located within *GBA*. rs35749011 has eQTL effects on *GBA* identified by the IPDGC database[243].

The final two SNPs, rs356220 and rs5019538, are located downstream of *SNCA*. *SNCA* encodes α-synuclein, which is central to PD pathogenesis[242]. The IPDGC database[243] indicates that rs5019538 has eQTL effects on *SNCA*. Notably, rs356220 had the strongest association to PD in the original WTCCC GWAS[50]. Therefore, there is sufficient evidence that has previously associated these six variants with PD through connections to PD risk genes.

5.4.1 Allele-specific regulatory changes in the heart atrial appendage conferring PD risk

We also identified that eQTLs specific to the heart atrial appendage make a reproducible and substantial contribution to the risk of developing PD. There is an increasing research literature that is indicating a close relationship between cardiovascular health and PD development[251–256]. Notably, the use of antihypertensive drugs and physical exercise both significantly lower PD risk[255,256]. Studies of PD patients have also identified abnormal blood flow patterns in brains[257], with atrial fibrillation (AF) being strongly related to early-stage PD[251]. Moon et al. identified that patients with PD have an increased risk of AF, with a threefold increased risk (HR: 3.06, 95% CI: 1.20-7.77) of AF in younger PD patients (age: 40-49 years)[258]. The heart atrial appendage is a trigger site of AF[259] and highly associated with hypertension and stroke[260–263]. Finally, it is argued that the perturbation of the brain blood supply networks by AF promotes tissue inflammation and damage[264], leading to PD pathogenesis.

Amongst the 15 eQTL features that combined to make the Heart Atrial Appendage's contribution to the risk of developing PD, the up-regulation of *EAF1-AS1* (a long non-coding mRNA) made the greatest contribution. Elevated *EAF1-AS1* transcript levels have previously been identified by differential gene expression analyses in brain tissue samples from PD patients[234]. It is interesting to speculate that the impact of this change is mediated through the interaction of *EAF1-AS1* with *EAF1*. Notably, *EAF1* has been associated with both neural development[265] and TGF-β signalling[266], which is a key pathway in many cardiac physiological processes[267]. As such, the deregulation of *EAF1-AS1* might impact on cardiac health. We propose that future studies should investigate the regulatory impacts of *EAF1-AS1* on *EAF1* and the consequences of alterations in expression levels on heart function and PD disease. We contend that understanding this relationship may help to decipher the complex interactions connecting cardiovascular fitness and PD pathogenesis.

Similar to our work, Li *et al.*[268] used LD score regression (LDSC) analysis[145,146] to identify enrichments of PD risk signals in six GTEx[126] central nervous system tissue types. However, three subsequent studies using LDSC have failed to reproduce Li *et al.*'s results[144,269,270]. LDSC focuses on measuring the risk enrichment of genes uniquely expressed in each GTEx tissue[145,146]. In contrast, the advantage of our model is that it does not contain this constraint and instead identifies the risk associated with the expression of all genes modulated specifically by PD SNPs in different or multiple GTEx tissues. We therefore hypothesise that the fact that Li *et al*. did not identify any signals in heart tissues is likely due to the differences of starting presumptions and employed methodologies.

5.4.2 PD models 1 and 2 identify the same contributors to PD

Comparing the results from PD model-1 and 2, which were generated using 290 and 90 PD SNPs respectively, yielded informative findings. The higher predictive power observed for PD model-1 (Appendices: Supplementary Table 17) can potentially be explained by the fact that the final model included more features (827 vs 308). However, given that PD model-1 leveraged 290 PD associated SNPs to start, the result also suggests that the 90 SNPs, originally identified as part of the Nalls *et al.* PRS analysis[33], do in fact contain the major genetic components that are associated with the risk of developing PD. The finding that both models consistently identified the same SNPs and heart atrial appendage eQTLs as the top contributors to the risk of developing PD further confirms the significance of these weak but informative results.

5.4.3 Constraints of our work

We acknowledge several limitations within our work. Firstly, the low predictive power of the models, in part, is due to the sample sizes and SNPs that were present within the cohorts we used to train and validate our models. We also acknowledge that the individuals in the included datasets are predominantly of European descent, and thus the significance of our findings are limited to this ethnicity. One limitation that impacts the vast majority of PD research is the lack of consistency in diagnostic criteria from one cohort to the other, and our study is not exempt from this.

The limitations within our study do not detract from the strengths of our model which included the fact that contributing features were: 1) easily identifiable; 2) validated across three independent cohorts; and 3) consistently identified genomic regions that are unanimously recognised as being associated with PD (*e.g. SNCA*).

Our approach provides a significant advance over other previously reported methods. The novelty revolves around the ability of our method to: 1) rank the contributions that SNPs make to a phenotype through regulatory changes; 2) identify the tissues in which these changes are occurring; and 3) include effects from variants that do not have detectable eQTLs in the reference library that is used in the assay. Finally, the consistency between models and the ability to filter extraneous SNPs (*e.g.*, T1D eQTLs) out of the final predictor is another strength of this study. The higher predictive power observed for PD model-1 (Supplementary Table 17) may be explained by the observation that the final model included more features (827 vs 308). However, given that PD model-1 leveraged 290 PD-associated SNPs, the result also suggests that the 90 SNPs, originally identified as part of the Nalls *et al.* PRS analysis[33], do in fact contain the major genetic components that are associated with the risk of developing PD. Therefore, while other genetic signals clearly remain to be identified, the finding that both models consistently identified the same SNPs and heart atrial appendage eQTLs as the top contributors to the risk of developing PD further confirms the significance of these observations.

5.4.4 Conclusion

In conclusion, we applied machine learning algorithms to predict the risk of developing PD by integrating PD-associated SNPs with information on genome organisation, tissue-specific eQTLs and the genotypes of PD cases and controls. This enabled us to identify and rank the pivotal variants and tissue-specific eQTL effects that may contribute to the risk of developing PD. We also identified a novel association between variation in *EAF1-AS1* gene regulation in the heart atrial appendage and the risk of developing PD risk. We contend that our findings provide new insights into the involvement of cardiovascular function in the risk of PD.

Chapter 6: General Discussion

The onset and progression of complex diseases such as T1D and PD are influenced by a combination of genetic and environmental factors[34,188,271–273], where genetic elements provide the underlying platforms for the environmental factors to result in disease[188,273]. The machine learning approach I took was designed to deconstruct the relative contributions of SNPs to the risk of developing two disparate diseases: T1D and PD. T1D is an early-onset disorder with a strong genetic component[271]. Alternatively, PD is a late onset disease that results from a strong environmental influence and weak genetic influence[20]. For both diseases, the associated variants are located in non-coding DNA regions. Therefore, it is likely that these SNPs have an indirect impact on gene expression that results in a cumulative impact on complex gene regulation processes[40].

Every individual's phenotype is the result of a delicate balance of inputs and outputs from a network of different tissues and organs[274] and their interaction(s) with the environment. Therefore, the risk of developing both T1D and PD can be considered the cumulative result of the impact of disease associated genetic variants (through gene regulation) on each tissue or organ[275]. Insights into the functional aspects of tissue/organ-specific gene regulation, and the impact of genetic variants, can be ascertained using publicly available data obtained using disparate methods (e.g. Hi-C and eQTL assays)[131,137,140,276]. However, each method individually provides only a limited snap-shot of the total picture. Using machine learning predictor models, information from each data type can be interpreted as layers of information, whereby the impact of each genetic variant can be interrogated for its influence on each layer, resulting in a complex individualized understanding of disease risk[83,98,277].

In this thesis, I have developed a regularized logistic regression predictor model which integrates biologically relevant information from four forms of biological input data: disease-associated variants (GWAS), genome structure (Hi-C), tissue-specific gene expression (eQTL), and individual genotype data. My model uses two levels of feature selection to apply the most relevant features from the GWAS, Hi-C, and eQTL layers to the individual genotype data to best predict tissue-specific impacts on gene expression and disease risk. This results in the discovery and assignment of crucial disease- and tissue-specific risk for each disorder with the potential to establish personalized tissue-specific biomarkers for early detection and intervention in T1D and PD.

6.1 Data integration and building the predictor models

6.1.1 T1D

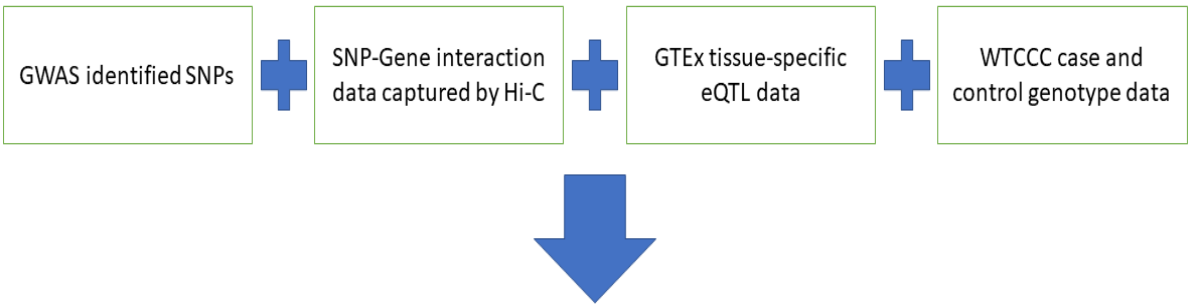
For my investigation into the genetics of T1D risk, 313 T1D SNPs from a collection of GWAS studies[19,164–170] were used as the starting point of the investigation. CoDeS3D[133] was then used to map the potential regulatory interactions of the T1D SNPs and genes. This mapping identified 8005 tissue-specific SNP-gene associations. To characterize the impacts of these T1D related tissue-specific eQTL effects is challenging. Using normalized effect size (NES) information from the GTEx study[275], the T1D variant modulated tissue-specific eQTL effects were estimated in individual samples from the Wellcome Trust T1D genotype data[70] (1960 cases and 2933 controls), which included 253 of the 313 T1D SNPs and their 6307 eQTL effects. Of the 253 T1D variants, 29 did not have any eQTL associations detected. However, it remains possible that these 29 T1D associated variants modulate T1D risk through gene regulation in tissues, cell types, at developmental stages, or by non-gene regulatory processes (*e.g.* protein-protein interactions) that are not represented within the GTEx study[275].

In my predictor model building, I applied two levels of data selection to identify only the informative data from the T1D individual eQTL effect matrix: 1) the Mann Whitney U test[158] in combination with the BY[163] procedure for controlling the FDR[158] was used to evaluate the association independently for each T1D variant (removing 4288 non-related features); and 2) elastic net regularization[90] was used to pick the subset of essential features that together deliver the best prediction[67]. Logistic regression models were chosen to generate the predictor, as the models that are generated by this platform can be assessed for result interpretation and evaluation of the component contributions[157].

The final T1D predictor model (T1D model-2) was built with the T1D optimized hyperparameters from the complete filtered eQTL matrix, selecting 134 tissue-specific eQTL effects and 6 SNPs without known eQTL effects (Figure 1). This model delivered an AUC of 0.774 with training data. While the AUC result was lower than other regularized predictors based on SNP data[67], my T1D model-2 mainly utilized eQTL effects across tissues to model T1D risk, and it was chosen as the best predictor from my modelling with the WTCCC derived eQTL data.

6.1.2 PD

In my PD study, 290 PD variants from a collection of GWAS studies were downloaded from the GWAS catalog database[278]. CoDeS3D[133] was used to identify 18,041 tissue-specific SNP-gene associations. Layering the Wellcome Trust PD genotype data[50] (1698 cases and 2668 controls) with the GTEx[275] tissue-specific NES information resulted in the inclusion of 281 from the 290 PD SNPs and their 17,829 regulated eQTL effects. After Mann-Whitney U test[158] with the BY control filtering[163], 11,288 variant derived features were selected. The final PD predictor used optimized hyperparameters (PD model-1) from the whole eQTL matrix that achieved an in-sample prediction AUC of 0.627 from 827 tissue-specific eQTLs and 6 SNPs with no eQTL effect (Figure 6-1).



