Universidade do Minho
Escola de Engenharia
Departamento de Informática

Débora Alves Antunes

Computational methods for the identification of genetic variants in complex diseases

April 2021

Universidade do Minho
Escola de Engenharia
Departamento de Informática

Débora Alves Antunes

Computational methods for the identification of genetic variants in complex diseases

Master dissertation
Master Degree in Bioinformatics

Dissertation supervised by
Miguel Francisco Almeida Pereira Rocha
Joel Perdiz Arrais

April 2021

Acknowledgments

First of all, I'd like to thank my advisors Miguel Rocha and Joel Arrais for all the support they gave me during the development of this work. Specially to Joel who always tried to provide me the best learning opportunities. He gave me the liberty and responsibility to develop my work, but was always available to help me, even with all the things happening in his life, like the birth of his second son. For both of them I give my best wishes.

I want to thank Daniel Martins, a doctoral student that was always there to help me. He had so much work to do, but always found a time to answer my questions or listening to my concerns. I wish you a future full of happiness and good opportunities.

To all my friends and colleagues at the Master's degree, who made me feel at home in Braga, a city where I didn't know anyone. Thank you for all the laughs, all the support and all the moments. I will cherish them for the rest of my life.

To my friends from Biology and EFC/AAC, I thank you for all for being there for me and providing me moments to relax. Specially Santos, that not only heard all my concerns and gave me advice, as remembered me constantly that there are no bioinformatics without "bio".

Finally, I thank my family, particularly my father and mother, for always trying to give me the best opportunities. Even during the tough times, they never stop believing in me, and for that I'm forever grateful.

Statement of Integrity

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledged the Code of Ethical Conduct of the University of Minho.

Abstract

Complex diseases, as Type 2 Diabetes, are not only affected by environmental factors but also by genetic factors involving multiple variants and their interactions. Even so, the known risk factors are not sufficient to predict the manifestation of the disease. Some of these can be discovered with Genome-Wide Association Studies that detect associations between variants, such as Single-Nucleotide Polymorphisms, and phenotypes, but other approaches, like Machine Learning, are needed to identify their effects and interactions. Even though these methods can identify important patterns and produce good results, they are changeling to interpret.

In this project, we developed a predictor for complex diseases that uses datasets from Genome-Wide Association Studies to help the identification of new genetic markers associated with Type 2 Diabetes. The pipeline developed integrates gene regions and protein-protein interaction networks in datasets of variants, extracts new features, and employs machine learning models to predict risk of disease.

This study showed the models can predict the risk of disease and using gene regions and protein-protein interaction networks improves the models and provides new information about the biology of the disease. From these models it was possible to identify new genes and pathways of interest which, with further investigation, could lead to the development of new strategies for diagnosis, prevention and treatment of Type 2 Diabetes.

Keywords: Complex diseases, Type 2 Diabetes, Genetics, Genome-Wide Association Study, Machine Learning, Bioinformatics.

Resumo

Doenças complexas, como Diabetes Tipo 2, são tanto causadas por fatores ambientais como por fatores genéticos que envolvem múltiplas variantes e as interações entre elas. Mesmo assim, os fatores de risco conhecidos não são o suficiente para prever a manifestação da doença. Alguns destes fatores podem ser descobertos em *Genome-Wide Association Studies* que detetam associações entre variantes, como polimorfismos num único nucleotídeo, e fenótipos, contudo são necessárias outras abordagens, como por exemplo Aprendizagem Máquina, para identificar os seus efeitos e interações. Mesmo quando estes métodos conseguem identificar padrões e obter bons resultados, estes são difíceis de interpretar.

Neste trabalho, desenvolvemos um algoritmo para doenças complexas que utiliza dados obtidos em *Genome-Wide Association Studies* para auxiliar na identificação de novos marcadores genéticos associados à Diabetes Tipo 2. A abordagem desenvolvida combina conjuntos de dados de variantes com a infomação das regiões de genes e redes de interações entre proteínas, extrai novas características, e utiliza modelos aprendizagem de máquina para prever o risco de doença.

Este trabalho mostra que os modelos conseguem prever o risco de doença e que o uso de genes e de redes de interação entre proteínas melhora os seus resultados, assim como também fornecem novas informações sobre a biologia da doença. Usando esta abordagem é possivel identificar novos genes e redes metabólicas de interece, que com investigação adicional, podem levar a criação de novas estratégias de diagnóstico, prevenção e tratamento da Diabetes Tipo 2.

Palavras-Chave: Doenças complexas, Diabetes Tipo 2, Genética, Estudos de Associação no Genoma Completo, Aprendizagem Máquina, Bioinformática.

Contents

Abbreviations

ALT Alternative Allele. 30, 31
ANN Artificial Neural Networks. 15, 16
AUC Area Under Curve. 19, 39, 43, 44

CDCV Common Disease Common Variant. 21
CDRV Common Disease Rare Variant. 21
ChIP-Seq Chromatin Immunoprecipitation Sequencing. 5
CNV Copy Number Variant. 4, 5, 21

DAVID Database for Annotation, Visualization and Integrated Discovery. 39, 41

FDR False Discovery Rate. 27

GO Gene Ontology. 40
GT Genotypes. xi, 30, 31
GWAS Genome-Wide Association Study. vii, 1, 2, 21–24, 27, 46

HWE Hardy-Weinberg Equilibrium. 23, 24, 30

IBS Iberian Populations in Spain. 30
INDEL Insertion/Deletion. 4, 30

LD Linkage Disequilibrium. 21, 22, 24, 27

MAF Minor Allele Frequency. 23–25
ML Machine Learning. 1, 2, 10, 11, 14, 16, 28, 29, 34
MS Mass Spectrometry. 5

NGS Next Generation Sequencing. 5, 7

PCA Principal Component Analysis. 17, 24, 34
PPI Protein-Protein Interaction. xi, 33, 34, 45, 46

REF Reference Allele. 30, 31
RNA-Seq RNA Sequencing. 5
ROC Receiver Operating Characteristic. 19, 39

SNP Single-Nucleotide Polymorphism. 1, 4, 21–31
SVM Support Vector Machines. vii, 12, 34, 37, 39, 43

t-SNE t-distributed Stochastic Neighbor Embedding. 17, 34
T2D Type 2 Diabetes. vii, 2, 8, 9, 29, 37, 43–47

VCF Variant Call Format. 29–31

WB Western Blotting. 5
WES Whole-Exome Sequencing. 5
WGS Whole-Genome Sequencing. 5, 46

List of Figures

List of Tables

1 Introduction

1.1 Motivation

Nowadays, it is believed that both genetic and environmental factors have a role in nearly every human disease, though several diseases do not have significant risk factors associated. Those genetic factors can be a mutation in a single gene that is transmitted to the descendants following Mendelian rules, causing a Monogenic (or Mendelian) disease. Still, most times, the diseases are Multigenic (or complex) and are caused by mutations in a group of genes that individually do not have any effect [1].

Since the first time that the whole human genome was sequenced in 2001, new approaches have been developed that contributed to the discovery of these genetic factors. A Genome-Wide Association Study (GWAS) is an analysis that identifies which genetic markers across the genome of different individuals, usually Single-Nucleotide Polymorphism (SNP), are associated with their traits. This approach has shown important results in the discovery of variants in Monogenic diseases, but in more complex diseases this contribution is less effective as, not only several variants have a part in the phenotype's development, but also the environmental factors can alter these traits [2, 3].

There are models that use SNP to test an individual's risk of developing a disease. The Polygenic Risk Score combines the effect sizes of the genetic variants and calculates a score used to predict risk of disease, and even though it is an easy and effective model, it does not account for complex data and their interactions [4]. Machine Learning (ML) models identify patterns from data and make predictions using statistical assumptions and mathematical algorithms. This approach can be supervised, when the prediction is based on training examples, or unsupervised, if no labels are given [5]. Though these models account for complex data interactions, they need big datasets to make correct predictions, are more difficult to apply and the effects of the genetic variants are challenging to interpret [4]. Lately, some evidence has appeared that introducing biological information into ML models helps to improve the disease's risk prediction, identify new genetic variants interactions and learn more about the underlying biological network [5].

Even with these approaches, the understanding of these diseases is limited, namely, in which specific genetic factors are involved, how far they interact with themselves and the environment and how they relate to the phenotypes. Innovative strategies are needed to better explain these diseases, since this can lead to the discovery of new biomarkers and treatments for a disease.

1.2 Goals

The main goal of this project is to develop a complex disease predictor from Genome-Wide Association Study datasets, that can identify new genetic variants associated with the phenotype, specifically, Type 2 Diabetes. In more detail, we aim to:

1. Develop a pipeline that can use datasets of genotypes to predict risk of disease;
2. Relate variants with their biological function using several databases, preparing datasets to be used in the disease predictor;
3. Create ML models that can predict the risk of disease based in the effect of genetic variants.

1.3 Structure

The chapter Biological Data and Disease Biology describes the genome, the flow of genetic information within a biological system and some of the data provided by these processes. This second chapter also performs a quick showdown of the available sequencing technologies and, in the end, characterizes a complex disease, Type 2 Diabetes, that will be a case study for this project.

The chapter Machine Learning explains what are Machine Learning algorithms and their categories, gives details on multiples models and specify some methods for model evaluation.

The fourth chapter shows a methodology for GWAS data, including preprocessing and association tests commonly used, and highlights some advantages and challenges of the use of Machine Learning in this context.

The pipeline of this study is described in the fifth chapter, the Development. Starts with the description of the datasets used, as well as its processing. Then it explains the integration of the biological information and the extraction of new features.

In the sixth chapter, all the results are shown, as well as quick description. Lastly, the Discussion and Conclusion explain the results taking into account the context of the problem, describes some advantages and disadvantages and suggest future approaches for these type of studies.

2 Biological Data and Disease Biology

Each organism has a unique genetic sequence essential for the functioning of all cells and respective processes. The genetic information is carried by DNA molecules named nucleotides and is structured in chromosomes. The human genome has more than 3 billion nucleotides organized in 23 pairs of chromosomes, 1 of them being the sexual chromosomes, XY for male and XX for female. Every nucleotide is formed by a sugar, deoxyribose, one of the four nitrogenous bases, adenine (A), thymine (T), cytosine (C) and guanine(G), and at least one phosphate group. A gene is a sequence of nucleotides that encodes a final product, generally RNA or a protein [6].

The RNA is composed of nucleotides formed by a different sugar named ribose, and with an uracil base (U) that substitutes the thymine. Besides being used for the synthesis of proteins, RNA molecules have also structural, regulatory, and catalytic functions on the cell. Proteins are complex molecules built from amino acids used for countless processes, for example, catalyzation of reactions, structure, transport, signaling, storage, gene regulation, among others. All these molecules are involved in many reactions and form complex pathways [6].

Although organisms of the same species have very similar genomes, there are variants that make each organism unique. These variants can be Single-Nucleotide Polymorphisms (SNPs), Insertion/Deletions (INDELs), substitutions, inversions, translocations or Copy Number Variants (CNVs) and are described in Table 1 [7].

2.1 Omics Data

The cell biology is complex and involves a variety of molecules and interactions. All of these elements can provide important data to understand the processes within the cell, tissue, organ or even organism. Omics refers to any biological study field that makes a global and comprehensive analysis of a specific set of molecules [8].

There are different types of omics data: genomics, epigenomics, transcriptomics, proteomics, metabolomics and microbiomics [8, 9]. Their characteristics and methods for detection are specified in Table 2.

Table 1.: Specific types of variants.

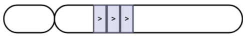
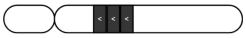
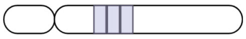
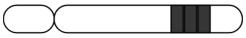
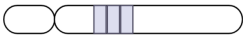
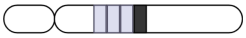
Variant	Description	Example: Reference	Example: Alternative
Single-Nucleotide Polymorphism	Isolated changes in a single base compared to the reference genome	TTG**A**CGTA	TTG**A**CGTA
Insertion	Addition of a sequence of nucleotides	TTGACGTA	TTGA**TG**CGTA
Deletion	Loss of a set of nucleotides	TTG**AC**GTA	TTGGTA
Substitution	Alteration of a set of bases, maintaining the same length	TTG**AC**GTA	TTG**TA**GTA
Inversion	Reversion of sequences of DNA compared to the reference genome		
Translocation	Relocation of a region of nucleotide sequence to a new position		
Copy Number Variant	Local repetition of segments of DNA with an elevated number of bases		

Table 2.: Types of omics data, their characteristics and methods for detection.

Omics	Description	Technologies
Genomics	It is the study of the whole genome and its main goal is to identify genetic variants that play an important role in a disease, a response to a treatment or a patient prognosis	Microarrays; Whole-Genome Sequencing (WGS); Whole-Exome Sequencing (WES); Targeted Sequencing
Epigenomics	This field studies chemical modifications of the DNA that are associated with the regulation of gene transcription	Chromatin Immunoprecipitation Sequencing (ChIP-Seq)
Transcriptomics	Focuses on detecting, quantifying and examining RNA transcripts	Microarrays; RNA Sequencing (RNA-Seq)
Proteomics	It is used to identify and quantify peptides encoded by the genome	Mass Spectrometry (MS); Western Blotting (WB)
Metabolomics	Detects and/or quantifies metabolites to learn more about a biological system	NMR Spectroscopy and Liquid or Gas Chromatography associated with Mass Spectrometry (MS)
Microbiomics	It is a field that analyses a microbial community to detect environmental alterations	Specific Next Generation Sequencing (NGS)

2.2 Sequencing Techniques

For many years the most used technique for DNA Sequencing was Sanger sequencing, that is based on the selective incorporation of labelled dideoxynucleotides (ddNTPs) in DNA chains during *in vitro* DNA replication [10, 11]. These ddNTPs are similar to deoxyribonucleotides (dNTPs) but without the 3' hydroxyl group. Because the lack of this group, ddNTPs cannot bond with the 5' phosphatate group of the next nucleotide, and the DNA chain cannot be extended further. After the amplification of the DNA sample, Sanger sequencing performs four simultaneous and very similar reactions, each containing a different ddNTP to be incorporated, producing DNA fragments of different sizes. Theses fragments are ran through a polyacrylamide gel electrophoresis and the result shows the position of the different nucleotides [12] (Fig. 1).

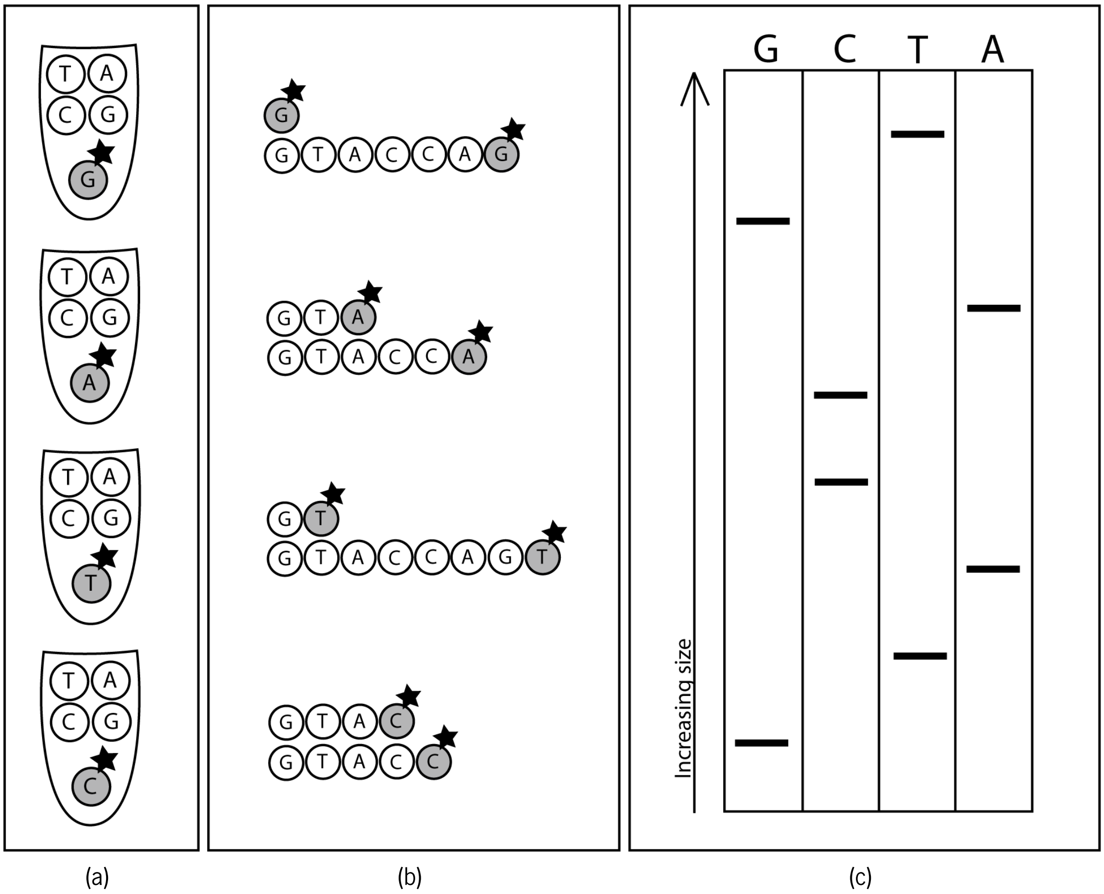


Figure 1.: Sanger sequencing. (a) DNA polymerization reactions, each containing all dNTPs and a different labelled ddNTP that once incorporated stops the reaction (b) These reactions produce DNA fragments of different sizes. (c) Polyacrylamide gel electrophoresis shows the order of the nucleotides based in the sizes of the DNA fragments.