	WTCCC T1D eQTL matrix (1960 cases and 2933 controls)	WTCCC PD eQTL matrix (1698 cases and 2668 controls)
WTCCC matrix features	6,336	17,829
Mann Whitey selected features	2,048	11,288
The final model features	140	833
In-sample AUC	0.774	0.627

Figure 6-1: Schematic of data integration and predictor model building for the T1D and PD studies performed in this thesis

6.2 Predictor model validations

Overfitting predictors is a common problem for predictor modelling[89]. An overfitted predictor lacks generalizing power and achieves excellent prediction with training data but has poor prediction accuracy with independent datasets[89]. To prevent overfitting, I employed internal cross-validation followed by independent dataset validation in my predictor modelling approach. Cross-validation and validation using independent test datasets can confirm the predictor performance[76,79,161]. I incorporated extensive validations in my predictor model building using cross-validation and external independent test datasets to ensure the developed predictors were not overfitted and retained good generalizing power in disease predictions. Conversely, many machine learning predictor developments employed only cross-validation or validations without external data[67,279,280].

The list of validation steps in the T1D study:

- 10 fold cross-validation with the training data for searching the T1D optimized model hyperparameters
- 20% of the T1D eQTL matrix was used as independent data for validating T1D model-1
- 10 repeated 5-fold cross-validation to create 50 predictors with the T1D optimized model hyperparameters for calculating the variation of the AUCs
- The T1D model-2 was validated using the UK Biobank[155] derived test dataset (30 test cohorts)

The list of validations performed in the PD study:

- 10 fold cross-validation with the full PD eQTL matrix for searching the PD optimized model hyperparameters
- 5 repeated 10-fold cross-validation to create 50 predictors with the PD optimized model hyperparameters for calculating the variation of the AUCs
- The PD model-1 was validated using the UK Biobank[155] derived test dataset (30 test cohorts)
- The PD model-1 was validated using the NeuroX-dbGap[51] derived test dataset

6.3 Regularized predictors identify genetic elements conferring complex disease risk

6.3.1 T1D model results

For each disease predictor, the absolute values of the feature model weights were used as the data feature contributions to the individual risk prediction. Ranking the tissues in T1D model-2 by their risk contributions, the results show that the eQTLs at lung (rs6679677 down-regulated *AP4P1-AS1* transcript levels in the lung, providing 13.3% to the risk prediction) and SNPs without detectable eQTL effects were the top two major T1D risk contributors.

Notably, the deterioration of lung functions has long been observed with T1D, known as diabetic lungs[281]. This is due to lungs acting as the primary interface for environmental factors in T1D[16,201,202], where respiratory infections might play a direct role in the onset of T1D[16]. Beyond T1D, lung-environment interactions could play a key role in promoting a variety of autoimmune diseases[201], such as auto-reactive T cells in Multiple Sclerosis[282]. Odoardi *et al.* experimented with green fluorescent protein tagged myelin basic protein reactive T cells in Lewis rats[282,283]. After stimulation, the tagged T cells were observed to reside in the lung for being reprogrammed of their gene expression profiles[282]. Subsequently, the reprogrammed auto-reactive T cells invaded the brain and caused inflammation[282]. The results reveal lungs could be involved in regulating auto-immune T cell activities. Nevertheless, the lung regulation functions have not been demonstrated with human models. The results were not attracting a lot of interest.

rs6679677 was the top variant associated with T1D reported from the original WTCCC GWAS study[70]. It is highly related to persistent multiple auto-antibody elevations including islet autoantibody[167] and has eQTL regulation effects with multiple immune regulating genes in whole blood samples[206–208]. rs6679677 is in significantly high LD ($r^2 > 0.9$)[284] with the well-studied R620W missense mutation of *PTPN22*[213,285–287]. Hence, the disease contributions of rs6679677 are overlooked and much of the focus is on R620W and *PTPN22*. Therefore, the disease contributions of rs6679677 could be shared between its regulatory role found here and the confounding genetic association with the R620W variant and its resulting effect on *PTPN22* function.

Studies show that long coding RNAs often participate in the regulation of genes located on opposite strands[288,289]. Employing small interfering RNAs, Stojic *et al.* have demonstrated that the suppression of *GNG12-AS1*transcription in cancer cells could up-regulate the transcription of *DIRAS3* located at the opposite strand showing the gene regulatory functions of antisense long coding RNAs.

AP4P1-AS1, a long non-coding RNA gene, is located in the opposite strand of *PTPN22* and in a DNA region reported to be strongly associated with a variety of autoimmune diseases[186]. *PTPN22* is involved in regulating T cell activation and receptor signalling[290] and its SNPs are highly associated with many autoimmune diseases such as Rheumatoid Arthritis (RA)[213,285,287,291]. We need further in vitro experiments with T cells to investigate the regulatory relationship between *AP4P1-AS1* and *PTPN22*, especially in immune cells in the lungs. For example, CRISPR-Cas9 experiments[292] altering rs6679677 or knockdown of *AP4P1-AS1* and qPCR[293] of *PTPN22* to measure changes in its transcription in T cells.

6.3.2 PD model results

Tissue ranking from PD model-1 indicated that eQTLs in Heart Atrial Appendage and 6 SNPs with unknown eQTL effects were the top two groups conferring PD risk. The top two rankings were also confirmed by the 50 regularized predictors created with the optimized PD model hyperparameters and PD model-2 created from the 90 variants of Nalls *et al*.[33] The model tissue ranking results imply a vital role for heart functions in PD pathogenesis, which has previously been shown heart disorders and PD as closely related[251–256]. The Heart Atrial Appendage is a trigger site of atrial fibrillation (AF)[259] and plays a crucial role in AF related stroke[261]. Moreover, AF is associated with early PD development[251]. With 15,375 PD patients, Hong et al. have studied the associations of AF within PD two-year before onset and two-year after. The results show that AF is significantly related to the before and early PD onset (adjusted odd ratio 1.15, 95% confidence interval: 1.04 - 1.28) but not the after. Evidence shows that AF can disrupt blood flow networks in brains promoting inflammation and tissue damage[264]. PD patients have been shown to have abnormal brain blood flow patterns[257]. I reason that the SNP modulated eQTLs in the Heart Atrial Appendage could promote AF leading to brain blood flow alterations and PD related brain tissue damages.

The 6 SNPs with unknown eQTL effects are all related to critical PD and Lewy body risk factor genes, including *INPP5F, GBA, SNCA* and *CNTN1*[30,31,242,247,248]. The two main eQTL risk contributors in Heat Atrial Appendage (rs7617877 and rs6808178) both up-regulated *EAF1-AS1,* a long non-coding RNA gene. Elevated *EAF1-AS1* transcript levels have been observed in PD patient brain tissues[234]. The opposite *EAF1* gene is involved in many critical cardiac biological pathways[265–267].

6.4 From GWAS SNPs to target genes

Many genetic variants have been discovered recently that are associated with the risk of developing complex diseases. These analyses have used large conglomerated case and control cohorts[19,51,171,294]. These genetic variants were selected and incorporated into predictor models to model individual complex disease risk and achieved acceptable predictive power[19,93,295]. By analyzing six different sets of European variant data (24,454 individual samples), Abraham *et al.*[93] applied the Lasso-SVM machine learning algorithm to select informative SNPs and create a model to predict the personal risk of Crohn's disease with AUC = 0.9 [93]. In another study, Sharp *et al.*[19] analyzed 6481 cases and 9247 controls from the Type 1 Diabetes Genetics Consortium and found 67 SNPs highly associated with T1D risk[19] and created the GRS2 predictive model that delivered T1D predictions with AUC = 0.92 [19]. Both predictive models above were developed for medical applications[19,93]. However, despite predicting individual disease risk accurately, these models were not designed to deconstruct how the related SNPs act to influence complex disease risk.

Makarios *et al.*[124] leveraged a mixture of three types of data (SNPs, Transcriptomics, Clinical and demographic) to develop a powerful predictor model that was used to distinguish essential genetic elements conferring PD risk. Twelve machine learning algorithms were tested on 49 different data combinations[124] and built an AdaBoost classifier model that achieved an impressive prediction performance (AUC = 0.85) on the validation data[124]. However, interpreting the model components became problematic. Firstly, the predictor contained 71 SNPs and 596 protein-coding transcripts[124]. Secondly, most of the predictive power was derived from the clinical and demographic data[124]. Notably, there were no clear connections between the selected SNPs and protein-coding transcripts. Hence, the predictor failed to shed light on the SNP modulated mechanisms. By contrast, the logistic regression modelling approach adopted in my regularised predictor modelling provides a clear and straightforward platform to evaluate the combinatorial relationships between my model elements.

GWAS have shown that more than 90% of the GWAS identified SNPs are located in intergenic regions[40,296], and they are most likely involved in gene regulation to influence diseases[40,296]. My modelling approach was focused on deciphering complex disorders through tissue-specific gene regulation associated with disease-related SNPs. Today, most studies still connect the related SNPs to their target genes based on *cis*-regulation (+/- 1 Mb from the genes)[33,51,147,294]. After analysing around 8 million SNPs from 13,708 cases and 95,282 controls, Nalls *et al.*[51] detected 26 risk variants for PD. Subsequently, they assessed the associations of the risk SNPs with genes surrounded by 1Mb apart in the methylation and expression data of brain tissues for locating the risk SNP targeted regulation[51]. The Nalls *et al.*[51] results implicated the cis regulated genes of the 26 crucial PD risk SNPs[51], but the study missed out on the other equally important *trans* regulated genes.

Although *cis* gene regulation is an essential component of disease aetiology, many studies demonstrate that SNPs can be involved in *trans* and *cis* gene regulation[297–299]. And, both SNP related *trans* and *cis* regulation have vital roles in promoting disorders[297–299]. Hi-C experiments are designed to describe the genomic 3D structural organization by capturing the proximal and distal DNA interactions that occur within cells[131–133]. And, the DNA interactions are significantly involved in gene regulation[132,223]. Fadason *et al.*[133] utilised Hi-C data to identify the putative SNP-gene contacts and thus calculate the *trans* and *cis* acting eQTL-gene regulatory connections involving SNPs related to obesity and diabetes in various tissues[133]. Fadason *et al.*[133] found evidence to support the hypothesis that SNPs located within *IGF2BP2* act as cis eQTLs in the regulation of *IGF2BP2* within thyroid tissues[133]. The SNPs were also associated with trans-regulation of three genes that were previously linked with obesity and diabetes (*i.e. RBM47*, *KIAA1430* and *DIS3L2*) in three different tissue types (Whole Blood, Hypothalamus, and Lung)[133]. The Fadason *et al.*[133] results of eQTLs in *IGF2BP2* have depicted a comprehensive view of the SNP controlled regulatory networks that impact both obesity and diabetes[133]. Using the same Fadason *et al.*[133] approach of recognizing the SNP-gene contacts, I used Hi-C data in my studies to identify the genes that were physically interacting in cis and trans for eQTL testing with the T1D- and PD-associated risk variants in order to discerning their tissue-specific genetic regulatory functions.

6.5 From GWAS SNP to tissue effects

GTEx studies have established that disease-associated SNPs can act as eQTLs to affect gene regulation in various tissues, and thus they can impact disease development[126,127,140,275]. It is well recognized that Type 2 diabetes is modulated by adipose tissue and the liver[300,301]. Lungs have been shown to promote auto-reactive T cell infiltration into brains during multiple sclerosis[282].

To estimate tissue-specific disease impacts, Gamazon *et al.*[143] applied LDSC to estimate risk contributions using disease-associated SNP information[143]. In their analysis, GTEx tissue-specific expression data were analysed to identify groups of genes that were expressed significantly only in one GTEx tissue. These genes were then used as the proxy for the contribution of each GTEx tissue in the risk evaluation[143]. The SNPs in the identified gene regions plus 100kb surrounding in each GTEx tissue were mapped to a full set of GWAS SNPs using LD scores[143]. The risk contributions were assigned to each GTEx tissue by regression of the LD scores[143]. Gamazon *et al.* examined the WTCCC T1D genotype data by LDSC with HLA and non-HLA genes[143]. Their non-HLA gene analysis identified the lung as the top tissue risk contributor to T1D[143], which is consistent with my T1D regularised predictor model results.

LDSC and my regularised predictor model method fundamentally differ in their starting presumptions and approaches for qualifying the tissue-specific disease risk. LDSC tissue risk estimations were calculated based on the LD strengths of the SNPs inside the selected tissue expressed only genes with GWAS SNPs[143]. LDSC only can assess tissue risk as a whole of each tissue type, but not the individual genetic elements. On the other hand, in my modelling approach, the risk contributions of each identified tissue-specific eQTL effect were first estimated with the case and control individual genotype data. Subsequent to this, the tissue risk contributions were calculated as the sum of the tissue-specific eQTL effect contributions within each tissue. Hence, my regularised predictor models also reveal the critical eQTL risk elements inside the related tissues.

Mendelian Randomisation (MR) is a computational method that is often used to confirm the causal relations of eQTL or mQTL effects with diseases[33,138,302]. MR utilises disease-associated SNPs and SNP-QTL effect associations to assess the connections of QTL effects and disorders[33,138,302]. In a meta-analysis of 17 PD GWAS datasets, Nalls *et al*.[33] identified 90 SNPs that were significantly associated with PD risk[33]. PRS analysis was employed to validate the risk contributions of the 90 SNPs[33]. The *cis*-regulation approach was adopted to recognise the associated QTL effects[33]. 237 genes near 70 of the 90 SNPs were nominated to evaluate their QTL effects in the brain and whole blood on PD risk by MR[33]. The method confirmed 151 of the 237 genes impacting PD through the expression or methylation changes were regulated by the SNPs of interest[33].

In my PD study, I applied a regularised predictor model analysis to interrogate the 90 SNPs discovered by Nalls *et al*. Using Hi-C data, I established trans and cis SNP-gene connections for 76 of the 90 SNPs and selected 3194 related eQTL effects with 518 genes across 49 GTEx tissues. The Mann Whitney U test[158] with the BY control[163] filtering and machine learning elastic net regularisation[90] selected 308 related tissue-specific eQTL effects (9 SNPs connecting to 95 genes across 49 GTEx tissue) as having causal impacts on PD risk. By applying logistic regression modelling, I could estimate the PD risk contributions of each inferred tissue-specific effect using the case and control genotype data. My final model, PD model-2, recognised seven PD risk contributors (7 eQTL effects: 4 SNPs connecting to 6 genes) of the heart atrial appendage tissue. The eQTL effects modulated by the PD SNPs in the heart atrial appendage were consistently shown to have substantial PD risk contributions by 50 regularised logistic regression predictors created from randomisation of the WTCCC[50] derived training data. The results suggest that the PD associated SNPs modulate heart functions to impact PD development. My analysis has deconstructed the impact of the 90 SNPs on PD and enabled the identification of potential tissue-specific mechanisms for PD genetic architecture.

6.6 Limitations of my study

The approach I took to implement the regularized predictor modelling analysis has several limitations. The first limitation arises from the use of the WTCCC case and control genotype datasets[50,70] in the predictor model training. The WTCCC associated limitations are largely due to the microarray technology used to produce the genotype data set, the sample biases within the data set, and sample size. Predictor models interpret and reveal the information within the training data[89]. Hence, the models will get all the predictive power but also incorporate the biases and errors that are inherent to the training data[89]. WTCCC T1D and PD genotype datasets[50,70] were acquired for training the regularized logistic predictors. Both case and control datasets were created more than ten years ago[50,70], and the genotyping microarrays used for measuring the genotypes was only able to handle 500k SNPs that are less powerful relatively to the genotyping arrays used in more updated GWAS studies measuring more than 2 million SNPs[172,303]. In the WTCCC studies[50,70], two cohorts (1958 Birth Cohort and UK Blood Services) of genotype data were used to generate common control datasets for various GWAS studies. Since the two control cohorts are not specifically selected and matched to each disease, they might include control samples that are less suitable as controls (*e.g.*, undetected disease). This would decrease the predictive power of the models by including incorrectly phenotyped samples into the training set. The sample sizes of the WTCCC data[50,70] used in both studies were below 5000. This is sufficient to detect significant effects but may be underpowered to identify minor signals from data features in the training data. This would bias feature selection and weighting in the final models.

Another limitation to my model training is that its ability to be used to generalize to trans-ethnic datasets remains unclear. The WTCCC individual samples[50,70] were mainly collected from the UK, as well as the UK biobank and PD samples used for validation. Thus, the sample data could be overfitted or biased towards T1D and PD effects specific to the European populations that dominate these cohorts. Ideally, a training dataset with bigger sample size and better genotype coverage generated by high density microarrays could help improve the predictive power of my algorithm and increase its ability to utilize the rich SNP information to explain the individual disease risk. Nevertheless, the WTCCC T1D and PD cohorts were the best datasets available to me when I started my research.

Another limitation of my approach is that the regularized predictor modelling utilizes tissue-specific eQTL effects to model disease risk. Genetic variants can influence diseases through many different mechanisms[180,181] (*e.g.*, protein-protein interactions[304], missense coding mutation[305] and DNA methylation[306]). While the predictor modelling includes SNP genotype data when the disease-related variants have no known eQTL effects, the modelling cannot account for the disease signals other than eQTL effects if the related SNPs have multiple pathways/mechanisms through which they modulate diseases (*e.g.*, a variant could act through both a gene- and protein-regulatory mechanism but only the gene-level was captured here). The limited scope of the eQTL gene regulation results in a lower predictive performance of my models when compared to those of PRS, but with the advantage of capturing the potential gene regulatory mechanisms behind my disease risk predictions. Moreover, my models were built based on the eQTL information from GTEx tissues[275] with limited tissue types (44 or 49 depending on the version used). As such, my predictor modelling cannot capture the risk signals from the related SNPs acting through eQTL effects in tissues or cell types not included in the GTEx database[275]. Additionally, the GTEx tissue sampling protocol includes whole tissues, meaning multiple cell types (including blood infiltration) are captured in each "tissue". Moreover, the GTEx tissue samples were mainly taken from white European males with old age which could make my results biased to a biologically special group. However, GTEx data was the largest tissue-specific eQTL database available to me for my research.

My modelling approach only relies on genetic information to estimate the disease risk of individuals, which becomes a limitation when the targeted disease only has weak genetic influence/heritability like PD. In the PD study, the PD regularized predictors resulted in limited predictive power, implying the models had less information available to them to predict the individual disease risk. This limitation has also affected other PD studies using LDSC to identify the disease risk enrichments in GTEx tissues[145,146]. A literature review of studies utilizing GTEx for tissue-specific enrichments found that three out of four PD studies could not find any risk enrichments in GTEx tissues[144,268–270]. By contrast, tissue prediction by LDSC in T1D with WTCCC T1D genotype data was able to detect strong risk enrichment in lungs[143]. Of course, this may also represent a limitation due to the conglomeration of the cases and controls within the cohorts[225].

The simple logistic regression algorithm employed in my approach could be considered to be another limitation because the regression models only assume the addition of the data feature effects as independent weights on the prediction model[76,79,157]. Thus, as the genetic features might not be entirely independent (*e.g.*, linked features, complex interactions between SNPs, multiplicative effects), the regression models may not fully account for the linked risk signals, leading to lower predictive power[89]. Nonetheless, the logistic regression model provided a straightforward platform for interpreting the eQTL effects in the T1D and PD studies, which illustrated the mechanisms underpinning the genetic architecture of both diseases.

Finally, all the conclusions and results of the regularized predictor modelling were drawn based on the estimated statistical disease associations of the genetic data in the combined eQTL matrices of the T1D and PD studies. Most of the SNP-gene interactions used as inputs in my models have not been empirically or independently validated. My models mitigated this limitation by ranking the effects of the data associations and conservatively only choosing the top-ranked genetic elements for interpreting the model results. Further future experiments are required to validate the findings from the predictor modelling of the T1D and PD studies.

6.7 Future directions

In the T1D study, my final model-2 identified eQTL rs6679677 as downregulating *AP4B1-AS1* in the lungs and that this was associated with conferring significant T1D risk. *AP4B1-AS1* could have a regulatory role on the opposite strand gene *PTPN22*. The eQTL rs6679677 locus has been shown to have enhancer activities in lung cells using a luciferase enhancer assay. Yet, many questions still need to be answered. Lungs are an organ that contains many different types of cells including T cells that play an important role in autoimmune disorders[290,307,308]. Evidence also suggests that resident T cell activities can be modulated in lungs[282]. In vitro experiments that investigate the regulatory effects of *AP4B1-AS1* transcription levels on *PTPN22* expression in lungs and T cells could be performed, using plasmids to up-regulate *AP4B1-AS1* transcription and RNA interference to knock down *AP4B1-AS1* transcription. Ideally, cells would be generated from autopsy samples of lungs and immune cells could be taken from deceased T1D patients to verify the effect of rs6679677 on *AP4B1-AS1* expression in T1D diseased lung tissue. These studies should also utilize the genotype and gene expression data to validate the causal impact of the eQTL effect on T1D by Mendelian randomization experiments[138,148].

Similarly, in the PD study, PD model-1 recognized rs7617877 and rs6808178 as both up-regulating *EAF1-AS1* transcript levels within the Heart Atrial Appendage. *EAF1-AS1* may also have a gene regulatory role on *EAF1* which is located on the opposite strand. I propose to use luciferase enhancer assay experiments with human cardiomyocyte cell lines (*e.g.*, AC16) to validate the enhancer activity of the loci marked by rs7617877 and rs6808178. Furthermore, plasmid and RNA interference experiments should be performed to overexpress and knockdown *EAF1-AS1* and thus investigate the effects of these putative regulatory regions on human heart cells and *EAS1* gene regulation. Again, autopsy samples from the heart and heart atrial appendage from deceased PD patients (*e.g.*, follow-up on IPDGC subject data) could be collected for a Mendelian randomization study to understand the relationship between *EAF1-AS1* transcription, organ remodelling and PD disease[85].The PD model-1 results suggest a strong association of heart atrial appendage function with PD risk. Currently, heart atrial appendage resection or occlusion surgery is a treatment of choice for atrial fibrillation patients to lower stroke risk. To validate my findings, a prospective epidemiological study could be performed to determine whether heart atrial appendage surgery changes the blood flow patterns in brains in a way that would promote tissue damage leading to PD or other neurodegenerative disorders. The study could employ magnetic resonance imaging to monitor the blood flow pattern changes in brains and their effects on PD development, providing invaluable insights into how heart functions impact brains, leading to disorders.

There are four key ways to improve my regularized predictor modelling analysis: 1) using bigger and better case and control genotype data for model building; 2) improving the selection of informative disease related SNPs; 3) increasing the eQTL tissue or cell types in modelling; and 4) employing advanced machine learning models to capture the full complexity of the genetic data.

For example, in my PD study, the NeuroX-dbGap[51] genotype data that was used for PD model-2 validation was designed and created in 2014 with 5,353 cases and 5,551 controls[51]. The microarrays used to generate the NeuroX-dbGap[51] genotype data were specially designed to cover only reported PD associated DNA regions, which makes the SNP imputation of the microarray variants very difficult. The missing SNPs only can be searched for using strong LD related proxies. However, this is imperfect, and a future dataset with this many samples but whole genome variant coverage would help with the evaluation of the PD predictive power, whilst providing more accurate estimates of the model results. Furthermore, the GWAS PD SNPs were mainly from European samples. In future studies, including SNPs discovered from other ancestries in future studies can improve the robustness of the disease predictions over different heterogenous populations[86], and also using advanced statistical platforms with SNP knockout procedure for selecting informative independent SNPs will help identify casual SNPs and their physiological impacts[85]. Similarly, many newly created eQTL datasets are publicly available for research use. For example, DICE is an immune cell eQTL database including 13 various immune cell types[276]. If DICE eQTL data[276] can be integrated into my modelling, my predictors would be able to utilize the eQTL information to identify the related immune cell types and estimate their contributions to disease risk. Many advanced machine learning algorithms (*e.g.*, SVM, Random Forrest and Deep Learning[61,309,310]) have been explored in biological data analysis applications and have achieved good results[88,98,120]. Adopting these advanced machine learning algorithms could help my predictors utilize the complex and interrelated genetic information more efficiently to enhance the risk predictive power.