Currently, Sanger sequencing is only cost-effective for small sequencing studies, and Next Generation Sequencing (NGS) technology arose as a cheaper method with a high sample throughput, discovery power and sensitivity. NGS has several sequencing platforms and all of them need a sequence library preparation. In most cases, this pre-processing starts with the fragmentation of the DNA molecules into reads with a specific size, followed by the attachment of adapters to each end of the read. Some of these technologies require an amplification of the reads [11, 13]. There are four main approaches to sequence the template: pyrosequencing, sequencing by synthesis, sequencing by ligation and ion semiconductor sequencing.

Pyrosequencing uses a bioluminescence signal to detect the base and needs an initial amplification because it is not sensitive enough to detect the light of individual nucleotides [13]. This amplification is made by attaching one DNA molecule per bead followed by an emulsion PCR. During synthesis, a manipulated DNA polymerase extends the primer pausing after incorporating each complementary dNTP. When a dNTP is added, a cascade of enzymatic reactions is catalyzed, generating peaks of light. The record of the time and intensity of these peaks gives information about the order of the nucleotides [11, 12].

Depending on the technology, sequence by synthesis, may need and initial amplification step [11]. Illumina platform makes use of a solid support with attached primers, promotes hybridization between these molecules and the DNA fragments and uses bridge amplification to form clusters of repeated DNA fragments. During a last synthesis, one by one, a fluorescently labelled dNTP with a blocked 3' end is incorporated in a growing DNA chain (each base has a different color). With each incorporation, the signal is recorded, the 3' end terminator and fluorophore are removed, and the cycle is repeated [11, 12, 13]. Other sequence by synthesis technologies use the same principle, but with some variations, like immobilized DNA templates or DNA polymerase, instead of primers, and one-colour labels instead of four-colour [11].

Sequence by ligation is a cyclic method that requires amplification of the DNA reads made by emulsion PCR and uses DNA ligase to hybridize fluorescently labelled two-base-encoded probes to the DNA template. First, one primer is added to a specific position of the sequence, followed by a set of cycles that consist in hybridization, visualization of the label and cleavage of the end segment of the probe along with the fluorophore. Subsequently, primers for minus one position are added and the process repeats. At the end, the signals are aligned by the positions and with this colour-space map the DNA fragment can be decoded [11, 13].

The last approach, ion semiconductor sequencing, also needs emulsion PCR amplification. This method uses two primers, one is attached to the bead and the other is in the solution, that hybridizes with the DNA template ensuring an uniform orientation of the molecules. During DNA polymerization, each time a dNTP is incorporated, a hydrogen ion is released causing changes in the pH. This sequencer has a sensor that detects these changes. This way, when no dNTP is incorporated, the pH does not change, if one dNTP is incorporated, a change is detected, and if over one dNTP is incorporated, the change in pH is proportional. Although, this is a method that sequence by synthesis, because no changes are made in the dNTPs and no fluorescence is used, turns it in a whole new approach [13].

2.3 Case Study - Type 2 Diabetes

Diabetes mellitus is a disorder characterized by high blood sugar levels (hyperglycemia) caused by a deficit in the production, or defective usage, of an essential hormone called insulin. Low levels of insulin for a long period of time cause damages in several organs and lead to life-threatening complications like cardiovascular diseases, nerve damage, kidney failure and sight impairment [14, 15].

This condition affects 463 million adults worldwide, and is responsible for 4.2 million deaths. Global health expenditure on diabetes is about USD 760 billion (689 billion euros) annually. There is evidence that the prevalence of this disorder increases by age, is slightly lower in women than men, and more common in individuals that live in urban areas than in rural areas. Worldwide, the prevalence changes by regional distribution, being North America and Caribbean the region with the highest values and Africa with the lowest [14].

Type 2 Diabetes (T2D) is a subtype of diabetes that accounts for 90% of diabetes worldwide. In this subtype the hyperglycemia is caused by insulin resistance which means that the cells cannot respond to the hormone, leading not only to high blood sugar levels but also an increase in insulin production. Typically, T2D affects adults, however, there is a growth in the number of children and adolescents diagnosed with this disorder. This condition is very difficult to be diagnosed and in some cases might be completely symptomless. For this reason, it is estimated that one-third to one-half of the individuals with this disorder are undiagnosed [14, 16, 17].

T2D is associated with obesity, physical inactivity and inadequate diet and can be prevented or managed with better lifestyle and healthier habits such as a controlled diet, regular physical activity, decreasing or stopping the consumption of alcohol, tobacco and other substances, and the maintenance of a healthy body weight. Nowadays, the treatments available for this condition are oral medication like metformin, sulphonylureas, dipeptidyl peptidase 4 inhibitors and glucagon-like peptide 1 analogues or in more extreme cases, insulin injections [14, 16].

Besides this association with environmental factors, there are evidences that T2D is strongly influenced by genetic factors [18]. Acording to Stančáková and Laakso [19], until 2016, there were identified more than 80 variants for this condition, but those only explained about 10% of the T2D variability within a population, thus denoting a problem known as missing heritability.

Missing heritability ($\pi_{missing}$) (1) is defined as the proportion of heritability of a trait not explained by the set of known variants. To calculate the missing heritability is necessary to find the ratio of explained heritability ($\pi_{explained}$) (2).

$$\pi_{missing} = 1 - \pi_{explained} \tag{1}$$

$$\pi_{explained} = \frac{h^2_{known}}{h^2_{all}} \tag{2}$$

h^2_{all} is an inferred value for the maximum variance explained by the additive contribution of all the allele counts, including the discovered and undiscovered, whereas h^2_{known} only accounts for the contribution of known variants. Some explanations for the existent missing heritability suggest that there might be a few variants not yet found because they are only present in a small percentage of the population or have smaller effects. Others explore the fact that the calculation of missing heritability is based on an addictive model and do not account for more complex associations like gene-environment and gene-gene interactions [19, 20].

Although some genetic variants for T2D are known, they produce little information that can be used in a medical context. To increase the usefulness of this information in predicting the disease risk and in the development of new strategies for diagnosis, prevention and treatment is necessary to identify new variants and consider gene-gene and gene-environment interactions [19].

3 Machine Learning

Machine Learning (ML) is a branch of artificial intelligence that, as previously mentioned, extracts knowledge from data using statistical assumptions and mathematical algorithms.

The first step in the development of ML algorithms is the collection, preprocessing and transformation of the raw data into features. These features are used in the training of the algorithm and posterior testing of the model, or directly on the testing. Testing the model provides information about the performance, effectiveness and accuracy of the algorithm, allowing for its improvement. In the end, the ML algorithm is used to predict feature behavior, help understand underlying structures, among other usages [21].

Supervised learning and Unsupervised learning are two categories of ML used in a variety of studies and are explained in more detail below. However, there are not the only available categories of ML algorithms.

Semi-supervised learning uses a dataset with labelled samples and, in a higher quantity, unlabelled samples. The necessity for the development of this type of learning came from the difficulty of obtaining labelled data, that resulted in smaller datasets and consequently in models with inferior accuracy. With the incorporation of unlabelled samples in these datasets, easier to obtain, more information about theproblem is provided, creating models with higher quality [21, 22, 23].

In reinforcement learning there is an agent that interacts with an environment and tries to reach a goal related to the state of the environment. Through several steps the agent perceives the current state of the environment, executes an action. This action triggers a change in the state of the environment and produces a reward. In the next step the agent perceives the reward and the new state and chooses another action and so forth. The objective of this type of algorithms is to learn a policy, that is a mapping of the actions to be executed in the different states of the environment, by maximizing the numerical reward signal. The agent does not know the best actions to take, instead it must learn by a trial-and-error search. Depending on the type of problem, the action may not influence the immediate reward but subsequent rewards [21, 22, 24].

3.1 Supervised Learning

Supervised learning algorithms try to understand the relationship between the independent or "input" variables (x) and the dependent or "output" variables (y). It is called supervised because the output is known and used to train the ML model, guiding the recognition of patterns and the prediction of outputs in other data. Therefore, the dataset used is a set of N examples $(x^{(i)}, y^{(i)})$, $i = 1, ..., N$; and each element $x^{(i)}$ has n features that characterizes it: $x_j^{(i)}, j = 1, ..., D$ [22].

This algorithm is used in regression problems when y is a continuous value, or in classification problems, if y is a discrete value. The objective in regression is to calculate an output value for each example, whereas in classification is to assign it a label from a list of two (binary classification) or more possibilities (multiclass classification) [5, 25, 26].

In this approach, when a model is well-trained it can make predictions with high accuracy, finding hidden information in data not yet tested, however, the quality of the training data largely influences how well the model is trained [21].

A few of the most applied supervised learning models are described in the following sub-sections.

3.1.1 Linear models

Linear models try to predict the relationship between the input x and output y variables using a linear function and can be used for both regression and classification problems. The training process of these models aims to find the parameters w, associated with the features of x, that help the model make the most accurate predictions. Some linear models used are linear regression, logistic regression, Ridge regression and LASSO, which differ on how the model parameters are calculated and how the model complexity is managed [26].

Linear regression is used to find a linear association between the multiple features and the output. This is done by finding the optimal values w that minimizes the mean squared errors between the predicted outputs and the real outputs y [22, 26, 27]. Although linear regression is a simple model that only fits some data, it is the basis of many complex learning methods [27].

Logistic regression is an algorithm used in classification problems and predicts the probability of an input $x^{(i)}$ having a label $y^{(i)} = 1$. This model is based in a sigmoid function that calculates values between 0 (negative label) and 1 (positive label) and the estimation of the parameters w is usually done by maximum likelihood estimation [21, 22, 28].

Ridge regression and LASSO use an additional constraint for regularization, and besides trying to find the optimal values of w for the model, also tries to fit this new factor [26]. The aim of both methods is to correct the model and avoid overfitting, and they differ from each other in the way they penalize the parameters w. Ridge regression uses this constraint to make the parameters w close to zero. Hence, each feature will have the smallest effect possible on the outcome without impair the final prediction [26]. The constraint used in LASSO can make a parameter $w = 0$ and consequently distinguish between features to ignore and essential features. One downside is that it can create underfitting. This type of regularization is called L1 regularization, while the one used in ridge regression is called L2 regularization [22, 26].

3.1.2 Support Vector Machines (SVM)

SVM models could be used in regression problems but is more commonly used for classification. Using these model as a classifier is based on finding a hyperplane that best separates the data into two classes. These data have a p-dimensional space and the corresponding hyperplane is a plane of dimension $p - 1$ that divides it into two parts. From this hyperplane it is easy to determine on which side a sample $x^{(i)}$ lies [29].

Another characteristic of these models is the maximization of the margin of the hyperplane. This margin is the maximum distance between the closest examples of each class and the hyperplane. Finding this hyperplane with the largest margin is straightforward in cases where the data is separable. However, in many cases there are outliers that prevent the hyperplane to perfectly separate the data or the data cannot be separated by a plane [22].

To better fit the model it is necessary to choose the kernel which defines the set of mathematical functions that quantify the similarity between two samples $x^{(i)}$ and $x^{(i')}$. When the data is linearly separable, the best results are usually produced by a linear kernel (1) which is the simplest function and faster to train. However, in many cases, the data cannot be separated by a linear function. The sigmoid kernel (2), polynomial kernel (3) and radial basis function kernel (4), are based, respectively, in sigmoid, polynomial and radial functions, where $gamma$ is the kernel coefficient, C in the constant term and d is the polynomial degree.

$$K(x^{(i)}, x^{(i')}) = \sum_{n=1}^{p} x_n^{(i)} x_n^{(i')} \tag{1}$$

$$K(x^{(i)}, x^{(i')}) = tanh(\gamma \sum_{n=1}^{p} x_n^{(i)} x_n^{(i')} + C) \tag{2}$$

$$K(x^{(i)}, x^{(i')}) = (\gamma \sum_{n=1}^{p} x_n^{(i)} x_n^{(i')} + C)^d \tag{3}$$

$$K(x^{(i)}, x^{(i')}) = exp(-\gamma \sum_{n=1}^{p} (x_n^{(i)} - x_n^{(i')})^2) \tag{4}$$

3.1.3 Decision Trees

Decision Tree learning constructs a model based in recursive splitting rules (also called tests), with each node being a test in a feature, each branch connecting the possible decision value and the next feature and each final node (called leaf) showing the output value. Therefore, the predicted output can be determined by following the path from the starting node until a leaf is reached [21, 26]. These models can be used in classification and regression problems [25].

To build this type of model, all tests are explored, using the one that better divides the training data. An optimal split is the one that separates the different outputs into two regions with most accuracy. With each division, these regions are becoming purer until each has only one output. However, when this state is reached, the model already contains too much structure creating a risk of overfitting [25, 26].

To control this risk, some strategies used are *pre-pruning*, which stops the creation of the tree before this state of purity is reached, and *post-pruning* that after the creation of the tree removes or collapses less informative nodes [26]. The *pre-pruning* can be done by limiting the size of the tree or number of leaves, or by stopping the creation in a less pure state while *post-pruning* requires an analysis of the subtree and posterior raising or replacement [26, 27]. Although the first approach is more attractive because facilitates the process of creating a tree, saving time and computational power, some of the important relations between nodes can be lost, making the tree less informative. For this reason most trees are build and then pruned [27].

3.1.4 Ensemble Learning

Ensemble learning integrates various ML models with low accuracy to form a new model with high accuracy. Typically, the low accuracy models used are produced with simple algorithms and have difficulty in learning from complex datasets but when their results are combined, these limitations are surpass [22, 26].

Two methods of ensemble learning are boosting and bagging. Both can be used in classification problems, by choosing the label with highest averaged probability, or in regression problems simply by averaging the results from all the used models. The difference relies on the weight of this result.

In boosting, some models have a higher influence and consequently their results have a higher weight. This happens because a first model is created from the original training data, but the following models are trained to complement the previous model, focusing on correcting their errors [22, 27]. On the other end, in bagging all models are build based in different variations of the original training data. All of these variations have the same size and are created by randomly deleting and copy some samples. This way, all results have an equal weight [22, 27].

Random forest is a learning algorithm based in the bagging method that compiles multiple decision trees. The method used to train these models is similar to bagging because it uses variations of the original training data. However, more randomness is added using different sets of features to create them. This factor is important because decision trees use the strongest features to split the dataset and by working with these, all models would be highly correlated and the accuracy would have almost no improvement [22, 29].

3.1.5 k-Nearest Neighbors

This simple algorithm is applied in classification problems and uses a set of *k* objects to label the test sample. It is based on the assumption that close points have the same label, so the distance between the unlabeled sample and all labeled examples is calculated to identify the k closest neighbors. In the end, the label that occurs more often in the group is used to label the test sample [21, 28].

3.1.6 Artificial Neural Networks

This model was inspired in the brain neurons and the way they process information and learn from the environment. For this reason, Artificial Neural Networks (ANN) have several nodes connected by edges that receive an input from previous nodes and transmit an output to the next node. The final output is decided by performing multiple and successive calculations [26]. ANNs are organized in an input layer that receives all values from $x^{(i)}$, an output layer y and intermediate hidden layers with M nodes $b^{(k)}$, $k = 1, ..., M$. Each node $y^{(k)}$ receives as input the corresponding weights w^{ik} and calculates the output with a chosen activation function. The number of layers l is given by counting the number of non-output layers because the input layer is considered the first layer [21, 22, 28].

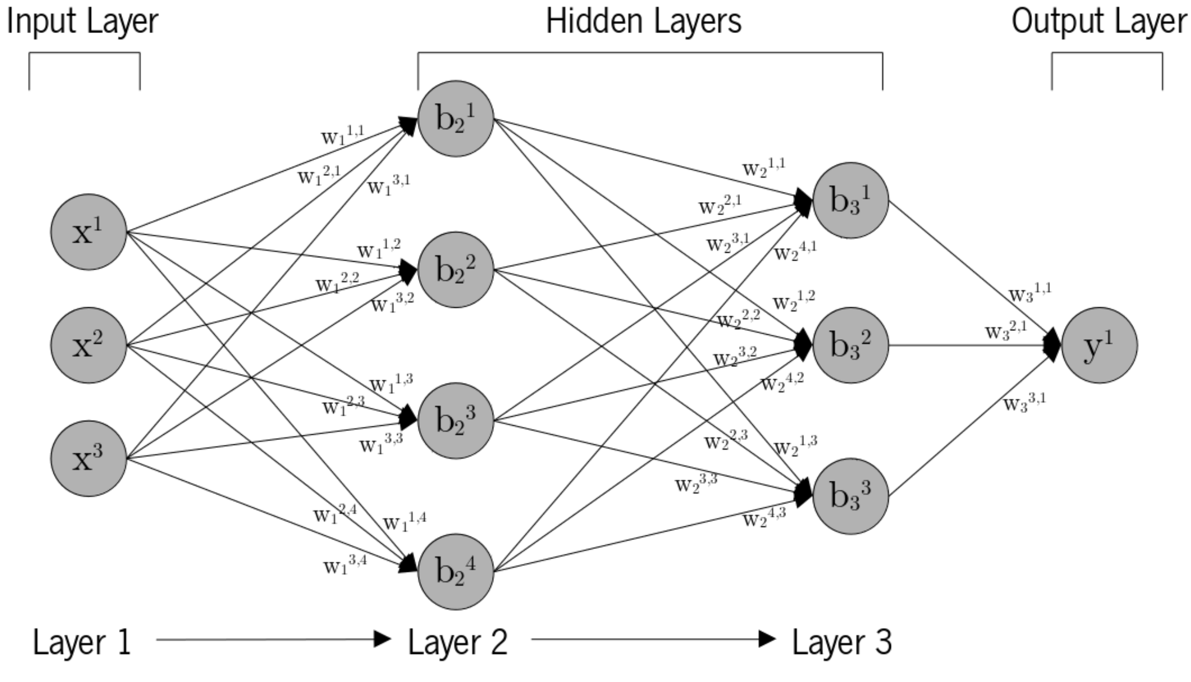


Figure 2.: Schema of an artificial neural network model.