My regularized predictor modelling analysis aims not to develop the most powerful predictors but to use good predictor models to interpret the functional roles of disease-related tissue-specific genetic elements in promoting individual disease risk. One of the applications is to employ my disease predictor models to reveal tissue-specific genetic elements according to individual genotypes. Other methods (*e.g.*, LDSC or PRS) can develop more refined predictor models to select individuals at risk with complex diseases. However, they lack the molecular understanding of why those model features were biologically relevant to the disease. Thus, integrating those methods within my predictor could leverage the benefits of better prediction while also exposing the important biological features underpinning those predictions. This would help to develop personalized healthcare bio-makers, specific to the tissues impacted early in the disease onset, for early detection and intervention to minimize the disease risk of individuals.

6.8 Conclusion

Complex disorders are the result of many genetic and environmental factors[273]. Genetic variants have been successfully used to reveal the disease risk of individuals but fail to illustrate the mechanisms promoting the diseases. In this thesis, I have developed a computational approach that integrates complex disease variants and their related tissue-specific gene regulation information with individual genotype data. The combined data were selected and analyzed by regularized logistic regression predictor models to estimate individual disease risk. After extensive validation, the best predictor model was chosen (according to the highest AUC) to reveal the essential variant related elements contributing to the complex disease risk. I applied the computational approach to study T1D and PD variants. My models have been able to identify major tissues, variants, and their patterns of tissue-specific gene regulation that were significantly associated with the T1D and PD disease risk. The regularized logistic regression models provided a clear platform for interpreting the genetic components of the model. These analyses implicate important insights into the genetic mechanisms acting on different tissues to modulate T1D and PD development.

The novelties of the regularised predictor modelling approach are the ability to distinguish *trans* and *cis* eQTL regulatory effects for disease-associated SNPs across tissues. With Mann Whitney U Test filtering[158] and machine learning regularisation, I can select and confirm the causal associations of the eQTL regulatory effects in multiple tissues. Moreover, my modelling can approximate the risk contribution of each tissue-specific eQTL regulatory effect for identifying the crucial tissue and their essential SNP modulated eQTL elements.

I suggest using in vitro experiments and autopsy patient tissue samples to verify my T1D and PD study findings. I also propose conducting an epidemiological study to investigate the impacts of heart atrial appendage functions on PD and neurodegenerative disorders to illustrate the interplay between dysfunction in the heart and subsequent disease manifestation in the brain. In conclusion, my model results provide novel and reproducible insights into the genetic underpinnings of T1D and PD, which can help to develop personalized tissue-specific bio-makers to identify and/or treat at risk individuals.

Appendices

The following tables are available on Figshare

Supplementary Table 1 to Table 9:

https://figshare.com/s/9d2ba00ca3bce4f24d47

Supplementary Table 10

https://figshare.com/s/421584a63292f1e4a6c7

Supplementary Table 11

https://figshare.com/s/a1bc5fd0d71b416d59c7

Supplementary Table 12

https://figshare.com/s/c2d4f9d38fd738d71c1d

Supplementary Table 13

https://figshare.com/s/7ac4eb48f3d57767165d

Supplementary Table 14

https://figshare.com/s/8f3138e94f2b1ad7dd3f

Supplementary Table 15

https://figshare.com/s/7ac4eb48f3d57767165d

Supplementary Table 16

https://figshare.com/s/dce6851987611d318743

Supplementary Table 17

https://figshare.com/s/aab1ddcc01d6ceaf9657

Supplementary Table 18

https://figshare.com/s/1f266739c3ff8b56cb0d

References

1. Pociot, F. *et al.* Genetics of type 1 diabetes: What's next? *Diabetes* **59**, 1561–1571 (2010).

2. Schork, N. J. Genetics of complex disease: Approaches, problems, and solutions. *Am. J. Respir. Crit. Care Med.* **156**, (1997).

3. International Diabetes Foundation. *International Diabetes Foundation*. *IDF Diabetes Atlas* (2017). doi:http://www.diabetesatlas.org/. (accessed 28 March 2018).

4. Zhou, B. *et al.* Worldwide trends in diabetes since 1980: a pooled analysis of 751 population-based studies with 4· 4 million participants. *Lancet* **387**, 1513–1530 (2016).

5. Armstrong, M. J. & Okun, M. S. Diagnosis and treatment of Parkinson disease: a review. *Jama* **323**, 548–560 (2020).

6. Sharpe, H. & Bradbury, S. Understanding excess body weight: New Zealand Health Survey. (2015).

7. Wright, A., Charlesworth, B., Rudan, I., Carothers, A. & Campbell, H. A polygenic basis for late-onset disease. *Trends Genet.* **19**, 97–106 (2003).

8. Matthysse, S. A general test of association for complex diseases with variable age of onset. *Genet. Epidemiol.* **19**, (2000).

9. Mobasseri, M. *et al.* Prevalence and incidence of type 1 diabetes in the world: A systematic review and meta-analysis. *Heal. Promot. Perspect.* **10**, 98–115 (2020).

10. Patterson, C. C. *et al.* Worldwide estimates of incidence, prevalence and mortality of type 1 diabetes in children and adolescents: Results from the International Diabetes Federation Diabetes Atlas, 9th edition. *Diabetes Res. Clin. Pract.* **157**, 107842 (2019).

11. Clark, M., Kroger, C. J. & Tisch, R. M. Type 1 diabetes: A chronic anti-self-inflammatory response. *Front. Immunol.* **8**, (2017).

12. Mallat, Z., Taleb, S., Ait-Oufella, H. & Tedgui, A. The role of adaptive T cell immunity in atherosclerosis. *J. Lipid Res.* **50**, S364–S369 (2009).

13. Eisenbarth, G. S. Type I Diabetes Mellitus A Chronic Autoimmune Disease. *Diabetes* **314**, 1360–1368 (1986).

14. Ounissi-Benkalha, H. & Polychronakos, C. The molecular genetics of type 1 diabetes: new

genes and emerging mechanisms. *Trends Mol. Med.* **14**, 268–275 (2008).

15. Nyaga, D. M., Vickers, M. H., Jefferies, C., Perry, J. K. & O'Sullivan, J. M. Type 1 Diabetes Mellitus-Associated Genetic Variants Contribute to Overlapping Immune Regulatory Networks. *Front. Genet.* **9**, 1–11 (2018).

16. Beyerlein, A., Wehweck, F., Ziegler, A. G. & Pflueger, M. Respiratory infections in early life and the development of islet autoimmunity in children at increased type 1 diabetes risk: Evidence from the BABYDIET study. *JAMA Pediatr.* **167**, 800–807 (2013).

17. George, C., Ducatman, A. M. & Conway, B. N. Increased risk of respiratory diseases in adults with Type 1 and Type 2 diabetes. *Diabetes Res. Clin. Pract.* **142**, 46–55 (2018).

18. Ram, R. *et al.* Systematic Evaluation of Genes and Genetic Variants Associated with Type 1 Diabetes Susceptibility. *J. Immunol.* **196**, 3043–3053 (2016).

19. Sharp, S. A. *et al.* Development and standardization of an improved type 1 diabetes genetic risk score for use in newborn screening and incident diagnosis. *Diabetes Care - Press* **42**, 200–207 (2019).

20. Mhyre, T. R., Nw, R., Boyd, J. T., Hall, G. & Room, C. Parkinson's Disease. *Subcell Biochem.* **65**, 389–455 (2012).

21. Lebouvier, T. *et al.* The second brain and Parkinson's disease. *Eur. J. Neurosci.* **30**, 735–741 (2009).

22. Pang, S. Y. Y. *et al.* The interplay of aging, genetics and environmental factors in the pathogenesis of Parkinson's disease. *Transl. Neurodegener.* **8**, 1–11 (2019).

23. Ray Dorsey, E. *et al.* Global, regional, and national burden of Parkinson's disease, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet Neurol.* **17**, 939–953 (2018).

24. Schapira, A. H. V., Chaudhuri, K. R. & Jenner, P. Non-motor features of Parkinson disease. *Nat. Rev. Neurosci.* **18**, 435–450 (2017).

25. Goldman, J. G. & Postuma, R. Premotor and nonmotor features of Parkinson's disease. *Curr. Opin. Neurol.* **27**, 434–441 (2014).

26. Hall, A., Bandres-Ciga, S., Diez-Fairen, M., Quinn, J. P. & Billingsley, K. J. Genetic risk profiling in parkinson's disease and utilizing genetics to gain insight into disease-related biological pathways. *Int. J. Mol. Sci.* **21**, 1–15 (2020).

27. Keller, M. F. *et al.* Using genome-wide complex trait analysis to quantify 'missing heritability' in Parkinson's disease. *Hum. Mol. Genet.* **21**, 4996–5009 (2012).

28. Bloem, B. R., Okun, M. S. & Klein, C. Parkinson's disease. *Lancet* **0**, (2021).

29. Polymeropoulos, M. H. *et al.* Mutation in the alpha-Synuclein Gene Identified in Families with Parkinson's Disease. *Science (80-.).* **276**, 2045–20047 (1997).

30. Riboldi, G. M. & Di Fonzo, A. B. GBA, Gaucher disease, and Parkinson's disease: from genetic to clinic to new therapeutic approaches. *Cells* **8**, 364 (2019).

31. Cao, M., Park, D., Wu, Y. & De Camilli, P. Absence of Sac2/INPP5F enhances the phenotype of a Parkinson's disease mutation of synaptojanin 1. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 12428–12434 (2020).

32. Zou, M. *et al.* Association analyses of variants of SIPA1L2, MIR4697, GCH1, VPS13C, and DDRGK1 with Parkinson's disease in East Asians. *Neurobiol. Aging* **68**, 159.e7-159.e14 (2018).

33. Nalls, M. A. *et al.* Identification of novel risk loci, causal insights, and heritable risk for Parkinson's disease: a meta-analysis of genome-wide association studies. *Lancet Neurol.* **18**, 1091–1102 (2019).

34. Chen, H. & Ritz, B. The search for environmental causes of Parkinson's disease: Moving forward. *J. Parkinsons. Dis.* **8**, S9–S17 (2018).

35. Gibbs, R. A. The Human Genome Project changed everything. *Nat. Rev. Genet.* **21**, 575–576 (2020).

36. Johnson, S. G. Genomic Medicine in Primary Care. in *Genomic and Precision Medicine (Third Edition)* 1–18 (Elsevier Inc., 2017). doi:10.1016/B978-0-12-800685-6.00001-1.

37. Laksman, Z. & Detsky, A. S. Personalized medicine: Understanding probabilities and managing expectations. *J. Gen. Intern. Med.* **26**, 204–206 (2011).

38. Spiegel, A. M. & Hawkins, M. 'Personalized medicine' to identify genetic risks for type 2 diabetes and focus prevention: Can it fulfill its promise? *Health Aff.* **31**, 43–49 (2012).

39. Auton, A. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).

40. Visscher, P. M. *et al.* 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am. J. Hum. Genet.* **101**, 5–22 (2017).

41. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, (2012).

42. Fadista, J., Manning, A. K., Florez, J. C. & Groop, L. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *Eur. J. Hum. Genet.* **24**, 1202–1205 (2016).

43. Panagiotou, O. A. *et al.* What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int. J. Epidemiol.* **41**, 273–286 (2012).