In binary classification and regression problems, the output layer y is composed of one node. If the ANN is a regression model the activation function is linear, but if it is a classification model the activation function is a logistic function. Multiple classification problems needs over one node in the output layer [22].

Deep Learning is a class of ML that uses ANN trained with multiple layers. Although increasing the number of layers can improve accuracy, too many layers can create overfitting. For this reason, the optimal number of layers varies according to the type of problem, the quality of the dataset among other reasons [21, 22, 26].

3.2 Unsupervised Learning

Unsupervised learning algorithms are constructed to find structure in datasets with a set of N inputs $x^{(i)}$, $i = 1, ..., N$ and no output (or label) associated [22]. These models are based in probability distribution, statistical tests, and other types of measurements and unlike supervised learning algorithms, there is no distinction between train and test data [21, 28]. Algorithms like this one are used in clustering, dimensionality reduction, anomaly detection, among others [22].

3.2.1 Clustering

The goal of clustering is to find the subgroups, named clusters, that better separate the data. This way, similar data is in the same cluster, whereas different data is in distinct clusters. From this partitioning, it is possible to predict to which group a certain example belongs. Models, like *k*-means Clustering and Hierarchical Clustering, differ from each other in their choice of what is similar and what is not, and in the way they present the output [26, 29].

A hierarchical cluster can be built from the bottom-up (agglomerative), in which the samples are assigned to an individual cluster and then the clusters are sequentially merged, or from the top-down (divisive), in which all data is assigned to a general cluster and consecutively separated into different clusters. Both methods define similarity based in the linkage between clusters. Single, complete and average linkage are defined as the minimum, maximum and average distance between the data of one cluster and the data of another cluster, respectively. The output is a dendrogram which is a tree-like visual representation of the samples and their similarity [25, 29].

The *k*-means Clustering method splits the dataset in k different clusters. This algorithm starts by randomly placing k centroids μ^j, $j = 1, ..., k$ into the feature space. The next step is a loop that assigns each $x^{(i)}$

to its closest μ^j, and recalculates the centroid positions. The loop stops when none of the observations is assigned to a different μ^j and, consequently, its position is unchanged. The output is a list with the assigned clusters [29, 30].

3.2.2 Reduction of Dimensionality

It is called dimensionality reduction to the process of representing data with a high dimensional space in a new space with a lower dimensionality. Some advantages of this process is that smaller dimensions require less computational power, lead to results with higher accuracy and simplifies data visualization [28]. Even though not all methods of dimensionality reduction use unsupervised learning, there are some important to refer.

One of these is Principal Component Analysis (PCA). PCA starts by computing new features called Principal Components (PCs) that are uncorrelated with the original features. These PCs are positioned in the data and will define a new coordinate system [21, 26]. The first PC is the axis that goes in the direction that better explains the variability of the data, the second PC is orthogonal to the first and goes in the direction of the second highest variance in the data, and the same occurs to the other dimensions. To reduce dimensionality the data is projected in this new coordinate system [22].

t-distributed Stochastic Neighbor Embedding (t-SNE) is another method for dimensionality reduction that embeds the information from higher dimensions to lower dimensions. This method focuses on preserving the local structure of the data, in contrast with PCA, that preserves the global structure. This technique has other advantages, one of them being the better handling of outliers [31, 32].

3.3 Model Evaluation

As previously mentioned the dataset is separated into a training set, to prepare the model for a given problem, and into a test set, to evaluate the error of the finalized model. The dataset can also be divided into a validation set that is used to optimize its parameters [27]. All the samples are divided between these three sets, without repetitions, using a proportion of 70% to 95% for the training set, and an equal proportion of 2.5% to 15% for both the validation and testing sets [22].

It is important to recognize if the model is underfitting or overfitting. The first means that the model has a high bias and maps poorly the trend of the data, having difficulty in predicting the output values of the training data. It may be caused by an oversimplification of the model or the use of little or uninformative features. On the other hand, overfitting means that the model has a high variance and does not generalize well the problem. On these cases, the model is very accurate in the prediction of the output in training sets but is unsuccessful in predicting outputs in the test set. A model too complex or too many features can create overfitting problems [22, 26].

One of the methods used to asses and enhance the performance of the model is cross-validation. In this method the training set is randomly divided in *k-folds*, or partitions, and sequentially one of the *folds* is used as a validation set while the others are used to train the model (Fig. 3). Each model is then evaluated by a metric of choice and in the end, all metrics are averaged [22, 26].

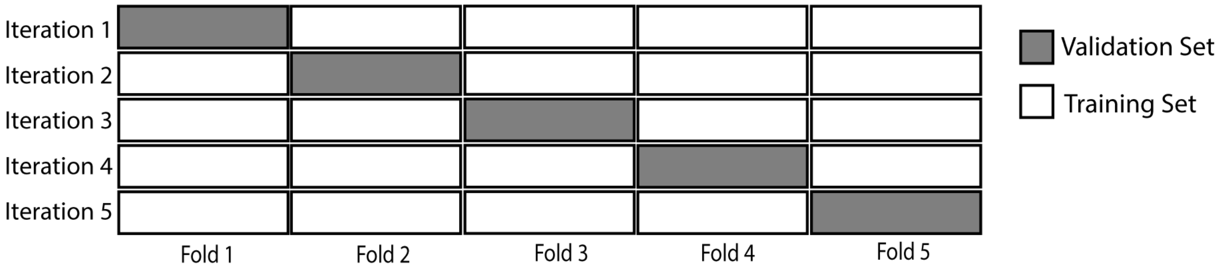


Figure 3.: Schema of division and sequential interactions within a 5-*fold* cross-validation.

A confusion matrix is a table used to examine the prediction's quality of the classification model. It compares the labels predicted with the real labels, organizing the results in four categories: true positives, false negatives, false positives and true negatives (Fig. 4) [21, 22, 26].

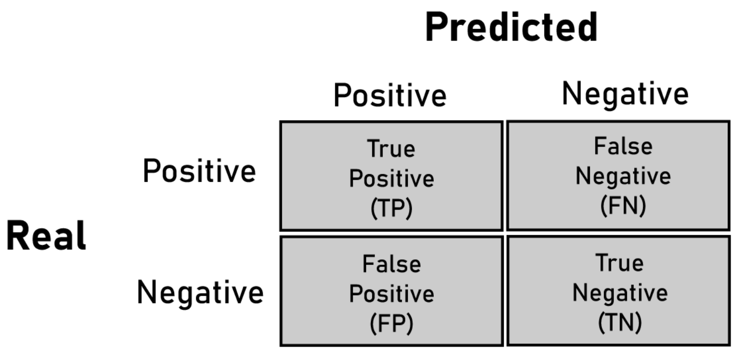


Figure 4.: Confusion matrix for binary classification where True Positive (TP) is the number of outputs correctly labeled positive, False Negative (FN) is the number of outputs incorrectly labeled negative, False Positive (FP) is the number of outputs incorrectly labeled positive and True Negative (TN) is the number of outputs correctly labeled negative.

In the Table 3 are described some metrics to evaluate the quality of the model for classification problems, using the values from the confusion matrix. For regression problems, the estimation methods described in Table 4 are calculated using the arrays of real outputs (y) and the correspondent predicted outputs ($\hat{y}$).

Table 3.: Principal metrics for model evaluation in classification problems where TP is the number of true positives, FN is the number of false negatives, FP is the number of false positives and TN is the number of true negatives.

Name	**Description**	**Formula**
Accuracy	Proportion of correct predictions among all predictions	$\frac{TP+TN}{TP+TN+FN+FP}$
Precision	Proportion of real positives among all predictions classified as positive, high values reveal a low number of false positives	$\frac{TP}{TP+FP}$
Recall or True Positive Rate (TPR)	Proportion of predicted positives among all real positives, high values reveal a low number of false negatives	$\frac{TP}{TP+FN}$
Specificity or True Negative Rate (TNR)	Proportion of predicted negatives among all real negatives	$\frac{TN}{TN+FP}$
False Positive Rate (FPR)	Proportion of real negative labels incorrectly predicted	$\frac{FP}{FP+TN}$
F1-score	Value that summarizes the relation between precision and recall	$2 \times \frac{Precision.Recall}{Precision+Recall}$

Receiver Operating Characteristic (ROC) curve is a graphic method for model evaluation that builds a summary for the classification performance in a range of thresholds using the values of the False Positive and True Positive rates for the x and y axis, respectively. If the model makes good predictions, the curve will be close to the top-left corner. The value of Area Under Curve (AUC) summarizes the results of the ROC curve. If the model makes perfect predictions, the AUC will be equal to 1. On the other hand, an AUC of less that 0.5 can be representative of a bad model or might suggest a mistake in the labelling of data [22, 26].

Table 4.: Principal estimation methods for model evaluation in regression problems where y is the array of real outputs and $\hat{y}$ are the predicted outputs.

Name	**Description**	**Formula**
Mean Squared Error (MSE)	Mean squared error between the real and predicted values	$\frac{1}{N}\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2$
Mean Absolute Error (MAE)	Mean absolute error between the real and predicted values	$\frac{1}{N}\sum_{i=1}^{N}\lvert y^{(i)} - \hat{y}^{(i)}\rvert$
Coefficient of Determination (R^2)	Measure of the quality of prediction for unseen samples by calculating the proportion of explained variance	$1 - \frac{\sum_{i=1}^{N}(y^{(i)} - \hat{y}^{(i)})^2}{\sum_{i=1}^{N}(y^{(i)} - \frac{1}{N}\sum_{i=1}^{N}y^{(i)})^2}$

4 Methods for detection of SNP associations

For Mendelian diseases, linkage studies based in the patterns of gene inheritance have been useful to localize their causal SNPs, but in complex disease these studies do not produce strong results [2].

The research on these diseases rests in two important hypotheses. The first, Common Disease Common Variant (CDCV), states that complex disease traits are established by common variants (frequency greater than 1%) in human populations [33, 34]. The second, Common Disease Rare Variant (CDRV), describes that the additive effect of multiple rare variants (frequency smaller than 1%) determines the complex disease phenotype [33, 35].

With the arrival of the Genome-Wide Association Study (GWAS), the CDCV hypothesis was proven by the identification of multiple common variants associated with several complex diseases. However, this association only accounts for a low percentage of the heritable component. The missing heritability could be accounted by environmental interactions, other types of variability like CNVs or even interactions between variants. Nevertheless, the CDRV could also be a possible explanation [33].

4.1 Genome-Wide Association Study

Genome-Wide Association Study (GWAS) is an analysis for the detection of associations between genetic variants, normally focused in SNPs, and phenotypes, inside a population.

GWAS genotyping is based on the principle of Linkage Disequilibrium (LD) [36]. LD quantifies the degree of nonrandom association between alleles of two or more different loci. This measurement can be affected by different forces like mutation, recombination, natural selection, population size, among others [3, 37, 38]. Considering two loci with the alleles M_1,m_1 and M_2, m_2, the LD between the alleles m_1 and m_2 is given in (1), being q_1 and q_2 the corresponding allele frequencies and q_{12} the frequency of the haplotype $m_1 m_2$. If $D = 0$ then $q_{12} = q_1 q_2$, implying that there is a random association between the two alleles. Although informative, the D value is highly dependent on population allele frequencies and for this reason there are two other statistics commonly used [2].

D' (2) is the ratio of D to its maximum absolute value and has a range of $[0, 1]$. If $D' = 1$ then at least one of the possible haplotypes is absent and no recombination has occurred since the last mutation. The squared correlation coefficient, r^2 (3), is used to assess linearity within the sample size and also has a range of $[0, 1]$. When there is a perfect association between the two alleles, $r^2 = 1$ [2, 37].

$$D = q_{12} - q_1 q_2 \tag{1}$$

$$D' = \begin{cases} \frac{D}{min[q_1 q_2, (1-q_1)(1-q_2)]} & if D < 0 \\ \frac{D}{min[q_1(1-q_2), (1-q_1)q_2]} & if D \geq 0 \end{cases} \tag{2}$$

$$r^2 = \frac{D^2}{q_1(1-q_1)q_2(1-q_2)} \tag{3}$$

Looking at these values is possible to find pairs of SNPs that are strongly correlated with each other, forming LD blocks. From these blocks are selected tag-SNPs that are genotyped in these studies [36]. GWAS use chip-based microarray technology and there are two principal platforms used in these studies, Affymetrix and Illumina, that differs in methodology, price and specificity [39].

One objective of these studies is to use these associations to predict the disease risk for a given individual and produce knowledge about the underlying biological molecules and processes for the development of new diagnosis, prevention and treatment strategies, However, these relationships are not easy to understand because of their complex pathways and growing number of variables [3, 40]. The use of GWAS in this type of studies has a few obstacles. Firstly, there is a high correlation between neighboring variants which hinders the identification of causal SNPs. Also, the large number of variants needs to be reduced before performing the analysis. These procedures are very prone to discard relevant variants. On top of that, this type of studies requires an extended quality control to ensure the reliability of the results.

4.2 Quality Control for GWAS Data

To regulate the presence of false positives or false negatives in the analysis of GWAS is necessary to perform quality control of the data. Some errors like missing data or mistakes in genotype calling can be made while performing GWAS and can be specific for the individuals or SNPs [2, 41, 42]. Referring to the individual-specific quality control problems, there may exist discordant sex information, missing genotypes or heterozygosity rate, duplicates and relatedness and divergent ancestry. As for SNP-specific, some common quality control problems are missing genotypes, deviation from the Hardy-Weinberg Equilibrium (HWE) and low Minor Allele Frequency (MAF) [41, 42].

Missing genotypes are a problem that can be addressed by removing individuals or SNPs with low quality. To choose the appropriate threshold is necessary to analyze the distribution of the data, making sure that only a small proportion of data is excluded, and the study is not compromised [41, 42].

The distribution of the mean heterozygosity can be examined to determine the heterozygosity rate across all individuals and detect whether the genetic data for a given individual is concordant with the HWE or not. On that regard, the Wright's inbreeding coefficient F relies on the observation of departures to that equilibrium, either motivated by excess of heterozygotes or homozygotes, and offers an advantage because it is independent of the MAF [42]. The inbreeding coefficient of an offspring F_O (4) is calculated from the inbreeding coefficient of the ancestor F_A and the number of generations connecting each parent (n and n') [43]. Taking into account the non-sexual chromosomes and calculating the mean value of heterozygosity, it is possible to identify outliers. Anomalously high heterozygosity is an indicator of sample contamination and anomalously low heterozygosity of inbreeding, and for this reason, these outliers are removed from the data [41].

$$F_O = \sum[(\frac{1}{2})^{n+n'+1}(1+F_A)] \tag{4}$$

With the data from the X chromosome is possible to check for discordance in sex information. Because males only have one copy of this chromosome, they do not have heterozygous SNPs associated with the X chromosome. On the other hand, females have two copies of this chromosome and for this reason a higher heterozygosity. A discordance in gender is reported when the heterozygosity rate does not match the reported sex. For males, this ratio should be lower than 0.2, for females should be higher than 0.8 [41, 42]. Although many studies do not account for difference in gender, a discordant sex may suggest an error during sequencing that will affect the downstream investigation, and for this reason should be taken into account [2].

Duplicated or related individuals share some same alleles, so including them in the study creates an over representation of genotypes and a bias in the allele frequencies of the population [41]. For quality control of duplicates and relatedness it is necessary to prepare a new dataset, pruned for LD and then calculate the IBS (identity by state) for each pair of individuals by averaging the proportion of shared alleles. The degree of shared ancestry is called IBD (identity by descent) and is calculated from IBS. $IBD = 1$ represents duplicates or monozygotic twins, $IBD = 0.5$ means a first-degree relation, $IBD = 0.25$ is for second-degree relatives and an $IBD = 0.125$ describes a third-degree relation. The threshold is normally set to an $IBD = 0.1875$, an intermediate between second and third-degree relatives [41, 42].

The population structures should be equally distributed in the case and control groups or association caused by ancestry may be encountered. A graphic representation of a PCA made on these datasets detect different population stratification, and after identification of these outliers, the corresponding individuals are removed [41].

The Hardy-Weinberg Equilibrium (HWE) suggests that the allele frequencies are constant across generations. This law represents the frequencies of the genotypes AA, Aa and aa as p^2, $2pq$ and q^2, being p the allele frequency of A and $q = p - 1$ the allele frequency of a [44]. In GWAS studies, SNPs associations can create a deviation from the HWE, however, extreme deviations are considered errors in the genotype calling [2, 42] The threshold for extreme deviation is dependent of the number of SNPs tested [2].

Other filter used as quality control is Minor Allele Frequency (MAF) and when is low implies that rare genotypes are present in the dataset. Besides reducing the variant call certainty and augmenting the number of false positives by changing the properties of the statistical tests, an association between these genotypes is difficult to detect and for these reasons it is better to remove them from the dateset [2, 42]. The threshold used is dependent of the sample size n and could be around a MAF$= 10/n$ [42].

4.3 Single SNP Association Analysis

In single SNP association analysis, an association between each SNPs and the phenotype is tested. In some cases, the SNPs are treated as independent and the LD structure is ignored. Some of the methods used are generalized linear models, statistic tests like χ^2 test and Cochran-Armitage trend test and Bayesian approaches [2, 45].