44. Edwards, T. L. *et al.* Genome-Wide association study confirms SNPs in SNCA and the MAPT region as common risk factors for parkinson disease. *Ann. Hum. Genet.* **74**, 97–109 (2010).

45. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes The Wellcome Trust Case Control Consortium. *Nat. Genet.* **July**; **39(**, 857–864 (2007).

46. Grant, S. F. A. *et al.* Follow-up analysis of genome-wide association data identifies novel loci for type 1 diabetes. *Diabetes* **58**, 290–295 (2009).

47. Mishra, R. *et al.* Relative contribution of type 1 and type 2 diabetes loci to the genetic etiology of adult-onset, non-insulin-requiring autoimmune diabetes. *BMC Med.* **15**, 88 (2017).

48. Pociot, F. Type 1 diabetes genome-wide association studies: Not to be lost in translation. *Clin. Transl. Immunol.* **6**, 1–7 (2017).

49. Baranzini, S. E. & Oksenberg, J. R. The Genetics of Multiple Sclerosis: From 0 to 200 in 50 Years. *Trends Genet.* **xx**, 1–11 (2017).

50. Spencer, C. C. A. *et al.* Dissection of the genetics of Parkinson's disease identifies an additional association 5' of SNCA and multiple associated haplotypes at 17q21. *Hum. Mol. Genet.* **20**, 345–353 (2011).

51. Nalls, M. A. *et al.* Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. *Nat. Genet.* **46**, 989–993 (2014).

52. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2017).

53. Schierding, W. *et al.* GWAS on prolonged gestation (post-term birth): Analysis of successive finnish birth cohorts. *J. Med. Genet.* **55**, (2018).

54. Schierding, W. & O'Sullivan, J. M. Connecting SNPs in Diabetes: A Spatial Analysis of Meta-

GWAS Loci. *Front. Endocrinol. (Lausanne).* **6**, 1–6 (2015).

55. Schierding, W., Antony, J., Cutfield, W. S., Horsfield, J. A. & O'Sullivan, J. M. Intergenic GWAS SNPs are key components of the spatial and regulatory network for human growth. *Hum. Mol. Genet.* **25**, 3372–3382 (2016).

56. Fadason, T., Ekblad, C., Ingram, J. R., Schierding, W. S. & Justin, M. Physical Interactions and Expression Quantitative Traits Loci Identify Regulatory Connections for Obesity and Type 2 Diabetes Associated SNPs. *Front. Genet.* (2017) doi:10.3389/fgene.2017.00150.

57. Abraham, G. & Inouye, M. Genomic risk prediction of complex human disease and its clinical application. *Curr. Opin. Genet. Dev.* **33**, 10–16 (2015).

58. Wray, N., Goddard, M. & Visscher, P. Prediction of individual genetic risk to disease from genome-wide association studies. *Genome Res.* **17**, 1520–1528 (2007).

59. Jostins, L. & Barrett, J. C. Genetic risk prediction in complex disease. *Hum. Mol. Genet.* **20**, 182–188 (2011).

60. Wang, X. *et al.* Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction. *J. Diabetes* **8**, 24–35 (2016).

61. Wei, Z. *et al.* From Disease Association to Risk Assessment: An Optimistic View from Genome-Wide Association Studies on Type 1 Diabetes. *PLoS Genet.* **5**, e1000678 (2009).

62. Manolio, T. A. Genomewide Association Studies and Assessment of the Risk of Disease. *N Engl J Med* **363**, 166–176 (2013).

63. Kruppa, J., Ziegler, A. & König, I. R. Risk estimation and risk prediction using machine-learning methods. *Hum. Genet.* **131**, 1639–1654 (2012).

64. Kooperberg, C., LeBlanc, M. & Obenchain, V. Risk prediction using genome-wide association studies. *Genet. Epidemiol.* **34**, 643–652 (2010).

65. Wray, N. R. *et al.* Research Review: Polygenic methods and their application to psychiatric traits. *J. Child Psychol. Psychiatry Allied Discip.* **55**, 1068–1087 (2014).

66. Dudbridge, F. Power and Predictive Accuracy of Polygenic Risk Scores. *PLoS Genet.* **9**, (2013).

67. Okser, S. *et al.* Regularized Machine Learning in the Genetic Prediction of Complex Traits. *PLoS Genet.* **10**, (2014).

68. Lyall, L. M. *et al.* Seasonality of depressive symptoms in women but not in men: a cross-sectional study in the UK Biobank cohort. *J. Affect. Disord.* **229**, 296–305 (2018).

69. Amin, N., Van Duijn, C. M. & Janssens, A. C. J. W. Genetic Scoring Analysis: A way forward in Genome Wide Association Studies? *Eur. J. Epidemiol.* **24**, 585–587 (2009).

70. Wellcome Trust Case Control Consortium. Genome-wide association study of 14 000 cases of seven common diseases and 3 000 shared controls. *Nature* **447**, 661–678 (2007).

71. Wei, Z. *et al.* Large sample size, wide variant spectrum, and advanced machine-learning technique boost risk prediction for inflammatory bowel disease. *Am. J. Hum. Genet.* **92**, 1008–1012 (2013).

72. Casson, R. J. & Farmer, L. D. M. Understanding and checking the assumptions of linear regression: A primer for medical researchers. *Clin. Exp. Ophthalmol.* **42**, 590–596 (2014).

73. Furlong, L. I. Human diseases through the lens of network biology. *Trends Genet.* **29**, 150–159 (2013).

74. Huang, Y. & Wang, P. Network Based Prediction Model for Genomics Data Analysis. *Stat Biosci* **4**, 1–23 (2012).

75. Clayton, D. G. Prediction and interaction in complex disease genetics: Experience in type 1 diabetes. *PLoS Genet.* **5**, 1–6 (2009).

76. Dasgupta, A., Sun, Y. V., König, I. R., Bailey-Wilson, J. E. & Malley, J. D. Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience. *Genet. Epidemiol.* **35**, 5–11 (2011).

77. Singh, G. & Samavedham, L. Unsupervised learning based feature extraction for differential diagnosis of neurodegenerative diseases: A case study on early-stage diagnosis of Parkinson disease. *J. Neurosci. Methods* **256**, 30–40 (2015).

78. Worachartcheewan, A. *et al.* Predicting Metabolic Syndrome Using the Random Forest Method. *Sci. World J.* **2015**, 1–10 (2015).

79. Mehta, P. *et al.* A high-bias, low-variance introduction to Machine Learning for physicists. *Phys. Rep.* **810**, 1–124 (2019).

80. Yuan, Y. Step-sizes for the gradient method. *AMS IP Stud. Adv. Math.* **42**, 785 (2008).

81. Vihinen, M. How to evaluate performance of prediction methods? Measures and their interpretation in variation effect analysis. *BMC Genomics* **13 Suppl 4**, (2012).

82. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, (2015).

83. Nguyen, T.-T., Huang, J., Wu, Q., Nguyen, T. & Li, M. Genome-wide association data classification and SNPs selection using two-stage quality-based Random Forests. *BMC Genomics* **16**, S5 (2015).

84. Schaffer, C. Technical Note: Selecting a Classification Method by Cross-Validation. *Mach. Learn.* **13**, 135–143 (1993).

85. Sesia, M., Bates, S., Candès, E., Marchini, J. & Sabatti, C. False discovery rate control in genome-wide association studies with population structure. *Proc. Natl. Acad. Sci. U. S. A.* **118**, 1–12 (2021).

86. Cavazos, T. B. & Witte, J. S. Inclusion of variants discovered from diverse populations improves polygenic risk score transferability. *Hum. Genet. Genomics Adv.* **2**, 100017 (2021).

87. Abraham, G., Kowalczyk, A., Zobel, J. & Inouye, M. Performance and Robustness of Penalized and Unpenalized Methods for Genetic Prediction of Complex Human Disease. *Genet. Epidemiol.* **37**, 184–195 (2013).

88. López, B., Torrent-Fontbona, F., Viñas, R. & Fernández-Real, J. M. Single Nucleotide Polymorphism relevance learning with Random Forests for Type 2 diabetes risk prediction. *Artif. Intell. Med.* 3–9 (2017) doi:10.1016/j.artmed.2017.09.005.

89. Ho, D. S. W., Schierding, W., Wake, M., Saffery, R. & O'Sullivan, J. Machine Learning SNP Based Prediction for Precision Medicine. *Front. Genet.* **10**, 1–10 (2019).

90. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J R Stat Soc Ser B* **67**, 301–320 (2005).

91. Shigemizu, D. *et al.* The construction of risk prediction models using GWAS data and its application to a type 2 diabetes prospective cohort. *PLoS One* **9**, (2014).

92. Shieh, Y. *et al.* Machine Learning–Based Gene Prioritization Identifies Novel Candidate Risk Genes for Inflammatory Bowel Disease. *Nat. Rev. Cancer* **12**, 1–12 (2017).

93. Abraham, G. *et al.* Accurate and Robust Genomic Prediction of Celiac Disease Using Statistical Learning. *PLoS Genet.* **10**, (2014).

94. Cox, D. R. The Regression Analysis of Binary Sequences. *J. R. Stat. Soc.* **20**, 215–242 (1958).

95. Yu, F., Rybar, M., Uhler, C. & Fienberg, S. E. Differentially-Private Logistic Regression for

Detecting Multiple-SNP Association in GWAS Databases BT - Privacy in Statistical Databases. in (ed. Domingo-Ferrer, J.) 170–184 (Springer International Publishing, 2014).

96. Niriella, M. A. *et al.* Lean non-alcoholic fatty liver disease (lean NAFLD): characteristics, metabolic outcomes and risk factors from a 7-year prospective, community cohort study from Sri Lanka. *Hepatol. Int.* (2018) doi:10.1007/s12072-018-9916-4.

97. Rosenblatt, F. Principles of neurodynamics. perceptrons and the theory of brain mechanisms. (1961).

98. Montañez, C. A. C., Fergus, P. & Chalmers, C. Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs. in *2018 International Joint Conference on Neural Networks (IJCNN)* 1–8 (2018).

99. Xue, L., Tang, B., Chen, W. & Luo, jiesi. Prediction of CRISPR sgRNA activity using a deep convolutional neural network. *J. Chem. Inf. Model.* acs.jcim.8b00368 (2018) doi:10.1021/acs.jcim.8b00368.

100. Yu, W. Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. *BMC Med. Inform. Decis. Mak.* **10**, (2010).

101. Corinna, C. & Vladimir, V. Support-Vector Networks. *Mach. Learn.* **20**, 273–297 (1995).

102. Han, J. The Design of Diabetic Retinopathy Classifier Based on Parameter Optimization SVM. *2018 Int. Conf. Intell. Informatics Biomed. Sci.* **3**, 52–58 (2018).

103. Tibshirani, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B* 267–288 (1996).

104. Song, J. Y. *et al.* New Genomic Model Integrating Clinical Factors and Gene Mutations to Predict Overall Survival in Patients with Diffuse Large B-Cell Lymphoma Treated with R-CHOP. *Blood* **132**, 346 (2018).

105. Rashkin, S. R. *et al.* A Pharmacogenetic Prediction Model of Progression-Free Survival in Breast Cancer using Genome-Wide Genotyping Data from CALGB 40502 (Alliance). *Clin. Pharmacol. Ther.* **0**, 1–8 (2018).

106. Quinlan, J. R. Induction of decision trees. *Mach. Learn.* **1**, 81–106 (1986).

107. Geurts, P., Irrthum, A. & Wehenkel, L. Supervised learning with decision tree-based methods in computational and systems biology. *Mol. Biosyst.* **5**, 1593–1605 (2009).

108. Li, Q., Diao, S., Li, H., He, H. & Li, J. Y. Applying decision trees to establish risk rating

model of breast cancer incidence based on non-genetic factors among Southwest China females. *Zhonghua Zhong Liu Za Zhi* **40**, 872–877 (2018).

109. Breiman, L. E. O. Random Forest. *Mach. Learn.* **45**, 5–32 (2001).

110. Dai, J. Y. *et al.* Case-only Methods Identified Genetic Loci Predicting a Subgroup of Men with Reduced Risk of High-grade Prostate Cancer by Finasteride. *Cancer Prev. Res.* canprevres-0284 (2018).

111. Freund, Y. & Schapire, R. E. Experiments with a New Boosting Algorithm. *Proc. 13th Int. Conf. Mach. Learn.* 148–156 (1996) doi:10.1.1.133.1040.

112. Li-ping, X. U., Jia, L. I. & Lin, F. Establishment of model of adaboost classifier and evaluation of harmful mutations in non-coding regions of liver cancer cells. *J. SHANGHAI JIAOTONG Univ. (MEDICAL Sci.* **35**, 819 (2015).

113. Capriotti, E., Calabrese, R. & Casadio, R. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics* **22**, 2729–2734 (2006).

114. Zhang, D. & Shen, D. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. *Neuroimage* **59**, 895–907 (2012).

115. Cruz, J. A. & Wishart, D. S. Applications of machine learning in cancer prediction and prognosis. *Cancer Inform.* **2**, 59–77 (2006).

116. Palaniappan, S. & Awang, R. Intelligent heart disease prediction system using data mining techniques. *2008 IEEE/ACS Int. Conf. Comput. Syst. Appl.* 108–115 (2008) doi:10.1109/AICCSA.2008.4493524.

117. Mieth, B. *et al.* Combining multiple hypothesis testing with machine learning increases the statistical power of genome-wide association studies. *Sci. Rep.* **6**, 1–14 (2016).

118. Boulesteix, A. L., Janitza, S., Kruppa, J. & König, I. R. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2**, 493–507 (2012).

119. Touw, W. G. *et al.* Data mining in the life science swith random forest: A walk in the park or lost in the jungle? *Brief. Bioinform.* **14**, 315–326 (2013).

120. Jiang, R., Tang, W., Wu, X. & Fu, W. A random forest approach to the detection of epistatic

interactions in case-control studies. *BMC Bioinformatics* **10**, 1–12 (2009).

121. Austin, P. C., Tu, J. V., Ho, J. E., Levy, D. & Lee, D. S. Using methods from the data-mining and machine-learning literature for disease classification and prediction: A case study examining classification of heart failure subtypes. *J. Clin. Epidemiol.* **66**, 398–407 (2013).

122. Chen, X. & Ishwaran, H. Random forests for genomic data analysis. *Genomics* **99**, 323–329 (2012).

123. Hu, J. Automated detection of driver fatigue based on AdaBoost classifier with EEG signals. *Front. Comput. Neurosci.* **11**, (2017).

124. Makarious, M. B. *et al.* Multi-Modality Machine Learning Predicting Parkinson ' s Disease. *bioRxiv* (2021).

125. Suzuki, D. T. & Griffiths, A. J. F. *An introduction to genetic analysis.* (WH Freeman and Company., 1976).

126. Aguet, F. *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).

127. Ardlie, K. G. *et al.* The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. *Science (80-.).* **348**, 648–660 (2015).

128. Devecchi, A. *et al.* The genomics of desmoplastic small round cell tumor reveals the deregulation of genes related to DNA damage response, epithelial–mesenchymal transition, and immune response. *Cancer Commun.* **38**, 1–14 (2018).

129. Benavente, C. A. *et al.* Chromatin remodelers HELLS and UHRF1 mediate the epigenetic deregulation of genes that drive retinoblastoma tumor progression. *Oncotarget* **5**, 9594 (2014).

130. Grubert, F. *et al.* Genetic Control of Chromatin States in Humans Involves Local and Distal Chromosomal Interactions. *Cell* **162**, 1051–1065 (2015).

131. Pal, K., Forcato, M. & Ferrari, F. Hi-C analysis: from data generation to integration. *Biophys. Rev.* **11**, 67–78 (2019).

132. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).

133. Fadason, T., Ekblad, C., Ingram, J. R., Schierding, W. S. & O'Sullivan, J. M. Physical interactions and expression quantitative traits loci identify regulatory connections for obesity and type 2 diabetes associated SNPs. *Front. Genet.* **8**, (2017).

134. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science (80-.).* **326**, 289–293 (2009).

135. Delaneau, O. *et al.* Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science (80-.).* **364**, (2019).

136. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science (80-.).* **369**, 1318–1330 (2021).

137. Ramani, V. *et al.* Mapping three-dimensional genome architecture through in situ DNase Hi-C. *Nat Protoc.* **11**, 2104–2121 (2016).

138. Gleason, K. J., Yang, F. & Chen, L. S. A robust two-sample transcriptome-wide Mendelian randomization method integrating GWAS with multi-tissue eQTL summary statistics. *Genet. Epidemiol.* **45**, 353–371 (2021).

139. Parker, S. C. J. *et al.* Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proc. Natl. Acad. Sci.* **110**, 17921 LP – 17926 (2013).

140. Ongen, H. *et al.* Estimating the causal tissues for complex traits and diseases. *Nat. Genet.* **49**, (2017).

141. Nica, A. C. & Dermitzakis, E. T. Expression quantitative trait loci: present and future. *Philos. Trans. Biol. Sci.* **368**, 1–6 (2013).

142. Rusu, V. *et al.* Type 2 Diabetes Variants Disrupt Function of SLC16A11 through Two Distinct Mechanisms. *Cell* **170**, 199-212.e20 (2017).

143. Gamazon, E. R. *et al.* Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet.* **50**, 956–967 (2018).

144. Reynolds, R. H. *et al.* Moving beyond neurons: the role of cell type-specific gene regulation in Parkinson's disease heritability. *npj Park. Dis.* **5**, (2019).

145. Finucane, H. K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet.* **50**, 621–629 (2018).

146. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).

147. Pavlides, J. M. W. *et al.* Predicting gene targets from integrative analyses of summary data from GWAS and eQTL studies for 28 human complex traits. *Genome Med.* **8**, 4–9 (2016).

148. Porcu, E. *et al.* Mendelian randomization integrating GWAS and eQTL data reveals genetic determinants of complex and clinical traits. *Nat. Commun.* **10**, 1–12 (2019).

149. Winkler, C. *et al.* Feature ranking of type 1 diabetes susceptibility genes improves prediction of type 1 diabetes. *Diabetologia* **57**, 2521–2529 (2014).

150. Pahikkala, T., Okser, S., Airola, A., Salakoski, T. & Aittokallio, T. Wrapper-based selection of genetic features in genome-wide association studies through fast matrix operations. *Algorithms Mol. Biol.* **7**, 1–15 (2012).

151. Laraway, S., Snycerski, S., Pradhan, S. & Huitema, B. E. An Overview of Scientific Reproducibility: Consideration of Relevant Issues for Behavior Science/Analysis. *Perspect. Behav. Sci.* **42**, 33–57 (2019).

152. Argyraki, A. *et al.* Reproducibility and Research Integrity. *2015 IEEE Summer Top. Meet. Ser. SUM 2015* **10**, 1–13 (2018).

153. Wilson, G. *et al.* Good enough practices in scientific computing. *PLoS Comput. Biol.* **13**, e1005510 (2017).

154. Wilson, G. *et al.* Best Practices for Scientific Computing. *PLoS Biol.* **12**, (2014).

155. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).

156. Quinlan, J. R. Learning Logical Definitions from Relations. *Mach. Learn.* **5**, 239–266 (1990).

157. Dreiseitl, S. & Ohno-Machado, L. Logistic regression and artificial neural network classification models: A methodology review. *J. Biomed. Inform.* **35**, 352–359 (2002).

158. McKnight, P. E. & Najab, J. Mann Whitney U Test. *Corsini Encycl. Psychol.* 1 (2010).

159. Benjamini, Y. & Yekutieli, D. The control of the false discovery rate in multiple testing under dependency. *Ann. Stat.* 1165–1188 (2001).

160. Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**, 714–721 (2009).

161. Browne, M. W. Cross-Validation Methods. *J. Math. Psychol.* **44**, 108–132 (2000).

162. Fadason, T., Schierding, W., Lumley, T. & O'Sullivan, J. M. Chromatin interactions and expression quantitative trait loci reveal genetic drivers of multimorbidities. *Nat. Commun.* **9**, 1–13 (2018).

163. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).

164. Frederiksen, B. N. *et al.* Evidence of Stage- and Age-Related Heterogeneity of Non-HLA SNPs and Risk of Islet Autoimmunity and Type 1 Diabetes: The Diabetes Autoimmunity Study in the Young. *Clin. Dev. Immunol.* **2013**, 1–8 (2013).

165. Steck, A. K. *et al.* Improving prediction of type 1 diabetes by testing non-HLA genetic variants in addition to HLA markers. *Pediatr. Diabetes* **15**, 355–362 (2014).

166. Bonifacio, E., Warncke, K., Winkler, C., Wallner, M. & Ziegler, A.-G. Cesarean Section and Interferon-Induced Helicase Gene Polymorphisms Combine to Increase Childhood Type 1 Diabetes Risk. *Diabetes* **60**, 3300–3306 (2011).

167. Sharma, A. *et al.* Identification of non-HLA genes associated with development of islet autoimmunity and type 1 diabetes in the prospective TEDDY cohort. *J. Autoimmun.* **89**, 90–100 (2018).

168. Steck, A. K. *et al.* Can Non-HLA Single Nucleotide Polymorphisms Help Stratify Risk in TrialNet Relatives at Risk for Type 1 Diabetes? *J. Clin. Endocrinol. Metab.* **102**, 2873–2880 (2017).

169. Howson, J. M. M., Rosinger, S., Smyth, D. J., Boehm, B. O. & Todd, J. A. Genetic Analysis of Adult-Onset Autoimmune Diabetes. *Diabetes* **60**, 2645–2653 (2011).

170. Bonifacio, E. *et al.* Genetic scores to stratify risk of developing multiple islet autoantibodies and type 1 diabetes: A prospective study in children. *PLOS Med.* **15**, e1002548 (2018).

171. Oram, R. A. *et al.* A type 1 diabetes genetic risk score can aid discrimination between type 1 and type 2 diabetes in young adults. *Diabetes Care* **39**, 337–344 (2016).

172. Lamy, P., Grove, J. & Wiuf, C. A review of software for microarray genotyping. *Hum. Genomics* **5**, 304–309 (2011).

173. Quick, C. *et al.* Sequencing and imputation in GWAS: Cost-effective strategies to increase power and genomic coverage across diverse populations. *Genet. Epidemiol.* **44**, 537–549 (2020).

174. Loh, P.-R. *et al.* Reference-based phasing using the Haplotype Reference Consortium panel. *Nat. Genet.* **48**, 1443–1448 (2016).

175. Durbin, R. Efficient haplotype matching and storage using the positional Burrows-Wheeler

transform (PBWT). *Bioinformatics* **30**, 1266–1272 (2014).

176. McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279 (2016).

177. Kwak, S. G. & Kim, J. H. Central limit theorem: the cornerstone of modern statistics. *Korean J. Anesthesiol.* **70**, 144–156 (2017).

178. Nalls, M. A. *et al.* NeuroX, a fast and efficient genotyping platform for investigation of neurodegenerative diseases. *Neurobiol. Aging* **36**, 1605.e7-1605.e12 (2015).

179. Mohammadi, P., Castel, S. E., Brown, A. A. & Lappalainen, T. Quantifying the regulatory effect size of cis-acting genetic variation using allelic fold change. *Genome Res.* **27**, 1872–1884 (2017).

180. Cieply, B. & Carstens, R. P. Functional roles of alternative splicing factors in human disease. *Wiley Interdiscip. Rev. RNA* **6**, 311–326 (2015).

181. Al-Hasani, K., Mathiyalagan, P. & El-Osta, A. Epigenetics, cardiovascular disease, and cellular reprogramming. *J. Mol. Cell. Cardiol.* **128**, 129–133 (2019).