The generalized linear models estimate a linear relationship between SNPs and phenotypes. If the output is binary, a logistic regression model is created, but if the output is continuous, a linear regression model is used. One advantage of using this method is that the formula can be easily changed to account for additive, dominant and recessive effects or even add other factors [2].

The χ^2 (5) tests the null hypothesis (H_0) of no association with the disease [46], observing the deviance between n observed samples (O_i) and their theoretical values(E_i) [47]. In this case, each SNP is considered independent of the other, but when additive, dominant and recessive models are considered, the Cochran-Armitage trend test is used [46]. These types of tests need less computational power but with more complex data do not produce so satisfactory results [2].

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i + E_i)^2}{E_i} \tag{5}$$

The goal of the Bayesian method is to compute a p-value for the null hypothesis and uses a frequentist approach [48]. The accuracy of the results depend on the threshold used for MAF and the size of the dataset, because this approach uses the Bayes factor (8) that measures the probability not only of the null hypothesis (H_0) but also of the alternative hypothesis (H_1). Here the H_0 is the logarithm of the odds ratios (θ) between the heterozygote and the common homozygote θ_{het} and between rare and common homozygotes θ_{hom} are equal to zero (6), and the H_1 is that θ_{het} and θ_{hom} are equal to their prior distribution $t1$ and $t2$, respectively (7) [48].

$$H_0 : \theta_{het} = \theta_{hom} = 0 \tag{6}$$

$$H_1 : \theta_{het} = t1, \theta_{hom} = t2 \tag{7}$$

$$BF = \frac{P(data|H_1)}{P(data|H_0)} \tag{8}$$

4.4 Multiple SNP Association Analysis

Multiple SNP association analysis examines the link between the phenotype and the combine effect of multiple SNPs. Some methods used for single SNP association can be modified and used for multiple SNP association, like the linear models and the Bayesian approach. Various types of analysis have been proposed to account for these associations: haplotype-based methods, SNP-SNP interaction models and tests based in biological knowledge [2].

Often, in these studies, there are too many SNPs to examine, so it is necessary select a subset of features to analyze [49]. Haplotypes are groups of genes that are inherited together and can be used in association studies [38]. These methods test the association between haplotypes and phenotype. One reason for considering these methods is that proteins and their physical proprieties are determined from linear sequences of nucleotides, generally haplotypes inherited from the progenitors which gives them a direct biological relevance[49]. Also, relevant alleles but with a lower frequency might only be identified when studying whole regions [2].

The definition of epistasis differs according to the area of study considered, and in this context is important to differentiate between biological epistasis and statistical epistasis. The first occurs in each individual and accounts for interactions between genetic variants, biological molecules and pathways, and their effects in the phenotype, as the second involves the interactions between genetic variants that cause a deviation from the models that consider the independent effects of these variants inside a population [50, 51]. SNP-SNP interaction models are based in statistical epistasis since biological epistasis alone is very difficult to study and the connection between the two types is not fully understood [50].

The most direct approach to study these interactions is a thorough analysis of all combinations of SNPs. Though strategies like dimensionality reduction, data compression, and others, allow for exhaustive search in two-locus combinations, a higher number of locus needs a computational power not available at the time, so filtration of variants is required [51]. These variants can be filtered by their independent effects in which only SNPs above a certain threshold are used [52]. However, SNPs with low independent effect may have an important effect when combined with other variants [51]. Other method of filtration uses one nearest neighbor based approach that measure the distance between individuals according to the number of different genotypes between them and attributes a weight to each SNP. With this method, SNPs with a higher weight are used in the SNP-SNP interaction models [53].

Another approach for this problem analyzes the association between variants integrating various types of biological information, namely the genes, exons, proteins and pathways related to them, and their interactions. Although this method facilitates the interpretation of the biological meaning of these SNPs, is heavily dependent of the biological information quality, requiring a rigorous control of what information is used [51].

4.5 Correction for Multiple Testing

To avoid errors in this type of association studies, it is important to make a correction for multiple testing thus accurately estimate significance thresholds [2, 46].

Bonferroni correction is a simple method that adjusts the significance level α to α/N, being N the number of tested SNPs. This way, any hypothesis with a p-value inferior or equal to α/N is rejected. One disadvantage of this method is the assumption that all SNPs are independent, being relatively conservative for GWAS. Another approach is the use of the False Discovery Rate (FDR) to fix an expected proportion of false positives among the associations considered significant. However, this method is highly dependent of different factors and do not account for association between variants. Permutation procedures correct the null hypothesis by comparing the original p-values with the generated empirical distribution of p-values. This distribution is originated from repeating the original tests in randomly permutated data. This approach can preserve the LD structure, but is more computationally demanding than Bonferroni correction or FDR [46, 54].

4.6 Machine Learning Approaches

Machine learning approaches are valuable for processing large datasets, integrate various types of omics data and can be applied to a variety of problems like identification of binding sites, genes, biomarkers and prediction of functions, outcomes or phenotypes [55, 56, 57]. One important use for these approaches is the analysis of SNPs, produced in GWAS that can provide information about the relation between variants and phenotypes in complex diseases and be a means to improve diagnosis, treatments and prognosis accuracy [55].

Some ML models used in this context are Neural Networks, Support Vector Machines and Random Forests [57]. However, these models face some challenges and limitations in this framework. For starters, the development and choice of model needs to take into account the complex relationship between combinations of SNPs and in some cases, between other genetic variations and environmental factors [58].

Usually the datasets used have a small amount of samples but numerous features and when used, the models produce misleading results [55]. For this reason the ratio for a robust model should be 5 to 10 features per sample [59]. In case the number of features is excessive, there are algorithms that select which SNPs are going to be used by the model. Filtering algorithms favour the subset of features with better quality or more relevance. Wrapper algorithms choose iteratively a subset of features to be classified by the model using a deterministic or stochastic algorithm. Although filtering methods are faster, wrapper methods are more powerful because they do not discard any attributes [58].

Some ML models have a complex mathematical basis, and even when the results are very precise, if these statistical relations cannot be interpreted in a biological context, the underlying mechanism of a complex disease remains unknown. Without these new insights it is difficult to create alternative strategies for diagnosis, treatment or prevention of complex diseases [58].

5 Development

The goal of this work is to identify new variants associated with a complex disease, in this case T2D, by developing a predictor of disease risk that uses SNPs. This predictor will be based in ML classifiers that would be able to distinguish between healthy individuals (controls) and T2D patients (cases).

Considering that the biology of this problem was very complex, highly variable between individuals and influenced by many factors, by studying it from a computational perspective, it was essential to specify a few assumptions. The first one was that during a person's lifetime the genetic code did not endure many changes. This assumption is acceptable, although it is known that the DNA undergo some changes, most of the genome is conserved. Other assumption made was that an individual from the case dataset has a higher risk of developing T2D and, consequently, an individual from the control dataset has a low risk of developing the disease. If two individuals had been exposed to the same environment, we could claim that the outcome had been caused by genetic factors. However, during this study we did not had access to the physiological data or family history of the patients. Due to this, we could have patients with lower risk of disease that with an inadequate diet or a sedentary lifestyle developed T2D, or a patient with higher risk of disease that have not yet developed the disease because of a controlled diet and regular physical activity. Even so, for the purpose of this study, we assumed that the case individuals had a higher risk of T2D.

5.1 Initial Datasets

The first step for the realization of this project involved choosing and preparing the datasets that would be used in the rest of this work. The case dataset originated from a gzip compressed Variant Call Format (VCF) file with a size of 556.58 Mb (2.01 Gb when uncompressed) that contained information for 71 exomes from Portuguese patients diagnosed with T2D and 57 142 453 loci. These patients were between the ages of 48 and 80 years old and had no congenital diseases. Although, medical features as age, sex and body mass index as well as family history are known risk factors and some of them are strong predictors of T2D, the focus of this study were the genetic factors and only the genomic data was used hereafter.

An initial filtering restricted the type of variants to SNPs or INDELs, reducing the dataset to 250 869 variants. The quality control of these data revealed that from the 71 individuals, two of them were related ("Ex41" and "Ex51")(Fig. 9). Since using data from related individual can create bias in the allele frequencies, "Ex51" was removed from the dataset. The criteria relied on the missing information, being the sample with more missing variants excluded from the study. An analysis of the HWE was also performed and about 3.9% (9 792) of the variants were removed because they did not follow this theorem (Fig. 10). Lastly, it was observed that some sites had a low quality score (Fig. 11 and Table 11). This quality score represents the probability of an incorrect base call. About 12 776 variants (5.3%) were removed by setting a minimum quality score of 20 that corresponds to a 1 in 100 chance that the SNP call is erroneous. Other aspects of the dataset were analyzed, however there was no need to perform further filtering at this point (Fig. 12, 13, 14 and Table 11). After this processing, that selected the variants and samples, the case dataset had a size of 161.74 Mb (737.62 Mb when uncompressed) and included 228 301 variants, comprising 225 070 SNPs (98.6%) and 3 231 INDELs (1.4%).

The control dataset resulted from a selection of VCF files collected from the 1000 Genome Project [60] whose goal was to discover the largest number of genetic variants with frequencies of at least 1% across different populations. For this study, there was selected data from the Iberian Populations in Spain (IBS) in the Phase 3 release, being the closest ethnic group to the Portuguese individuals of the case dataset. This group was chosen to prevent bias that would tend to distinct case and controls based on populational genetic divergences instead of the disease's risk prediction. The 1000 Genome Project IBS subpopulation included 107 samples, all collected in the Spanish territory and are from individuals born in Spain that had the previous two generations born in the same area [61]. The control data was in gzip compressed files, divided in chromosomes, with a total size of 16.21 Gb (approximately 600 Gb when uncompressed).

The VCF file is a standard tab delimited text file used to save genotype data because of its scalability, flexibility and unambiguity allowing the storage of millions of variants and extra information across the several fields. The first nine columns are used to describe the variant, indicating the chromosome, position, id, reference (REF) and alternative alleles (ALT), several measures of quality and the format of the subfields reported for each sample. The remaining columns correspond to the reported information for each sample including the called genotypes (GT). The GT field encodes the alleles using numbers where 0 is REF, 1 is the first allele in ALT, 2 is the second allele in ALT, etc. The separator specifies if the alleles are phased ("|"), which means that it is known from which chromosome the allele came from, or if they are unphased

("/") and the order of chromosome pairs is not taken into account. Regarding the case dataset, the GT were unphased and represented by "0/0" for homozygous by REF, "0/1" and "1/0" for heterozygous, and "1/1" for homozygous by ALT.

5.2 Dataset Construction

Up to this point, there were two different VCF files, however, to proceed it was necessary to represent the data in a simpler format. The first step was to translate these complex GT into numbers that could be interpreted by the computational methods we intended to use. This way a genotype "0/0" was translated to 0, "0/1" or "1/0" to 1, and so forth according to Table 5.

Table 5.: Translation of the Genotypes (GT) found in the VCF file for each variant into a numerical format that computational methods can interpret.

Genotype	0/0	0/1	1/1	0/2	1/2	2/2	0/3	1/3	2/3	3/3	0/4	...
Translation	0	1	2	3	4	5	6	7	8	9	10	...

Since these two datasets had different dimensions, the merging of these files into one dataset was required. This was possible by finding the common variants. A variant was considered equal if their location (chromosome and position) was the same and their REF and ALT were matching. In these datasets were found 172 629 variants in common. Next, a column named "labels" was added with a number attributed to each sample, with 1's being cases and 0's being controls.

Even though the resulting dataset was in the correct format, there was a portion of missing values that could impair the subsequent analysis. For this reason all variants with a rate of more than 10% of missing values were removed from the dataset and the remaining missing values were imputed by the most frequent value of the column. The final dataset had 177 samples and 168 715 variants.

5.3 Test Variant Significance

The first analysis made was a single SNP association using the χ^2 test that measures the probability of association of each variant with the disease, considering each one of them independent. Before this statistical test, the data was tested for normality using D'Agostino and Pearson's test [62] and a p-value of

0.05. The results showed that of 168 715 variants, only 321 did not follow a normal distribution (Fig. 5). Therefore we considered that all variants followed a normal distribution.

The χ^2 test was performed in all variants with a p-value of 0.05, as in the normality test. From 168 715 variants, 9 427 presented significantly different distributions between cases and controls, which means that these variants had a statistical association with the phenotype (Fig. 6).

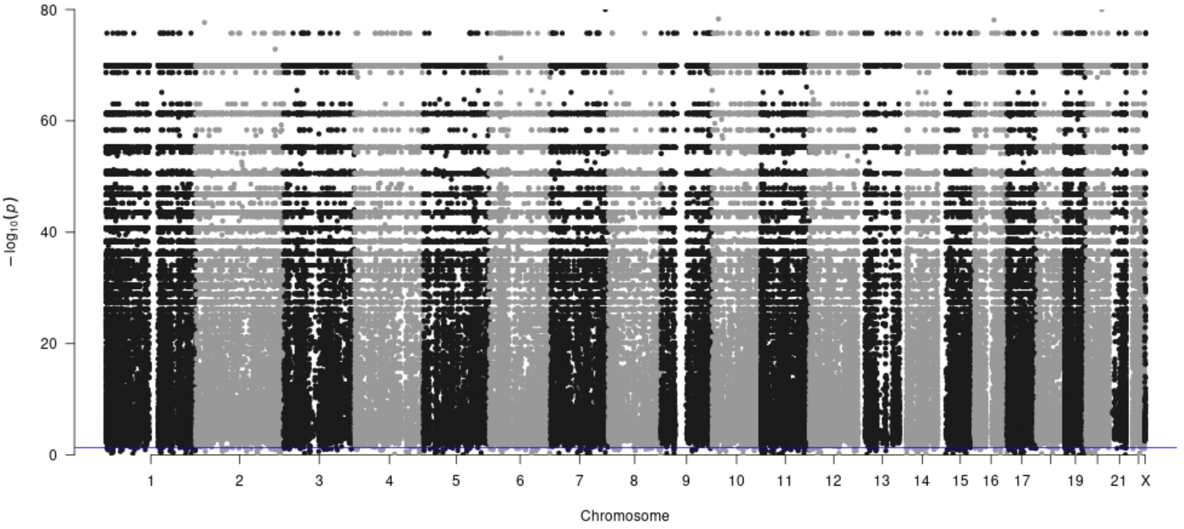


Figure 5.: Manhattan plot of the results from the D'Agostino and Pearson's normality test according to the position of each variant. Only 321 variants of 168 715 do not follow a normal distribution. The threshold $-\log_{10}(\text{p}-\text{value}) = 1.3$ is represented by a line.

5.4 Variant to Gene

The next step of this study was to combine the variants by translating them into genes. For this we used the package pyensembl, a Python interface that uses GTF and FASTA files from Ensembl. In this study we used as reference the GTF file from the GRCh37 version, and with a tool from this package we found the gene ID associated with the position of each variant.

At this point, we had a list of all variants, the associated gene if found, and the p-value from the χ^2 test. The variants were grouped in genes with an associated p-value corresponding to the average of all variants and when this p-value was less than 0.05 the gene was selected to continue the analysis. In total, 82 genes were selected and were referred henceforth by significant genes (Table 7).

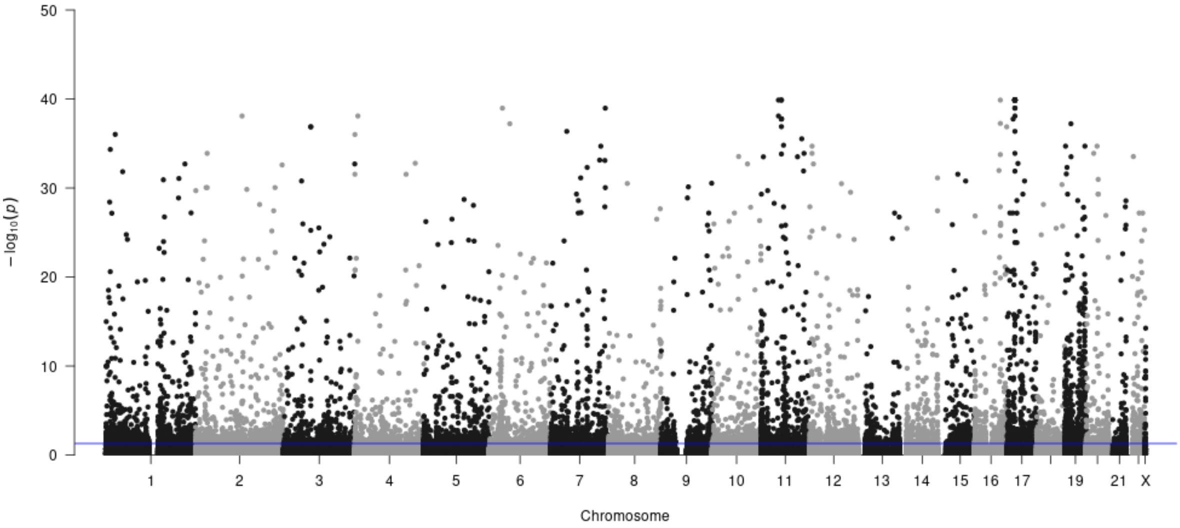


Figure 6.: Manhattan plot of the results from the χ^2 test according to the position of each variant. 9 429 variants of 168 715 had a significant association with the disease. The threshold $-\log_{10}(\mathrm{p} - \mathrm{value}) = 1.3$ is represented by a line.

Although the objective of this research is to identify new markers, we also studied the known risk genes. For this reason, after the gene grouping, other genes were selected. Using a list of 75 known risk genes from Type 2 Diabetes Knowledge Portal [63] (filter Prediction: T2D_related and CAUSAL), we selected the matching genes on our dataset. Since our data was only from the exome, not all risk genes were present, so at the end we had 67 risk genes to work with (Table 7).