182. Raschka, S. & Mirjalili, V. Python Machine Learning: Machine Learning and Deep Learning with Python. *Scikit-Learn, TensorFlow. Second Ed. ed* (2017).

183. Abraham, A. *et al.* Machine learning for neuroimaging with scikit-learn. *Front. Neuroinform.* **8**, 14 (2014).

184. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

185. Defazio, A., Bach, F. & Lacoste-Julien, S. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. *Adv. Neural Inf. Process. Syst.* **2**, 1646–1654 (2014).

186. Onengut-gumuscu, S. *et al.* Fine mapping of type 1 diabetes susceptibility loci and evidence for colocalization of causal variants with lymphoid gene enhancers. *Nat. Genet.* **47**, 381–386 (2015).

187. Rewers, M. *et al.* Newborn screening for HLA markers associated with IDDM: Diabetes Autoimmunity Study in the Young (DAISY). *Diabetologia* **39**, 807–812 (1996).

188. Krischer, J. P. *et al.* Genetic and Environmental Interactions Modify the Risk of Diabetes-Related Autoimmunity by 6 Years of Age: The TEDDY Study. *Diabetes Care* **40**, 1194–1202

(2017).

189. Hummel, S. & Ziegler, A. G. Early determinants of type 1 diabetes: experience from the BABYDIAB and BABYDIET studies. *Am. J. Clin. Nutr.* **94**, 1821S-1823S (2011).

190. Christ, M., Braun, N., Neuffer, J. & Kempa-Liehr, A. W. Time Series FeatuRe Extraction on basis of Scalable Hypothesis tests (tsfresh – A Python package). *Neurocomputing* **307**, 72–77 (2018).

191. Reiner, A., Yekutieli, D. & Benjamini, Y. Identifying differentially expressed genes using false discovery rate controlling procedures. *Bioinformatics* **19**, 368–375 (2003).

192. Melnikov, A., Zhang, X., Rogov, P., Wang, L. & Mikkelsen, T. S. Massively parallel reporter assays in cultured mammalian cells. *J. Vis. Exp.* 1–8 (2014) doi:10.3791/51719.

193. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).

194. R Core Team. R: A language and environment for statistical computing. *R Foundation for Statistical Computing* vol. 739 1–2630 (2014).

195. Salvatier, J., Wieckiâ, T. V & Fonnesbeck, C. PyMC3: Python probabilistic programming framework. *ascl* ascl-1610 (2016).

196. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).

197. Kruschke, J. K. Bayesian estimation supersedes the t test. *J. Exp. Psychol. Gen.* **142**, 573 (2013).

198. Orban, T. *et al.* Costimulation Modulation With Abatacept in Patients With Recent-Onset Type 1 Diabetes: Follow-up 1 Year After Cessation of Treatment. *Diabetes Care* **37**, 1069–1075 (2014).

199. Võsa, U. *et al.* Unraveling the polygenic architecture of complex traits using blood eQTL meta-analysis. *bioRxiv* **18**, 10 (2018).

200. Noble, J. A. & Valdes, A. M. Genetics of the HLA Region in the Prediction of Type 1 Diabetes. *Curr. Diab. Rep.* **11**, 533–542 (2011).

201. Zhao, C.-N. *et al.* Emerging role of air pollution in autoimmune diseases. *Autoimmun. Rev.* **18**, 607–614 (2019).

202. Hathout, E. H., Beeson, W. L., Ischander, M., Rao, R. & Mace, J. W. Air pollution and type 1 diabetes in children. *Pediatr. Diabetes* **7**, 81–87 (2006).

203. Lönnrot, M. *et al.* Respiratory infections are temporally associated with initiation of type 1 diabetes autoimmunity: the TEDDY study. *Diabetologia* **60**, 1931–1940 (2017).

204. Pociot, F., Kaur, S. & Nielsen, L. B. Effects of the genome on immune regulation in type 1 diabetes. *Pediatr. Diabetes* **17**, 37–42 (2016).

205. Alizadeh, B. Z. *et al.* MICA marks additional risk factors for Type 1 diabetes on extended HLA haplotypes: an association and meta-analysis. *Mol. Immunol.* **44**, 2806–2812 (2007).

206. Krummel, M. F. & Allison, J. P. CD28 and CTLA-4 have opposing effects on the response of T cells to stimulation. *J. Exp. Med.* **182**, 459–465 (1995).

207. Willerford, D. M. *et al.* Interleukin-2 receptor α chain regulates the size and content of the peripheral lymphoid compartment. *Immunity* **3**, 521–530 (1995).

208. Wang, N. *et al.* Negative regulation of humoral immunity due to interplay between the SLAMF1, SLAMF5, and SLAMF6 receptors. *Front. Immunol.* **6**, 1–13 (2015).

209. Konopacki, C., Pritykin, Y., Rubtsov, Y., Leslie, C. S. & Rudensky, A. Y. Transcription factor Foxp1 regulates Foxp3 chromatin binding and coordinates regulatory T cell function. *Nat. Immunol.* **20**, 232–242 (2019).

210. Gutierrez-Achury, J. *et al.* Functional implications of disease-specific variants in loci jointly associated with coeliac disease and rheumatoid arthritis. *Hum. Mol. Genet.* **25**, 180–190 (2016).

211. Hinks, A. *et al.* Dense genotyping of immune-related disease regions identifies 14 new susceptibility loci for juvenile idiopathic arthritis. *Nat. Genet.* **45**, 664–669 (2013).

212. Yu, X. *et al.* Structure, inhibitor, and regulatory mechanism of Lyp, a lymphoid-specific tyrosine phosphatase implicated in autoimmune diseases. *Proc. Natl. Acad. Sci.* **104**, 19767–19772 (2007).

213. Crabtree, J. N. *et al.* Autoimmune variant PTPN22 C1858T is associated with impaired responses to influenza vaccination. *J. Infect. Dis.* **214**, 248–257 (2016).

214. Inoue, F. & Ahituv, N. Decoding enhancers using massively parallel reporter assays. *Genomics* **106**, 159–164 (2015).

215. Clarke, C. E. *et al. UK Parkinson's Disease Society Brain Bank Diagnostic Criteria*. vol.

Appendix 1 (NIHR Journals Library, 2016).

216. Berg, D. *et al.* Prodromal Parkinson disease subtypes — key to understanding heterogeneity. *Nat. Rev. Neurol.* 1–13 (2021) doi:10.1038/s41582-021-00486-9.

217. Killcoyne, S. *et al.* Genomic copy number predicts esophageal cancer years before transformation. *Nat. Med.* **26**, 1726–1732 (2020).

218. Yong, S. Y., Raben, T. G., Lello, L. & Hsu, S. D. H. Genetic architecture of complex traits and disease risk predictors. *Sci. Rep.* **10**, 12055 (2020).

219. Sud, A., Kinnersley, B. & Houlston, R. S. Genome-wide association studies of cancer: current insights and future perspectives. *Nat. Rev. Cancer* **17**, 692–704 (2017).

220. Guo, X., Zhang, J., Cai, Z., Du, D. Z. & Pan, Y. Searching genome-wide multi-locus associations for multiple diseases based on Bayesian inference. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **14**, 600–610 (2017).

221. Chang, D. *et al.* A meta-analysis of genome-wide association studies identifies 17 new Parkinson's disease risk loci. *Nat. Genet.* **49**, 1511–1516 (2017).

222. Farrow, S. L. *et al.* Establishing gene regulatory networks from Parkinson's disease risk loci. *bioRxiv* 2021.04.08.439080 (2021) doi:10.1101/2021.04.08.439080.

223. Yu, J., Hu, M. & Li, C. Joint analyses of multi-tissue Hi-C and eQTL data demonstrate close spatial proximity between eQTLs and their target genes. *BMC Genet.* **20**, 43 (2019).

224. Duggal, G., Wang, H. & Kingsford, C. Higher-order chromatin domains link eQTLs with the expression of far-away genes. *Nucleic Acids Res.* **42**, 87–96 (2014).

225. Schierding, W. *et al.* Common Variants Coregulate Expression of GBA and Modifier Genes to Delay Parkinson's Disease Onset. *Mov. Disord.* **35**, 1346–1356 (2020).

226. Naushad, S. M. *et al.* Machine learning algorithm-based risk prediction model of coronary artery disease. *Mol. Biol. Rep.* **45**, 901–910 (2018).

227. Siitonen, A. *et al.* Genetics of early-onset Parkinson's disease in Finland: exome sequencing and genome-wide association study. *Neurobiol. Aging* **53**, 195-e7 (2017).

228. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).

229. de Leeuw, C. A., Mooij, J. M., Heskes, T. & Posthuma, D. MAGMA: Generalized Gene-Set

Analysis of GWAS Data. *PLoS Comput. Biol.* **11**, 1–19 (2015).

230. Jiao, X. *et al.* DAVID-WS: A stateful web service to facilitate gene/protein list analysis. *Bioinformatics* **28**, 1805–1806 (2012).

231. Wider, C. *et al.* Association of the MAPT locus with Parkinson's disease. *Eur. J. Neurol.* **17**, 483–486 (2010).

232. Tobin, J. E. *et al.* Haplotypes and gene expression implicate the MAPT region for Parkinson disease: The GenePD Study. *Neurology* **71**, 28–34 (2008).

233. Machiela, M. J. & Chanock, S. J. LDlink: A web-based application for exploring population-specific haplotype structure and linking correlated alleles of possible functional variants. *Bioinformatics* **31**, 3555–3557 (2015).

234. Dick, F. *et al.* Differential transcript usage in the Parkinson's disease brain. *PLoS Genet.* **16**, 1–24 (2020).

235. Boros, F. A., Maszlag-Török, R., Vécsei, L. & Klivényi, P. Increased level of NEAT1 long non-coding RNA is detectable in peripheral blood cells of patients with Parkinson's disease. *Brain Res.* **1730**, 146672 (2020).

236. Escott-Price, V. *et al.* Polygenic risk of Parkinson disease is correlated with disease age at onset. *Ann. Neurol.* **77**, 582–591 (2015).

237. Lill, C. M. *et al.* Impact of Parkinson's disease risk loci on age at onset. *Mov. Disord.* **30**, 847–850 (2015).

238. Riou, A. *et al.* Functional Role of the Cerebellum in Parkinson Disease: A PET Study. *Neurology* (2021) doi:10.1212/WNL.0000000000012036.

239. Seidel, K. *et al.* Involvement of the cerebellum in Parkinson disease and dementia with Lewy bodies. *Ann. Neurol.* **81**, 898–903 (2017).

240. Wu, T. & Hallett, M. The cerebellum in Parkinson's disease. *Brain* **136**, 696–709 (2013).

241. Berge-Seidl, V. *et al.* The GBA variant E326K is associated with Parkinson's disease and explains a genome-wide association signal. *Neurosci. Lett.* **658**, 48–52 (2017).

242. Siddiqui, I. J., Pervaiz, N. & Abbasi, A. A. The Parkinson Disease gene SNCA: Evolutionary and structural insights with pathological implication. *Sci. Rep.* **6**, 1–11 (2016).

243. Grenn, F. P. *et al.* The Parkinson's Disease Genome-Wide Association Study Locus Browser.

Mov. Disord. **35**, 2056–2067 (2020).

244. Kim, H. S., Li, A., Ahn, S., Song, H. & Zhang, W. Inositol Polyphosphate-5-Phosphatase F (INPP5F) inhibits STAT3 activity and suppresses gliomas tumorigenicity. *Sci. Rep.* **4**, 1–10 (2014).

245. Zhu, W. *et al.* Inpp5f is a polyphosphoinositide phosphatase that regulates cardiac hypertrophic responsiveness. *Circ. Res.* **105**, 1240–1247 (2009).

246. Zou, Y. *et al.* Gene-silencing screen for mammalian axon regeneration identifies Inpp5f (Sac2) as an endogenous suppressor of repair after spinal cord injury. *J. Neurosci.* **35**, 10429–10439 (2015).