5.5 Network of Protein-Protein Interactions

To research the relationship between these genes, we integrated them in a network of Protein-Protein Interactions (PPIs) that was represented by a graph whose nodes are genes and edges are interactions.

First, we downloaded human PPIs from a database, in this study we use BioGRID [64], a curated repository for biomedical interaction. These interactions were represented on a text file, and not only provided information about the interactors but also the interaction detection methods, confidence metrics, among others. The file was preprocessed, only selecting interactions with experimental data associated and with both interactors being human proteins. The final file with PPIs had 366 327 interactions.

To integrate all this information we uses a R package named *dmGWAS*. It enables the identification of subnetworks of highly connected genes, and consequently, the discovery of new candidate genes. Using as input the human PPIs file and a list of genes and respective p-values from our data, this tool implemented a dense module searching method and outputted a list of modules associated with the disease, ranked by significance. The top 50 modules were chosen and combined.

At this point, there was a subnetwork of PPIs significant to our study containing 252 genes. To select new candidate genes we used the R package *igraph*, applied three network metrics (Table 6) to each gene of this subnetwork and choose the genes that were in the top 100 of each metric. At the end, 77 genes were selected and are referred henceforth by central genes. From these 77 genes, three were in common with the known risk genes, CAV1, PCBD1 and WFS1 (Table 7).

Table 6.: Brief description of the three network metrics applied to each node of the PPI network.

Degree	Number of edges that a node has connected to it, including loop edges.
Betweenness	Number of shortest paths going through the node.
Closeness	Average length of the shortest path between the node and all other nodes in the graph.

5.6 Dimensionality reduction

At this point, there were 3 sets of genes selected, the 67 known risk genes, the 82 significant genes and the 77 central genes. To discover whether these sets of genes may be used to predict the disease, they were used as features in three ML models (SVM, decision tree and logistic regression). Even though these sets of genes are a fraction of the original dataset and considerably reduced the number of features, it was still composed of data from the variants that could overfit the ML models.

For this reason, a dimensionality reduction was applied to the dataset. As explained in a previous chapter, it transforms the data from a high number of features, some correlated and consequently redundant, into a reduced number of features (feature reduction) or a representation that maintains the overall properties of the original data (feature extraction).

First, feature extraction was applied to each gene, using the information present in the corresponding group of variants. We used two methods of dimensionality reduction, PCA and t-SNE, and, in both these methods, the first component was chosen to represent the data. Also, two statistics, mean and variance, were calculated to complement the information gather in the feature extraction. After this process, each one of the selected genes had only four features to describe it, so the final datasets, known risk genes, significant genes and central genes, had 268, 328 and 308 features, respectively.

Because the goal of this study is to identify new variants of interest and the datasets had a high number of genes, feature reduction was performed in two of these datasets, the significant genes and the central genes. For this we used Extremely Randomized Trees (Extra-Trees), which is a tree based ensemble method. This algorithm was trained 1000 times with each dataset, for every train the top 100 most important features were registered and, at the end, their frequency was calculated. For each dataset we selected the 25 features with higher frequency. The 25 features from the top 25 significant genes dataset belonged to 25 different genes, while the ones from the top 25 central genes dataset belonged to 12 different genes (Table 7).

Table 7.: Known risk genes, significant genes and central genes selected for this study. In the grey boxes are the genes selected in the dimensionality reduction. In bold are the common genes between the known risk genes and the central genes lists.

Known Risk Genes		Significant Genes		Central Genes	
ABCC8	PCSK1	AAMDC	MAST1	APP	MYC
AKT2	PDX1	AKT1S1	MSTN	ATXN1	NCL
ANGPTL4	PLCB3	AMMECR1L	MYPOP	BAG3	NEK6
ANKH	PNPLA3	ANP32A	NBPF14	BAIAP2	NFKBIA
APOE	POC5	B3GALNT1	NBPF4	BTRC	NSMF
APPL1	POLD1	BCL2L10	NDUFB6	CALM1	OPTN
BLK	PPARG	C10orf95	NGLY1	CASP1	**PCBD1**
BSCL2	PPP1R15B	C1orf162	NKX2-1	CASP8	PCNA
CAV1	PTF1A	C20orf202	NUFIP2	**CAV1**	PICK1
CDKN1B	QSER1	CDKN2C	OLIG1	CDC37	PIK3R1
CEL	RFX6	CGB5	OR13J1	CDH1	PIN1
EIF2AK3	RREB1	CHKA	OR2T5	CDK2	PLK1
ERAP2	SIX2	CLEC18A	P4HTM	CDKN1A	PPP1CA
GATA4	SIX3	CNOT11	PAGR1	CEP70	PTPN6
GATA6	SLC16A11	CSRP2	PARP11	DEAF1	RAC1
GCG	SLC19A2	CXCL13	PNMT	DISC1	RPS6KB1
GCK	SLC30A8	CXCL5	PNRC2	ENO1	SFN
GCKR	SLC5A1	DCAF16	POTED	ERBB2	SKP1
GIPR	TBC1D4	DDTL	PPP1R7	ESR1	SMAD3
GLIS3	TM6SF2	DEXI	PRAMEF13	GFAP	SPRED1
GLP1R	TRMT10A	DLEU1	PRDX6	GRB2	STK11
GRB10	WARS	DLX6	PRRT2	HLA-B	STX1A
HNF1A	**WFS1**	DOK1	PTRF	HNRNPC	SYK
HNF1B	WSCD2	GLIPR1	RAX	HSP90AB1	TGFBR2
HNF4A	ZFP57	GPR25	RNASE10	HSPA8	TNF
IGF1	ZNF771	HEXIM2	RNF182	HSPD1	TRAF6
IRS2		HFE2	RYBP	HTT	TRIM54
KCNJ11		HMOX2	S100A16	INCA1	TSC22D1
KLF11		HNRNPAB	SCG5	IQUB	UBC
LPL		HOXB8	SKIL	JPH3	UBE2Z
MC4R		HSD3B1	SOX21	KANK2	USP2
MNX1		ID2	SYT4	KCTD13	VCP
MTNR1B		IFNA13	TADA2B	KDR	**WFS1**
NAT2		IL33	TAF11	KIFC3	YWHAE
NEUROD1		IL36RN	TEX22	KRT34	YWHAG
NEUROG3		JOSD1	TLX1NB	LMO4	YWHAZ
NKX2-2		KCNMB4	TMEM178A	LNX1	
PAM		KRTAP5-1	TMEM60	LNX2	
PAX4		LSMEM1	TPST1	MAP3K1	
PAX6		MAFA	TRAM1L1	MDFI	
PCBD1		MAFF	WDR45B	MEOX2	

6 Results

As previously mentioned, this study was a case-control analysis with samples from individuals with and without T2D which translates to a computational problem of binary classification. The methods employed were used to detect the differences between the samples and to classify them accordingly. It was also necessary to analyze the results from a biological perspective. The following sections shown the results from these methods.

6.1 Classifier Optimization

When using classifiers, it is important to optimize the hyperparameters of the models to best fit the datasets. Hyperparameters are parameters that control the accuracy of the model but are not chosen when the algorithm is trained. To optimize these models it was performed a grid search for each algorithm and dataset. A grid search is an exhaustive search that compares the performance of each combination of parameters (given by the user) in a training data.

In this study we train the SVM, decision tree and logistic regression models for each dataset with the parameters shown in Table 8, using the class model_selection.GridSearchCV() from the package *scikit-learn* with a *5-fold* cross-validation.

Table 8.: Parameters and respective values tested for SVM, decision tree and logistic regression models during grid search.

SVM		**Decision Tree**		**Logistic Regression**	
kernel	[’linear’,’poly’,’rbf’,’sigmoid’]	n_estimators	[20,50,75,100]	penalty	[’l1’, ’l2’]
C	[0.25,0.4,0.5,0.55,0.75,1]	criterion	[’entropy’,’gini’]	C	[0.25,0.4,0.5,0.55,0.75,1]
tol	[1e-3,1e-4,1e-5]	min_samples_leaf	[1,2,3,5,10]	tol	[1e-3,1e-4,1e-5]
gamma	[25,50,75,100,150,’auto’]	min_samples_split	[2,4,5,8,10]	solver	[’liblinear’,’saga’]
degree	[1,2,3,5,10]	max_leaf_nodes	[2,20,50,75,100]	-	-

For the SVM model we used the class *svm.SVC* from the package *sklearn* and tested different parameters that highly influence the classifier. The first one is the *kernel* that define the set of mathematical functions used by the model. The *C* parameter is used for regularization, which prevent overfitting. By decreasing *C* the regularization is stronger. The stopping criterion is defined by *tol*, denoting that the algorithm will stop when the loss or score does not improve by at least the value of this parameter. The parameter *gamma* corresponds to a kernel coefficient for the non-linear functions and indicates the extent of the influence of each training example. Higher values of *gamma* tend to overfit the model. Lastly, the parameter *degree* is the degree of the *polynomial kernel* function and is ignored by all other kernels.

Referring to the Decision Tree model, it was used the class *ensemble.ExtraTreesClassifier* from the package *sklearn* and we also tested 5 different parameters. This model implements a number of randomized decision trees, defined by the parameter *n_estimators*, on various sub-samples of the dataset. In the end, the average of the results is calculated to improve accuracy and control overfitting. The parameter *criterion* measures the quality of the split by gini impurity, if *gini* is selected, or by information gain, if *entropy* is selected. Given that a leaf node is a node of the tree that do not have child nodes, the parameter *min_samples_leaf* sets the minimum number of samples required to be in each node when considering a split point, and the parameter *max_leaf_nodes* defines total number of leaf nodes in a tree. Finally, to set the minimum number of samples required to do a split is used the parameter *min_samples_split*.

The third model used in this study was the Logistic Regression implemented by the class *linear_model.LogisticRegression* from the package *sklearn*. The first parameters tested were *penalty* and *C*, both being used for regularization. In this model, the parameter *tol* also defines the tolerance for stopping criterion. The parameter *solver* sets the algorithm used in the optimization problem. Although there are more, in this study we only tested the *liblinear* and *saga* because they are the recommended for this type of study.

Although the grid search was performed in all the datasets and, in some cases, produced different results, the overall best set of parameters was selected and used for the study. The final parameters used are shown in the Table 9.

6.2 Models Evaluation

After preparing the datasets and the classifiers, it was necessary to assess the models. As previously mentioned, we prepare 5 different datasets for this study: the known risk genes that had 268 features, the significant genes with 328 features and the correspondent top 25 selected features, and the central

Table 9.: Parameters and respective values chosen for SVM, decision tree and logistic regression models after grid search.

SVM		Decision Tree		Logistic Regression	
kernel	linear	n_estimators	50	penalty	l1
C	0.25	criterion	entropy	C	0.4
tol	1e-3	min_samples_leaf	3	tol	1e-3
gamma	25	min_samples_split	5	solver	liblinear
degree	1	max_leaf_nodes	50	-	-

genes with 308 features and the correspondent top 25 selected features. For each dataset, the classifiers were run 1000 times using a 5-fold cross-validation and for the evaluation of the models we used three metrics, accuracy, F1-score and AUC. The results are shown in Fig. 7 and the ROC curves for each dataset and model can be seen in Fig. 15- 17.

Comparing the results obtained for each metric, the values were similar which shows the robustness of the results. The different models produced results with different accuracies, F1-scores and AUC. SVM models produced the worst results, followed by the Logistic Regression models. The best results were obtained using Decision Tree models, with mean values higher than 0.85. The differences between the results of the various datasets were consistent being the known risk genes dataset the one that produced the lower values. The best results were from the top 25 central genes dataset.

6.3 Functional Annotation

Previously, the significant genes and central genes dataset were selected, and after observing the results of the classifiers, we proceed to a functional annotation of the identified genes. Functional annotation is the process of adding a biological context to, for example, a variant, gene or protein. This information can be gene functions, known associations to a phenotype, metabolic pathways, among others.

To perform this functional annotation, we submitted three separate lists of genes, the known risk genes, the significant genes and the central genes, to the online platform Database for Annotation, Visualization and Integrated Discovery (DAVID) [65, 66]. This tool uses information from several databases to find the most relevant and over-represented biological terms related with the gene list provided. The annotations cover different categories as diseases, gene ontology terms, pathways, protein interactions, among others.

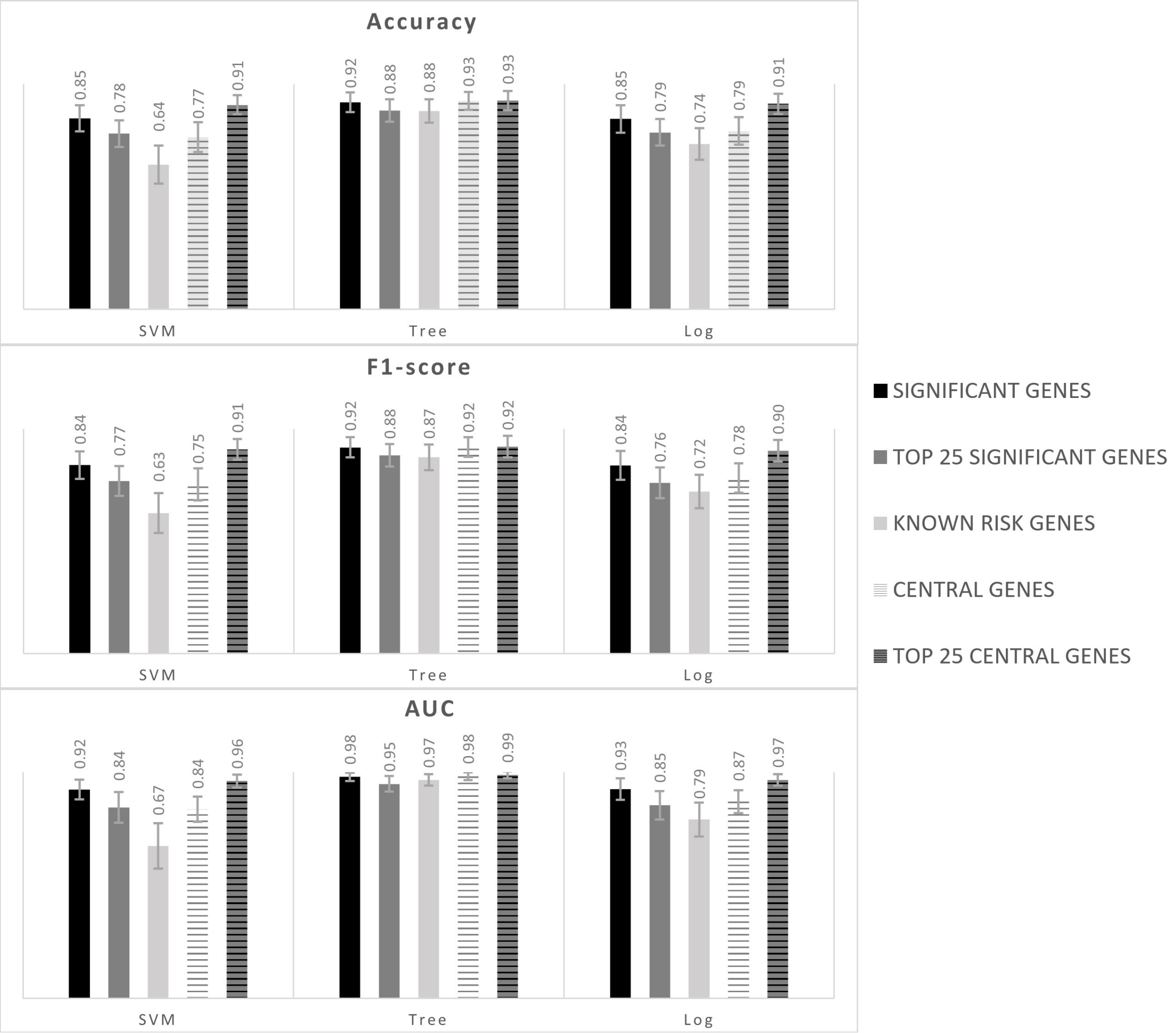


Figure 7.: Average accuracy, F1-score and area under curve, and respective standard deviation, of the SVM, Decision Tree (Tree) and Logistic Regression (Log) models, for significant genes (black), top 25 significant genes (dark grey), known risk genes (light grey), central genes (light stripes) and top 25 central genes (dark stripes).

Although the results from our submissions extend across the different categories and classes, the attention was focused in the results from the biological processes annotation from Gene Ontology (GO). All the genes from each list were grouped by the GO terms associated with them. These terms help to understand the biological activity of the gene but is not equivalent to the pathway. Some of the terms found on this analysis and the respective number of associated genes from each list can be observer in Fig. 8 (Complete list in Table 12).

From 29 683 biological process terms, there were found 160 terms in common between the genes of the lists. Looking at the terms with more associated genes, and knowing that most of the genes are different

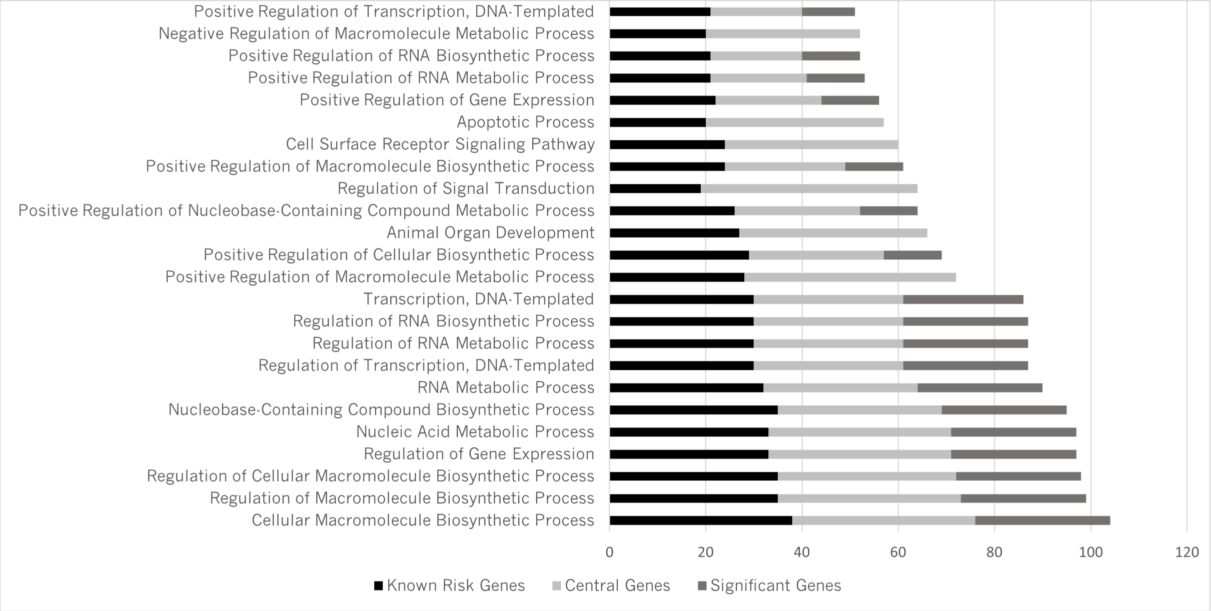


Figure 8.: Number of genes from the known risk genes, significant genes and central genes associated with the top biological processes found.

across the three lists, its observable that many of the identified genes (either significant or central) share the same terms as the known risk genes.

DAVID also highlights pathways that integrates two or more genes from the list. Even thought no pathway was provided for the significant genes, there were some pathways associated with the central genes. Some of most significant pathways are shown in Table 10, as well as brief description and the associated genes.

Table 10.: Top 9 significant pathways associated with the central genes.

Regulation of PLK1 Activity at G2/M Transition	**RHO GTPases activate PKNs**	**Chk1/Chk2(Cds1) mediated inactivation of Cyclin B:Cdk1 complex**
PLK1 is a protein that phosphorylates several proteins involved in transition from phase G2 to phase M of mitosis	RHO GTPases is a family of proteins involved in processes like changes in morphology and mitosis. They can bind to a PKN protein. PKN is involved in the regulation of cell cycle, receptor trafficking, vesicle transport and apoptosis	The kinases Chk1/Chk2(Cds1) are used as a checkpoint during mitosis, induced when there is DNA damage
YWHAE, CEP70, PLK1, UBC, BTRC, OPTN, YWHAG, SKP1	YWHAE, SFN, RAC1, YWHAZ, YWHAG	YWHAE, SFN, YWHAZ, YWHAG

Activation of BAD and translocation to mitochondria	**CLEC7A (Dectin-1) signaling**	**Translocation of SLC2A4 (GLUT4) to the plasma membrane**
BAD has an important role in apoptosis. Calcineurin activates BAD by dephosphorylation. After activation, it is transported to the external membrane of the mitochondria, realeasing the cytochrome C that is a factor for apoptosis	CLEC7A (or Dectin-1) is a protein involved with the detection of bacteria and fungal cells. CLEC7A signaling induces the production of cytokines and interleukins	GLUT4 is a protein encoded by the gene SLC2A4 and is responsible for the uptake of glucose from the bloodstream, when the level of insulin increases. When the insulin binds to the receptors of the cell it starts the movement of the vesicles containing glucose towards the plasmamembrane
YWHAE, SFN, YWHAZ, YWHAG	NFKBIA, SYK, TRAF6, UBC, BTRC, SKP1	YWHAE, SFN, RAC1, CALM1, YWHAZ, YWHAG

Constitutive Signaling by Ligand-Responsive EGFR Cancer Variants	**Regulation of signaling by CBL**	**GPVI-mediated activation cascade**
After activation, the EGFR receptor starts several signaling cascades that initiates the transcription of gene involved in apoptosis and cellular proliferation. Over-expression of the wild-type EGFR or EGFR cancer mutants results in aberrant activation of these signaling cascades, providing an advantage to cancer cells	CBL negatively regulates signaling pathways by targeting proteins with ubiquitin for proteasomal degradation	GPVI is a receptor of collagen that initiates a signaling cascate that lead to platelet activation
CDC37, UBC, GRB2, PIK3R1	SYK, UBC, GRB2, PIK3R1	SYK, PTPN6, GRB2, RAC1, PIK3R1

7 Discussion

The main goal of this study was the discovery of new risk factors for a specific complex disease, in this case, T2D, using a case-control analysis. Even though it is known that complex diseases have genetic factors associated to them, as there is evidence of its heritability, the specific genetic factors are unknown. According to the Human Phenotype Ontology (HPO) website [67], at the moment there are 169 gene associations with T2D, but they explain a small percentage of the heritability of the disease, meaning that there is still little information on the disease and its mechanisms.

Can these models predict the disease risk?

From a computational perspective, this problem is defined as the discrimination between a healthy individual (control) and an individual with T2D (case) only using the information provided by their genotypes. Since the two initial datasets were not genotyped by the same methods nor sequenced by the same machines, it was necessary to prevent a prediction based in these differences. By using a set of variants that corresponded to a gene, and not independent variants, the weight of a noisy or incorrectly genotyped variant was reduced, avoiding this bias. Also, when using a classification algorithm we monitor the importance of the features, detecting genes or variants that overfit the model. Besides these mechanisms, by performing a quality control on the case dataset, it was assumed that most of the variants were correctly genotyped and that there was no bias between the case and control groups.

Considering the results obtained in the models' evaluation, it is possible to observe that most of the models can successfully predict the disease risk of the samples (Fig. 7). Comparing the results between the different models, the Decision Tree models produced higher values (≥ 0.87) which shows that it could better address the complexity of the data. The Logistic Regression models and SVM models provided lower results. Both used linear functions for the classification, which indicates that these functions have a higher difficulty in explaining the underlying structure of the data.

When using the known risk genes as input, the models had lower accuracy, F1-score and AUC values. This was expected because it is known that these gene associations do not account for a high percentage of the heritability. Comparing the results obtained for the significant genes and for the central genes, the first, produced for the most part, slightly better values. Since the significant genes dataset had features that were extracted directly from the most significant genes of the original dataset, it is not surprising that it produced good results. When just the top 25 features were used, which had less than 8% of the full dataset's size, the three metrics kept relatively good values. It was expected lower values in the results from the central genes dataset, given that the features were extracted from genes central to a network of significance and generally not significant themselves. Even though, the values from accuracy, F1-score and AUC were lower than the values from the significant genes, the results were still good. When using the top 25 features from the central genes, which had less than 9% of the full dataset's size, there was noticeable a rising of the values, being the dataset with the best overall results. With only the 25 features of this dataset it was possible to predict the risk of disease with a good degree of success.

Does the use of genes and their integration in PPI networks offer an advantage for this type of studies?

In 2017, Boyle, Li, and Pritchard [68] proposed the omnigenic model to explain the complex diseases. Their model states that a reduced number of genes or pathways directly affect the traits of the disease, having specific roles its etiology, introducing the term "core genes". Furthermore, they state that when these "core genes" are damaged or lose function they have large effect sizes on the disease risk, since they play more direct roles.

In this study it is shown that even thought most of the genes identified are not on the lists of gene associations by HPO or T2D Knowledge Portal, they are involved in biological processes and pathways of interest for this disease. When looking into the GO terms for biological processes it is noticeable that many of the genes are grouped by the same terms as the known risk genes (Fig. 8 and Table 12). Even so these genes share these terms, this does not mean that they share the same pathway or interactors, however, they can be involved in similar functions.

Looking at the pathways selected from the central genes (Table 10), one of the most interesting is the Translocation of SLC2A4 (GLUT4) to the plasmamembrane. Since T2D is characterized by an insulin resistance. This pathway is related with T2D [69] and the identification of the genes in this pathway may provide important information.

Is it still advantageous to look into the PPIs networks, if the allele information is lost in the process?

By grouping the variants into genes, this pipeline loss allele information that is important for more specific genetic studies. Although it is possible to retrace some of the information, it becomes complex to understand, for instance, which specific mutations have an effect on the disease. This difficulty further increase when looking to the central genes, since their selection is based not on their genetic information but in their connection to significant genes.

By looking at the results it is possible to see that using these networks could provide more biological information into the mechanisms of the disease. According to the omnigenic model, identifying these "core genes" would lead the investigation to genes that have a higher effect size on disease risk. However, since these genes are affected by other processes, they might not be directly associated with any mutation, and for this reason their usefulness in the prediction of the disease risk might be limited.

Can this pipeline identify genetic variants of interest?

The central question of this study is if this disease predictor can identify new genetic markers for a complex disease, in this case, T2D. Looking at the list of central genes identified from the PPI networks, it is possible to identify three genes in common with the known risk genes list: CAV1, PCBD1 and WFS1 (Table 7). Features from two of these three genes were included in the top 25 central genes dataset, the one with the best results from the models' evaluation. Although this number of genes is low, this indicates that this pipeline has correctly selected genes of interest. Also, from the functional analysis of the central gene was identified a pathway that is directly associated with T2D. All this shows that this pipeline can, at some degree, identify genes of interest.

8 Conclusion

In this study we developed a T2D risk predictor that identified genes of interest for the disease, using a case-control analysis of variants' genotypes.

During the pipeline's development we faced some challenges. Firstly, there are ethical constrains around clinical studies which creates difficulties and delays obtaining patient data. For this reason, and because of the estimated duration of this investigation, we used a dataset shared by UC-BIOTECH which had a small number of samples from Portuguese patients only. This not only reduced the ratio samples/variants, as restricted the study to a specific ancestry. Another difficulty in this work was the computational power needed to perform the analysis. Since the datasets used had a total size of about 600 Gb when uncompressed, processing these files required not only a higher computational power, but an optimization of the code that handle them.

Nevertheless, in this study were identified 82 significant genes and 77 central genes. After dimensionality reduction, 25 significant genes and 12 central genes were highlighted and from the functional analysis of the central genes were identified pathways of interest. The models for risk prediction had good results, being the set of the top 25 features from the central genes dataset that had the higher values. Both the significant genes and the central genes selected in this study shared several biological processes' terms with the known risk genes which shows that the overall functions of the genes may be similar. A known pathway of interest for the disease, Translocation of SLC2A4 (GLUT4) to the plasmamembrane, was identified from the list of central genes provided. Lastly, from the PPI network were identified three genes that are already known risk genes.

All these results demonstrated that important biological information could be revealed when using gene regions and even more by integrating them in PPIs networks. Even though much of the allele information is lost by using this approach, gaining insight into the biology of the disease can be essential to developed new treatments and strategies for diagnosis and prevention.

From this point on, it would be necessary to validate the risk predictor using bigger datasets, from other complex diseases or even from not only GWAS studies, but, for example, Whole-Genome Sequencing (WGS). Another way this pipeline could be improved is by using clinical data from the patients, which would improve the risk variants' selection by taking into account the environmental factors. Also, using not only genomics, but other omics' data, for instance, transcriptomics and epigenomics, would increase the risk's prediction power of this pipeline. The results obtained also highlighted central genes from interesting pathways, so further investigation into this genes and pathways could reveal more information into the biology of T2D.

Bibliography

[1] Bertrand Jordan. "Genes and Non-Mendelian Diseases: Dealing with Complexity". In: *Perspectives in Biology and Medicine* 57.1 (2014), pp. 118–131.

[2] Andrew P. Morris and Lon R. Cardon. "Genome‐Wide Association Studies". In: *Handbook of Statistical Genomics*. Ed. by David Balding, Ida Moltke, and John Marioni. 4th. Wiley, July 2019, pp. 597–550.

[3] Peter M. Visscher et al. "10 Years of GWAS Discovery: Biology, Function, and Translation". In: *American Journal of Human Genetics* 101.1 (July 2017), pp. 5–22.

[4] Daniel Sik Wai Ho et al. "Machine learning SNP based prediction for precision medicine". In: *Frontiers in Genetics* 10.MAR (Mar. 2019).

[5] Diogo M. Camacho et al. "Next-Generation Machine Learning for Biological Networks". In: *Cell* 173.7 (June 2018), pp. 1581–1592.

[6] Bruce Alberts et al. *Essential Cell Biology*. 4th. New York, 2013, p. 864.

[7] Eleftheria Zeggini and Andrew Morris. *Assessing rare variation in complex traits: Design and analysis of genetic studies*. Springer, 2015, pp. 1–261.

[8] Yehudit Hasin, Marcus Seldin, and Aldons Lusis. "Multi-omics approaches to disease". In: *Genome Biology* 18.1 (2017).

[9] Andrew R. Joyce and Bernhard Palsson. "The model organism as a system: Integrating 'omics' data sets". In: *Nature Reviews Molecular Cell Biology* 7.3 (2006), pp. 198–210.

[10] F. Sanger, S. Nicklen, and A. R. Coulson. "DNA sequencing with chain-terminating inhibitors." In: *Proceedings of the National Academy of Sciences of the United States of America* 74.12 (1977), pp. 5463–5467.

[11] Michael L. Metzker. "Sequencing technologies the next generation". In: *Nature Reviews Genetics* 11.1 (2010), pp. 31–46.

[12] James M. Heather and Benjamin Chain. "The sequence of sequencers: The history of sequencing DNA". In: *Genomics* 107.1 (2016), pp. 1–8.

[13] H. P.J. Buermans and J. T. den Dunnen. "Next generation sequencing technology: Advances and applications". In: *Biochimica et Biophysica Acta - Molecular Basis of Disease* 1842.10 (2014), pp. 1932–1941.

[14] International Diabetes Federation. *IDF Diabetes Atlas*. 9th. Brussels, Belgium: International Diabetes Federation, 2019, p. 176.

[15] Rashmi B. Prasad and Leif Groop. "Genetics of type 2 diabetes—pitfalls and possibilities". In: *Genes* 6.1 (2015), pp. 87–123.

[16] Sudesna Chatterjee, Kamlesh Khunti, and Melanie J. Davies. "Type 2 diabetes". In: *The Lancet* 389.10085 (2017), pp. 2239–2251.

[17] Dharambir K Sanghera and Piers R Blackett. "Type 2 Diabetes Genetics: Beyond GWAS". In: *Journal of Diabetes & Metabolism* 3.198 (2012), p. 2012.

[18] Gonneke Willemsen et al. "The Concordance and Heritability of Type 2 Diabetes in 34,166 Twin Pairs From International Twin Registers: The Discordant Twin (DISCOTWIN) Consortium". In: *Twin Research and Human Genetics* 18.6 (2015), pp. 762–771.

[19] Alena Stančáková and Markku Laakso. "Genetics of type 2 diabetes". In: *Endocrine Development*. Ed. by C Stettler, E Christ, and P Diem. Vol. 31. Karger Publishers, 2016, pp. 203–220.

[20] Teri A. Manolio et al. "Finding the missing heritability of complex diseases". In: *Nature* 461.7265 (2009), pp. 747–753.

[21] Mariette Awad and Rahul Khanna. *Efficient learning machines: Theories, concepts, and applications for engineers and system designers*. 1st. Apress, 2015, pp. 1–248.

[22] Andriy Burkov. *The Hundred-Page Machine Learning Book*. Andriy Burkov, 2019, p. 160.

[23] Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-supervised Learning*. London: The MIT press, 2006, p. 508.

[24] Richard S. Sutton and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd. London: The MIT Press, 2018, p. 526.

[25] Abhijit Dasgupta et al. "Brief review of regression-based and machine learning methods in genetic epidemiology: The Genetic Analysis Workshop 17 experience". In: *Genetic Epidemiology* 35.SUPPL. 1 (2011), S5–11.

[26] Andreas C Müller and Sarah Guido. *Introduction to Machine Learning with Python*. 2016, p. 400.

[27] Ian H Witten et al. *Data mining : practical machine learning tools and techniques*. 4th. Morgan Kaufmann, 2017, p. 654.

[28] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. New York: Cambridge University Press, 2014, p. 397.

[29] Gareth James et al. *An Introduction to Statistical Learning*. Vol. 112. New York: Springer, 2013.

[30] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning Book*. MIT Press, 2016.

[31] Laurens Van Der Maaten and Geoffrey Hinton. "Visualizing data using t-SNE". In: *Journal of Machine Learning Research* 9.86 (2008), pp. 2579–2605.

[32] Wentian Li et al. "Application of t-SNE to human genetic data". In: *Journal of Bioinformatics and Computational Biology* 15.4 (2017), p. 1750017.

[33] Nicholas J. Schork et al. "Common vs. rare allele hypotheses for complex diseases". In: *Current Opinion in Genetics and Development* 19.3 (2009), pp. 212–219.

[34] David E. Reich and Eric S. Lander. "On the allelic spectrum of human disease". In: *Trends in Genetics* 17.9 (2001), pp. 502–510.

[35] J. K. Pritchard. "Are rare variants responsible for susceptibility to complex diseases?" In: *American Journal of Human Genetics* 69.1 (2001), pp. 124–137.

[36] Peter M. Visscher et al. "Five Years of GWAS Discovery". In: *The American Journal of Human Genetics* 90 (2012), pp. 7–24.

[37] Montgomery Slatkin. "Linkage disequilibrium - Understanding the evolutionary past and mapping the medical future". In: *Nature Reviews Genetics* 9.6 (June 2008), pp. 477–485.

[38] Jeffrey D. Wall and Jonathan K. Pritchard. "Haplotype blocks and linkage disequilibrium in the human genome". In: *Nature Reviews Genetics* 4.8 (2003), pp. 587–597.