247. Chatterjee, M. *et al.* Contactin-1 is reduced in cerebrospinal fluid of parkinson's disease patients and is present within lewy bodies. *Biomolecules* **10**, 1–15 (2020).

248. Guerreiro, R. *et al.* Investigating the genetic architecture of dementia with Lewy bodies: a two-stage genome-wide association study. *Lancet Neurol.* **17**, 64–74 (2018).

249. Anderson, C. *et al.* PLP1 and CNTN1 gene variation modulates the microstructure of human white matter in the corpus callosum. *Brain Struct. Funct.* **223**, 3875–3887 (2018).

250. Mata, I. F. *et al.* Large-scale exploratory genetic analysis of cognitive impairment in Parkinson's disease. *Neurobiol. Aging* **56**, 211.e1-211.e7 (2017).

251. Hong, C. T., Chan, L., Wu, D., Chen, W. T. & Chien, L. N. Association between Parkinson's disease and atrial fibrillation: A population-based study. *Front. Neurol.* **10**, (2019).

252. Scorza, F. A., Fiorini, A. C., Scorza, C. A. & Finsterer, J. Cardiac abnormalities in Parkinson's disease and Parkinsonism. *J. Clin. Neurosci.* **53**, 1–5 (2018).

253. Potashkin, J. *et al.* Understanding the links between cardiovascular disease and Parkinson's disease. *Mov. Disord.* **35**, 55–74 (2020).

254. Awerbuch, G. I. & Sandyk, R. Autonomic functions in the early stages of parkinson's disease. *Int. J. Neurosci.* **74**, 9–16 (1994).

255. Ascherio, A. & Tanner, C. M. Use of antihypertensives and the risk of parkinson disease. *Neurology* **72**, 578–579 (2009).

256. Fang, X. *et al.* Association of Levels of Physical Activity With Risk of Parkinson Disease: A Systematic Review and Meta-analysis. *JAMA Netw. open* **1**, e182421 (2018).

257. Teune, L. K. *et al.* Parkinson's disease-related perfusion and glucose metabolic brain patterns identified with PCASL-MRI and FDG-PET imaging. *NeuroImage Clin.* **5**, 240–244 (2014).

258. Han, S. *et al.* Increased atrial fibrillation risk in Parkinson's disease: A nationwide population-based study. *Ann. Clin. Transl. Neurol.* **8**, 238–246 (2021).

259. Di Biase, L. *et al.* Left atrial appendage: An underrecognized trigger site of atrial fibrillation. *Circulation* **122**, 109–118 (2010).

260. Stöllberger, C., Schneider, B. & Finsterer, J. Elimination of the left atrial appendage to prevent stroke or embolism?: anatomic, physiologic, and pathophysiologic considerations. *Chest* **124**, 2356–2362 (2003).

261. Hart, R. G. & Halperin, J. L. Atrial fibrillation and stroke: concepts and controversies. *Stroke* **32**, 803–808 (2001).

262. Turagam, M. K. *et al.* Epicardial Left Atrial Appendage Exclusion Reduces Blood Pressure in Patients With Atrial Fibrillation and Hypertension. *J. Am. Coll. Cardiol.* **72**, 1346–1353 (2018).

263. Du, W. *et al.* Large left atrial appendage predicts the ablation outcome in hypertensive patients with atrial fibrillation. *J. Electrocardiol.* **63**, 139–144 (2020).

264. Junejo, R. T., Lip, G. Y. H. & Fisher, J. P. Cerebrovascular Dysfunction in Atrial Fibrillation. *Front. Physiol.* **11**, (2020).

265. Liu, J. X. *et al.* Eaf1 and Eaf2 negatively regulate canonical Wnt/β-catenin signaling. *Dev.* **140**, 1067–1078 (2013).

266. Liu, J. X. *et al.* Transcriptional factors Eaf1/2 inhibit endoderm and mesoderm formation via suppressing TGF-β signaling. *Biochim. Biophys. Acta - Gene Regul. Mech.* **1860**, 1103–1116 (2017).

267. Yousefi, F. *et al.* TGF-β and WNT signaling pathways in cardiac fibrosis: Non-coding RNAs come into focus. *Cell Commun. Signal.* **18**, 1–16 (2020).

268. Li, Y. I., Wong, G., Humphrey, J. & Raj, T. Prioritizing Parkinson's disease genes using population-scale transcriptomic data. *Nat. Commun.* **10**, 1–10 (2019).

269. Gagliano, S. A. *et al.* Genomics implicates adaptive and innate immunity in Alzheimer's and Parkinson's diseases. *Ann. Clin. Transl. Neurol.* **3**, 924–933 (2016).

270. Bryois, J. *et al.* Genetic Identification of Cell Types Underlying Brain Complex Traits Yields

Novel Insights Into the Etiology of Parkinson's Disease. *Nat Genet* **52**, 482–493 (2021).

271. Nyaga, D. M., Vickers, M. H., Jefferies, C., Perry, J. K. & O'Sullivan, J. M. The genetic architecture of type 1 diabetes mellitus. *Mol. Cell. Endocrinol.* (2018) doi:10.1016/j.mce.2018.06.002.

272. Wang, Z., Xie, Z., Lu, Q., Chang, C. & Zhou, Z. Beyond Genetics: What Causes Type 1 Diabetes. *Clin. Rev. Allergy Immunol.* **52**, 273–286 (2017).

273. Mitchell, K. J. What is complex about complex disorders? *Genome Biol.* **13**, 1–11 (2012).

274. Sherwood, L. *Human physiology: from cells to systems*. (Cengage learning, 2015).

275. Lonsdale, J. *et al.* The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).

276. Schmiedel, B. J. *et al.* Impact of Genetic Polymorphisms on Human Immune Cell Gene Expression. *Cell* **175**, 1701-1715.e16 (2018).

277. Wei, L. *et al.* Improved and promising identificationof human microRNAs by incorporatinga high-quality negative set. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **11**, 192–201 (2014).

278. Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.* **42**, 1001–1006 (2014).

279. Ross, E. G. *et al.* The use of machine learning for the identification of peripheral artery disease and future mortality risk. *J. Vasc. Surg.* **64**, 1515-1522.e3 (2016).

280. Jadhav, S., Kasar, R., Lade, N., Patil, M. & Kolte, S. Disease Prediction by Machine Learning from Healthcare Communities. *Int. J. Sci. Res. Sci. Technol.* 29–35 (2019) doi:10.32628/ijsrst19633.

281. Mameli, C. *et al.* The diabetic lung: Insights into pulmonary changes in children and adolescents with type 1 diabetes. *Metabolites* **11**, 1–16 (2021).

282. Odoardi, F. *et al.* T cells become licensed in the lung to enter the central nervous system. *Nature* **488**, 675–679 (2012).

283. Pitarokoili, K., Ambrosius, B. & Gold, R. Lewis Rat Model of Experimental Autoimmune Encephalomyelitis. *Curr. Protoc. Neurosci.* **81**, 9.61.1-9.61.20 (2017).

284. Eriksson, N. *et al.* Novel associations for hypothyroidism include known autoimmune risk loci. *PLoS One* **7**, 1–8 (2012).

285. Zheng, P. & Kissler, S. PTPN22 silencing in the NOD model indicates the type 1 diabetes-associated allele is not a loss-of-function variant. *Diabetes* **62**, 896–904 (2013).

286. Galvani, G. & Fousteri, G. PTPN22 and islet-specific autoimmunity: What have the mouse models taught us? *World J. Diabetes* **8**, 330 (2017).

287. Eliopoulos, E. *et al.* Association of the PTPN22 R620W polymorphism with increased risk for SLE in the genetically homogeneous population of Crete. *Lupus* **20**, 501–506 (2011).

288. Zhou, M., Guo, X., Wang, M. & Qin, R. The patterns of antisense long non-coding RNAs regulating corresponding sense genes in human cancers. *J. Cancer* **12**, 1499–1506 (2021).

289. Latgé, G., Poulet, C., Bours, V., Josse, C. & Jerusalem, G. Natural antisense transcripts: Molecular mechanisms and implications in breast cancers. *Int. J. Mol. Sci.* **19**, (2018).

290. Adam Moser, Kevin Range, and D. M. Y. PTPN22 alters the development of T regulatory cells in the thymus. *J Immunol.* **23**, 1–7 (2012).

291. Rodríguez-Rodríguez, L. *et al.* The PTPN22 R263Q polymorphism is a risk factor for rheumatoid arthritis in Caucasian case-control samples. *Arthritis Rheum.* **63**, 365–372 (2011).

292. Ran, F. A. *et al.* Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* **8**, 2281–2308 (2013).

293. Postollec, F., Falentin, H., Pavan, S., Combrisson, J. & Sohier, D. Recent advances in quantitative PCR (qPCR) applications in food microbiology. *Food Microbiol.* **28**, 848–861 (2011).

294. Iwaki, H. *et al.* Genetic risk of Parkinson disease and progression: An analysis of 13 longitudinal cohorts. *Neurol. Genet.* **5**, (2019).

295. Redondo, M. J. *et al.* A type 1 diabetes genetic risk score predicts progression of islet autoimmunity and development of type 1 diabetes in individuals at risk. *Diabetes Care* **41**, 1887–1894 (2018).

296. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science (80-.).* **337**, 1190–1195 (2012).

297. Cheung, V. G. *et al.* Polymorphic cis- and Trans-Regulation of human gene expression. *PLoS Biol.* **8**, (2010).

298. Mahr, S. *et al.* Cis- and trans-acting gene regulation is associated with osteoarthritis. *Am. J. Hum. Genet.* **78**, 793–803 (2006).

299. Franzén, O. *et al.* Cardiometabolic risk loci share downstream cis- and trans-gene regulation across tissues and diseases. *Science (80-.).* **353**, 827–830 (2016).

300. Williamson, R. M. *et al.* Prevalence of and risk factors for hepatic steatosis and nonalcoholic fatty liver disease in people with type 2 diabetes: the Edinburgh Type 2 Diabetes Study. *Diabetes Care* **34**, 1139–1144 (2011).

301. Kohlgruber, A. & Lynch, L. Adipose tissue inflammation in the pathogenesis of type 2 diabetes. *Curr. Diab. Rep.* **15**, 1–11 (2015).

302. Cheng, W. W., Zhu, Q. & Zhang, H. Y. Identifying risk genes and interpreting pathogenesis for Parkinson's disease by a multiomics analysis. *Genes (Basel).* **11**, 1–14 (2020).

303. Bailey, J. C., Kinzy, T. & Schiltz, N. Genetic analysis of self-reported glaucoma from the Health and Retirement Study. *Invest. Ophthalmol. Vis. Sci.* **60**, 5419 (2019).

304. Ryan, D. P. & Matthews, J. M. Protein–protein interactions in human disease. *Curr. Opin. Struct. Biol.* **15**, 441–446 (2005).

305. Geisterfer-Lowrance, A. A. T. *et al.* A molecular basis for familial hypertrophic cardiomyopathy: A β cardiac myosin heavy chain gene missense mutation. *Cell* **62**, 999–1006 (1990).

306. Volkov, P. *et al.* A genome-wide mQTL analysis in human adipose tissue identifies genetic variants associated with DNA methylation, gene expression and metabolic traits. *PLoS One* **11**, e0157776 (2016).

307. Szodoray, P. *et al.* Altered Th17 cells and Th17/regulatory T-cell ratios indicate the subsequent conversion from undifferentiated connective tissue disease to definitive systemic autoimmune disorders. *Hum. Immunol.* **74**, 1510–1518 (2013).

308. Houtman, M. *et al.* T cells are influenced by a long non-coding RNA in the autoimmune associated PTPN2 locus. *J. Autoimmun.* **90**, 28–38 (2018).

309. Hapfelmeier, A. & Ulm, K. A new variable selection approach using Random Forests. *Comput. Stat. Data Anal.* **60**, 50–69 (2013).

310. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. *Nat. Genet.* **50**, 1171–1179 (2018).