[39] Johanna K. Distefano and Darin M. Taverna. "Technological issues and experimental design of gene association studies". In: *Methods in Molecular Biology*. Vol. 700. Springer, 2011, pp. 3–16.

[40] William S. Bush and Jason H. Moore. "Chapter 11: Genome-Wide Association Studies". In: *PLoS Computational Biology* 8.12 (2012).

[41] Carl A. Anderson et al. "Data quality control in genetic case-control association studies". In: *Nature Protocols* 5.9 (2010), pp. 1564–1573.

[42] Michael E. Weale. "Quality control for genome-wide association studies". In: *Methods in Molecular Biology* 628 (2010), pp. 341–372.

[43] Sewall Wright. "Coefficients of Inbreeding and Relationship". In: *The American Naturalist* 56.645 (1922), pp. 330–338.

[44] Oliver Mayo. "A century of Hardy-Weinberg equilibrium". In: *Twin Research and Human Genetics* 11.3 (2008), pp. 249–256.

[45] David J. Balding. “A tutorial on statistical methods for population association studies”. In: *Nature Reviews Genetics* 7.10 (2006), pp. 781–791.

[46] Geraldine M. Clarke et al. “Basic statistical analysis in genetic case-control studies”. In: *Nature Protocols* 6.2 (2011), pp. 121–133.

[47] Karl Pearson. “ X. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling ”. In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 50.302 (1900), pp. 157–175.

[48] Matthew Stephens and David J. Balding. “Bayesian statistical methods for genetic association studies”. In: *Nature Reviews Genetics* 10.10 (2009), pp. 681–690.

[49] Andrew G. Clark. “The role of haplotypes in candidate gene studies”. In: *Genetic Epidemiology* 27.4 (2004), pp. 321–333.

[50] Jason H. Moore and Scott M. Williams. “Traversing the conceptual divide between biological and statistical epistasis: Systems biology and a more modern synthesis”. In: *BioEssays* 27.6 (2005), pp. 637–646.

[51] Marylyn D. Ritchie. “Using biological knowledge to uncover the mystery in the search for epistasis in genome-wide association studies”. In: *Annals of Human Genetics* 75.1 (2011), pp. 172–182.

[52] David M. Evans et al. “Two-stage two-locus models in genome-wide association”. In: *PLoS Genetics* 2.9 (2006), pp. 1424–1432.

[53] Casey S. Greene et al. “Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions”. In: *BioData Mining* 2.1 (2009).

[54] Sandrine Dudoit, Juliet Popper Shaffer, and Jennifer C. Boldrick. “Multiple hypothesis testing in microarray experiments”. In: *Statistical Science* 18.1 (2003), pp. 71–103.

[55] Andrew Collins and Yin Yao. “Machine Learning Approaches: Data Integration for Disease Prediction and Prognosis”. In: *Applied Computational Genomics*. Ed. by Yin Yao. Singapore: Springer, 2018, pp. 137–141.

[56] Maxwell W. Libbrecht and William Stafford Noble. “Machine learning applications in genetics and genomics”. In: *Nature Reviews Genetics* 16.6 (2015), pp. 321–332.

[57] Ching Lee Koo et al. “A review for detecting gene-gene interactions using machine learning methods in genetic epidemiology”. In: *BioMed Research International* 2013.432375 (2013).

[58] Jason H. Moore, Folkert W. Asselbergs, and Scott M. Williams. “Bioinformatics challenges for genome-wide association studies”. In: *Bioinformatics* 26.4 (2010), pp. 445–455.

[59] Ray L. Somorjai, B. Dolenko, and R. Baumgartner. “Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: Curses, caveats, cautions”. In: *Bioinformatics* 19.12 (2003), pp. 1484–1491.

[60] Adam Auton et al. “A global reference for human genetic variation”. In: *Nature* 526.7571 (2015), pp. 68–74.

[61] Coriell Institute for Medical Research. *Iberian populations in Spain [IBS]*. url: https://www.coriell.org/1/NHGRI/Collections/1000-Genomes-Collections/Iberian-populations-in-Spain-IBS (visited on 11/09/2020).

[62] Ralph D’Agostino and E. S. Pearson. “Tests for departure from normality. Empirical results for the distributions of B and $\sqrt{b}$”. In: *Biometrika* 60.3 (1973), pp. 613–622.

[63] Type 2 Diabetes Knowledge Portal. *Curated T2D effector gene predictions*. url: http://t2d.hugeamp.org/method.html?trait=t2d%7B%5C&%7Ddataset=mccarthy (visited on 09/22/2020).

[64] Rose Oughtred et al. “The BioGRID interaction database: 2019 update”. In: *Nucleic Acids Research* 47.D1 (2019), pp. D529–D541.

[65] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources”. In: *Nature Protocols* 4.1 (2009), pp. 44–57.

[66] Da Wei Huang, Brad T. Sherman, and Richard A. Lempicki. “Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists”. In: *Nucleic Acids Research* 37.1 (2009), pp. 1–13.

[67] Sebastian Köhler et al. *Human Phenotype Ontology*. 2019. url: https://hpo.jax.org/app/browse/term/HP:0005978 (visited on 01/07/2021).

[68] Evan A. Boyle, Yang I. Li, and Jonathan K. Pritchard. “An Expanded View of Complex Traits: From Polygenic to Omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186.

[69] Michael Gaster et al. “GLUT4 is reduced in slow muscle fibers of type 2 diabetic patients: Is insulin resistance in type 2 diabetes a slow, type 1 fiber disease?” In: *Diabetes* 50.6 (2001), pp. 1324–1329.

A Support material

A.1 Initial Datasets: Quality Control Results

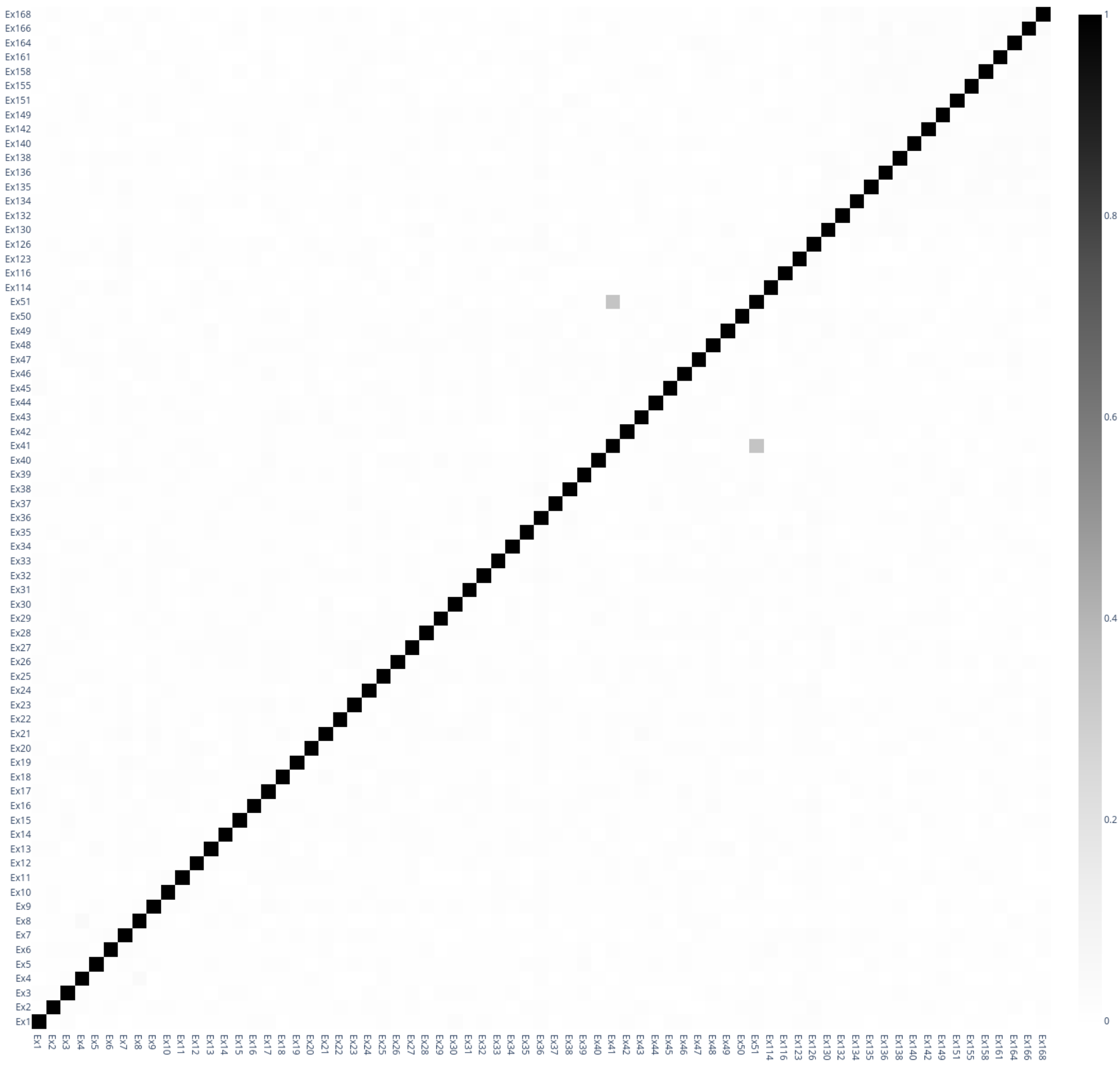


Figure 9.: Heatmap that represents the relatedness (Ajk) between all individuals of case's dataset, where $Ajk = 1$ means duplicates and $Ajk = 0$ means not related. The pair Ex41-Ex51 had $Ajk = 0.342$ and those individuals were considered related.

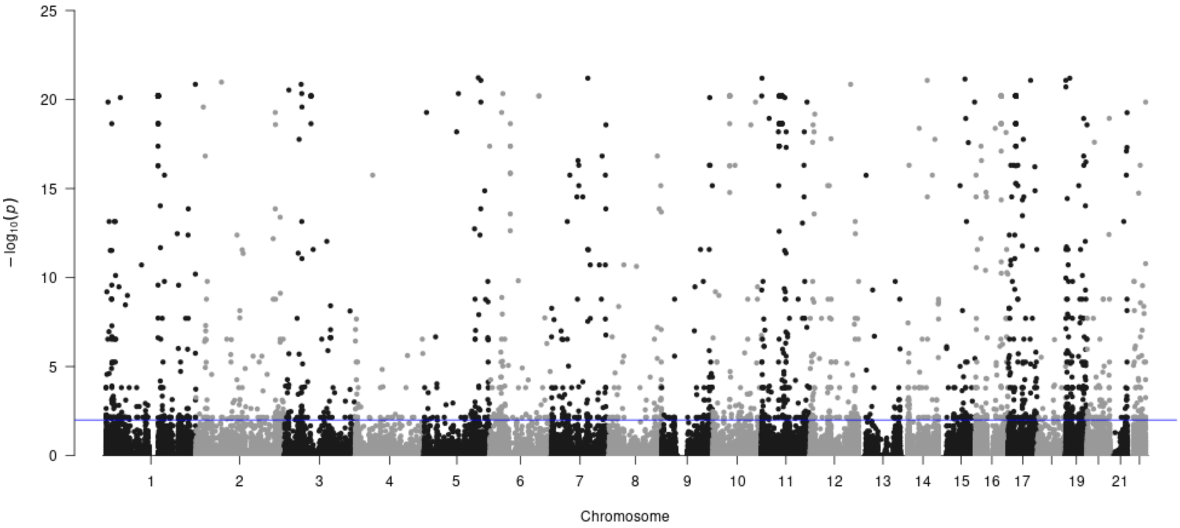


Figure 10.: Manhattan plot of the results from the Hardy-Weinberg Equilibrium according to the position of each variant. 9 792 variants of 250 869 did not follow the Hardy-Weinberg Equilibrium. The threshold $-\log_{10}(0.01) = 2$ is represented by a line.

Table 11.: Summary of the phred quality control, missingness and allele frequency statistics for the case dataset, calculated during quality control.

Statistic	**Minimum**	**1st Quadrant**	**Median**	**Mean**	**3rd Quadrant**	**Maximum**
Quality score	10.0	99.6	244.1	333.8	449.1	4003.3
Missingness	0.00000	0.00000	0.00000	0.03835	0.02899	0.97222
Allele Frequency	0.000000	0.007246	0.021127	0.093787	0.129630	0.500000

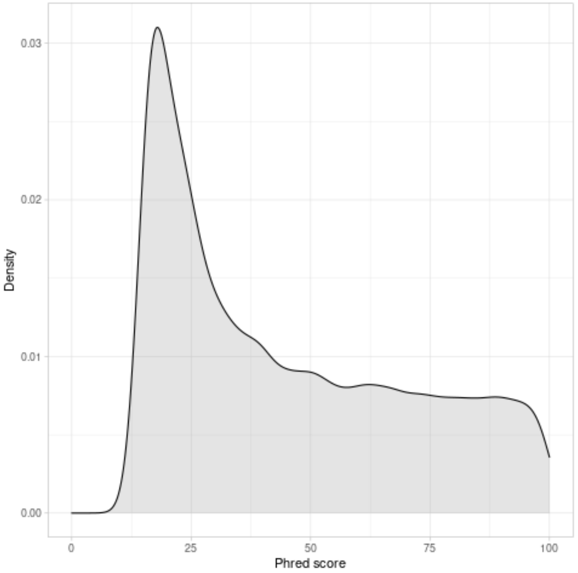


Figure 11.: Phred quality score of the variants in the case dataset, using a window from 0 to 100.

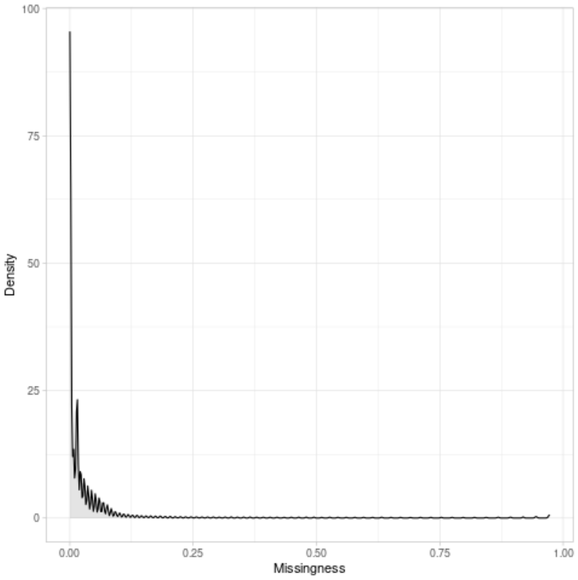


Figure 12.: Percentage of missing genotypes in the case dataset. Most variants have a lower number of missing genotypes.

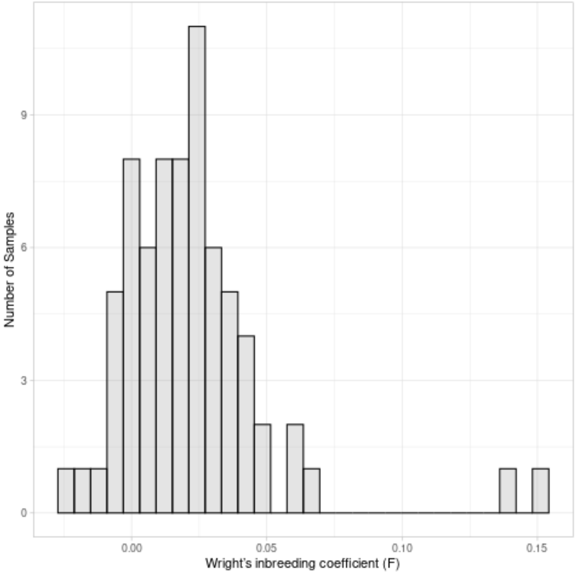


Figure 13.: Wright's inbreeding coefficient for sample in the case dataset.

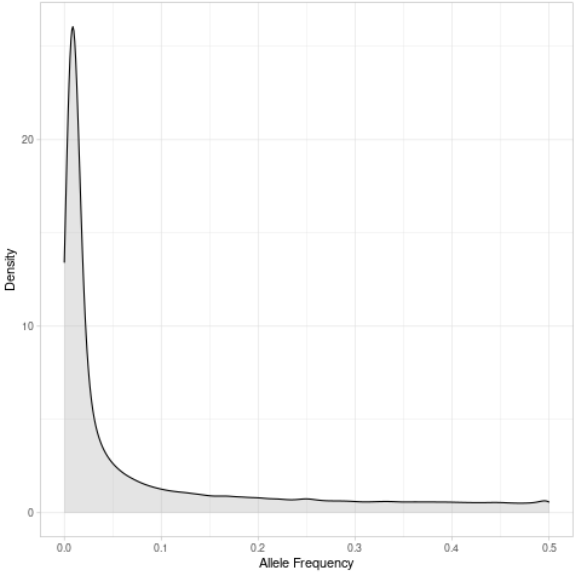
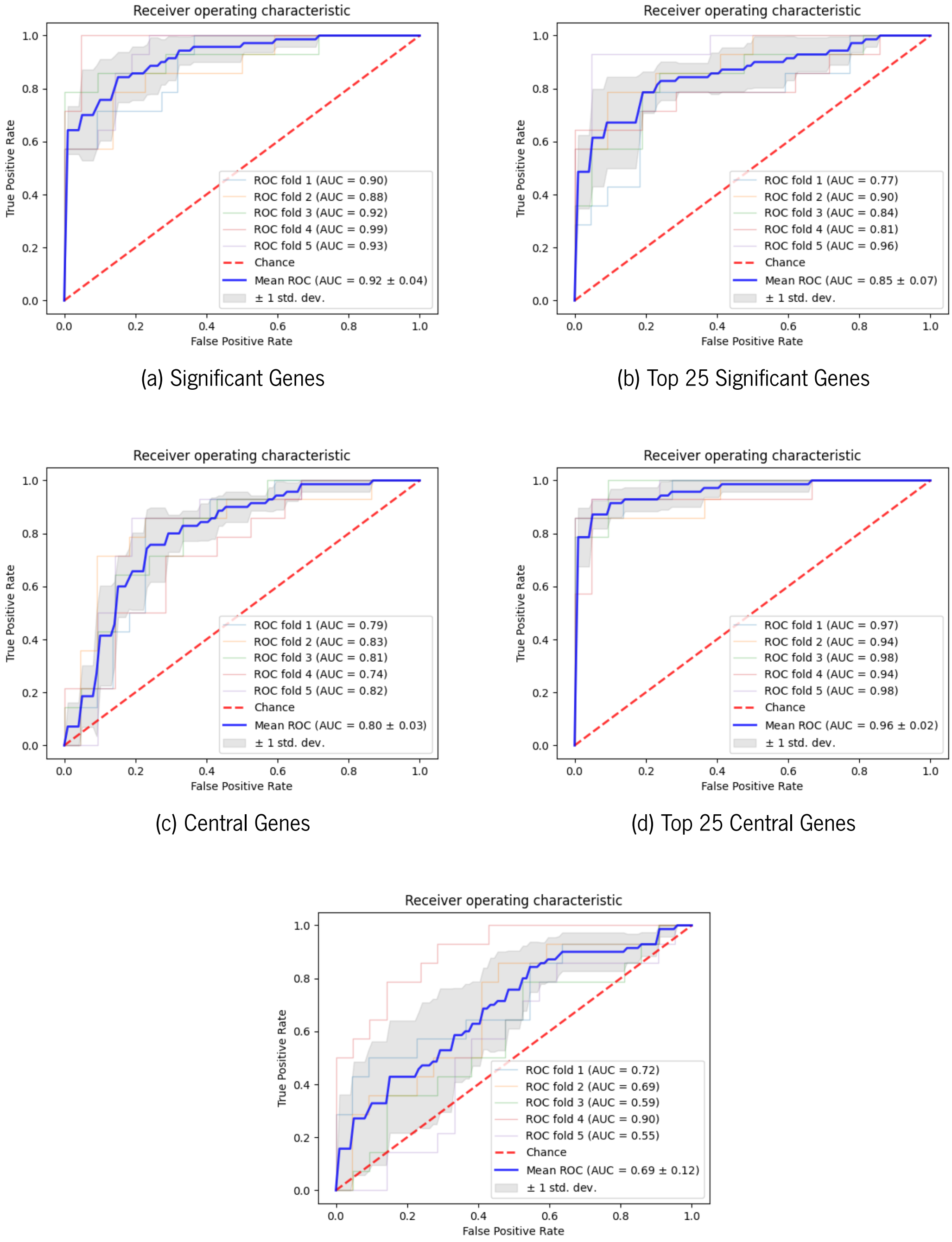


Figure 14.: Allele Frequency of variants in the case dataset.

A.2 Model Evaluation: ROC Curves Results



(a) Significant Genes

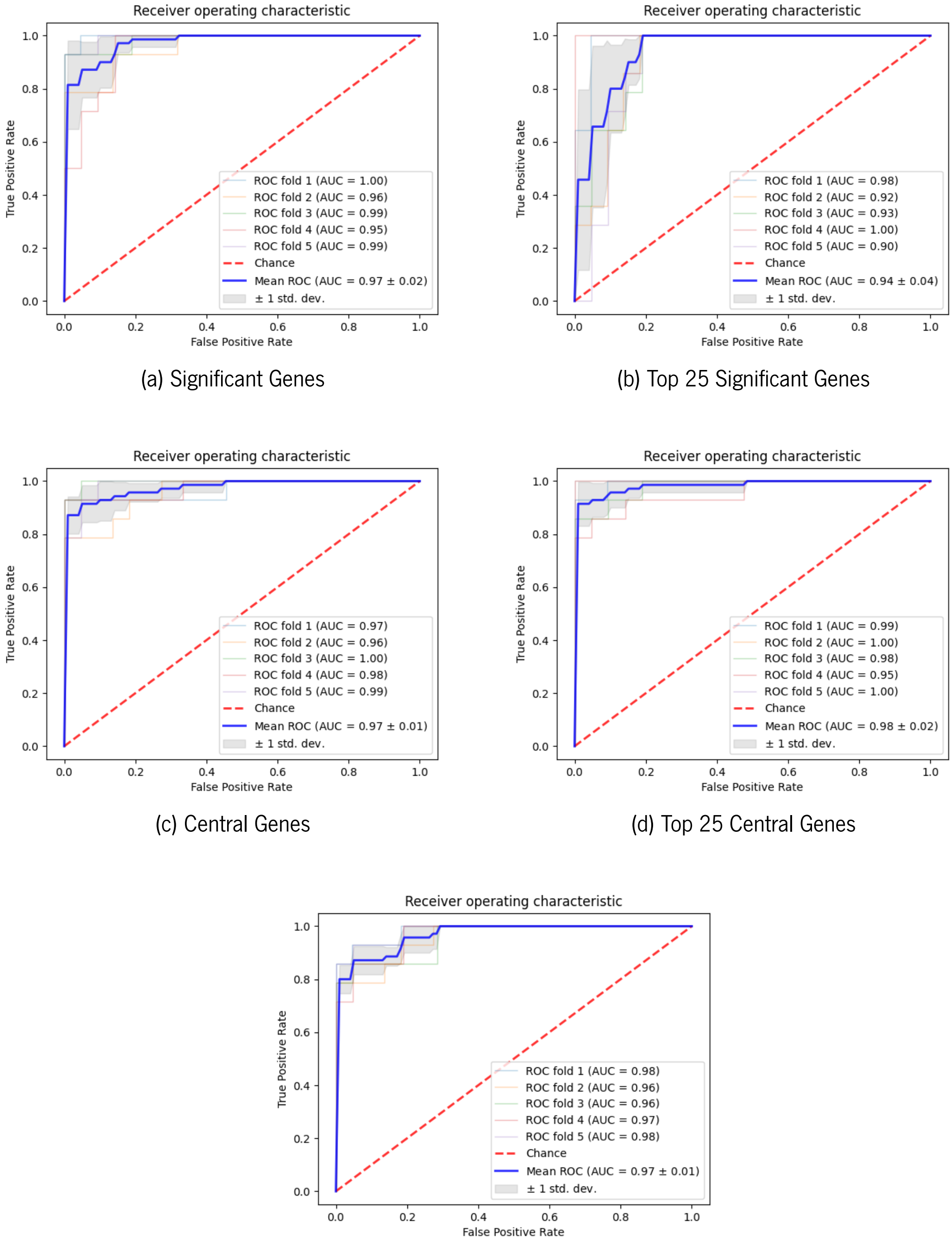
(b) Top 25 Significant Genes

(c) Central Genes

(d) Top 25 Central Genes

(e) Known Risk Genes

Figure 15.: ROC Curves results of each dataset (a) to (e) for SVM model.



(a) Significant Genes

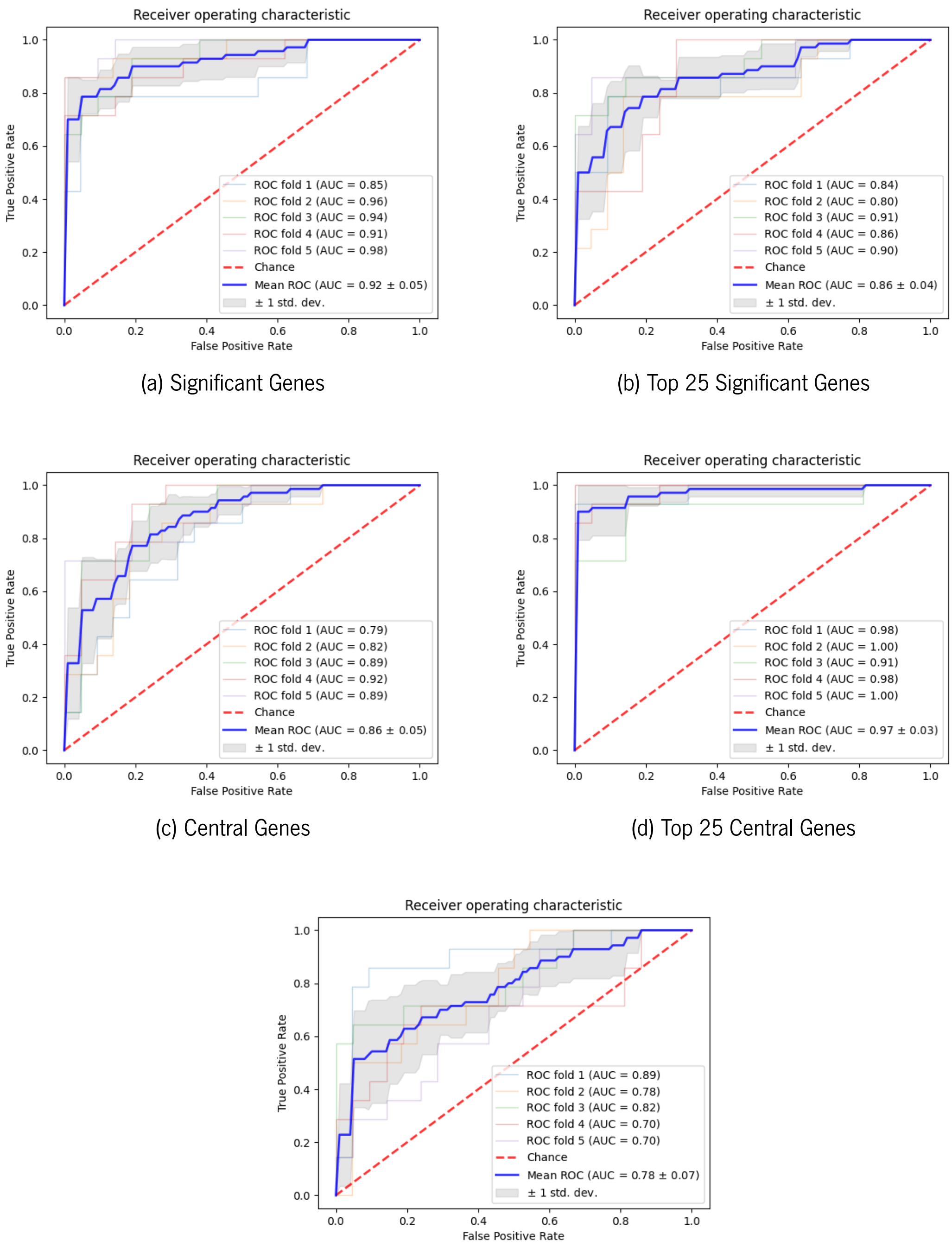
(b) Top 25 Significant Genes

(c) Central Genes

(d) Top 25 Central Genes

(e) Known Risk Genes

Figure 16.: ROC Curves results of each dataset (a) to (e) for Decision Tree model.



(a) Significant Genes

(b) Top 25 Significant Genes

(c) Central Genes

(d) Top 25 Central Genes

(e) Known Risk Genes

Figure 17.: ROC Curves results of each dataset (a) to (e) for Logistic Regression model.

A.3 Functional Annotation: DAVID Results

Table 12.: Number of genes from the known risk genes, significant genes and central genes associated with all biological processes found.

Name	Known Risk Genes	Central Genes	Significant Genes
Cellular macromolecule biosynthetic process	38	38	28
Regulation of macromolecule biosynthetic process	35	38	26
Regulation of cellular macromolecule biosynthetic process	35	37	26
Regulation of gene expression	33	38	26
Nucleic acid metabolic process	33	38	26
Nucleobase-containing compound biosynthetic process	35	34	26
RNA metabolic process	32	32	26
Regulation of transcription, DNA-templated	30	31	26
Regulation of RNA metabolic process	30	31	26
Regulation of RNA biosynthetic process	30	31	26
Transcription, DNA-templated	30	31	25
Positive regulation of macromolecule metabolic process	28	44	
Positive regulation of cellular biosynthetic process	29	28	12
Animal organ development	27	39	
Positive regulation of nucleobase-containing compound metabolic process	26	26	12
Regulation of signal transduction	19	45	
Positive regulation of macromolecule biosynthetic process	24	25	12
Cell surface receptor signaling pathway	24	36	
Apoptotic process	20	37	
Positive regulation of gene expression	22	22	12
Positive regulation of RNA metabolic process	21	20	12
Positive regulation of RNA biosynthetic process	21	19	12
Negative regulation of macromolecule metabolic process	20	32	
Positive regulation of transcription, DNA-templated	21	19	11
Regulation of phosphate metabolic process	17	34	
Phosphorylation	14	36	
Protein transport	27	22	
Regulation of programmed cell death	17	31	
Regulation of apoptotic process	17	30	
Positive regulation of cell communication	13	32	
Regulation of establishment of protein localization	20	20	
Cell development	15	25	
Positive regulation of transport	17	21	
Negative regulation of cell communication	14	24	
Negative regulation of cellular biosynthetic process	19	18	
Nervous system development	16	21	
Negative regulation of programmed cell death	14	23	
Negative regulation of cell death	14	23	
Negative regulation of macromolecule biosynthetic process	18	18	
Negative regulation of cellular macromolecule biosynthetic process	18	18	
Negative regulation of apoptotic process	14	22	
Regulation of protein transport	19	16	
Negative regulation of gene expression	17	18	
Negative regulation of signal transduction	12	23	
Positive regulation of phosphorus metabolic process	12	23	

Continued on next page

Table 12 continued from previous page

Name	Known Risk Genes	Central Genes	Significant Genes
Positive regulation of phosphate metabolic process	12	23	
Secretion	23	11	
Negative regulation of nucleobase-containing compound metabolic process	16	18	
Negative regulation of transcription, DNA-templated	16	17	
Negative regulation of RNA metabolic process	16	17	
Negative regulation of RNA biosynthetic process	16	17	
Regulation of kinase activity	7	24	
Secretion by cell	21	9	
Enzyme linked receptor protein signaling pathway	12	18	
Ion transport	17	12	
Cellular chemical homeostasis	15	12	
Positive regulation of cell death	8	19	
Protein secretion	20	6	
Cell migration	9	17	
Organ morphogenesis	13	12	
Gland development	13	12	
Positive regulation of establishment of protein localization	11	14	
Positive regulation of programmed cell death	6	19	
Regulation of protein secretion	18	6	
Central nervous system development	12	12	
Apoptotic signaling pathway	7	17	
Positive regulation of apoptotic process	6	18	
Cellular response to organonitrogen compound	12	11	
Cardiovascular system development	11	12	
Circulatory system development	11	12	
Epithelial cell differentiation	14	8	
Response to peptide	14	8	
Cellular response to hormone stimulus	12	10	
Positive regulation of protein transport	11	10	
Negative regulation of phosphorus metabolic process	8	13	
Negative regulation of phosphate metabolic process	8	13	
Response to peptide hormone	13	7	
Regulation of cell cycle process	6	14	
Cellular response to growth factor stimulus		14	6
Negative regulation of intracellular signal transduction	6	13	
Regulation of apoptotic signaling pathway	6	13	
Positive regulation of proteolysis	5	14	
Cellular response to peptide	11	7	
Positive regulation of cell development	8	10	
Cellular response to peptide hormone stimulus	10	7	
Embryonic morphogenesis	9	8	
Brain development	8	9	
Regulation of nervous system development	8	9	
Cellular ion homeostasis	6	11	
Vasculature development	8	8	
Regulation of ion transport	8	8	
Regulation of neurogenesis	7	9	
Embryo development ending in birth or egg hatching	6	10	
Response to steroid hormone	6	10	
Negative regulation of transferase activity	6	10	

Continued on next page

Table 12 continued from previous page

Name	Known Risk Genes	Central Genes	Significant Genes
Positive regulation of secretion	10	5	
Positive regulation of secretion by cell	10	5	
Response to insulin	9	6	
Negative regulation of transport	8	7	
Positive regulation of cellular catabolic process	5	10	
Positive regulation of protein secretion	10	4	
Regulation of WNT signaling pathway	6	8	
Regulation of neuron differentiation	6	8	
Embryonic organ development	6	8	
Protein import	5	9	
Hexose metabolic process	9	4	
Reproductive structure development	6	7	
Positive regulation of neurogenesis	6	7	
Positive regulation of nervous system development	6	7	
Reproductive system development	6	7	
Protein import into nucleus	4	9	
Protein targeting to nucleus	4	9	
Single-organism nuclear import	4	9	
Regulation of nucleotide metabolic process	8	4	
Carbohydrate transport	8	4	
Monosaccharide transport	8	4	
Angiogenesis	6	6	
Regulation of vesicle-mediated transport	6	6	
Regulation of metal ion transport	5	7	
Intrinsic apoptotic signaling pathway	5	7	
Cell cycle arrest		8	4
Regulation of glucose transport	7	4	
Positive regulation of nucleotide metabolic process	7	4	
Digestive system development	7	4	
Second-messenger-mediated signaling	6	5	
Positive regulation of homeostatic process	6	5	
Response to hypoxia	5	6	
Regulation of transmembrane receptor protein serine/threonine kinase signaling pathway		7	4
Energy reserve metabolic process	6	4	
Regulation of nucleotide biosynthetic process	6	4	
Positive regulation of nucleotide biosynthetic process	6	4	
Digestive tract development	6	4	
Urogenital system development	5	5	
Regulation of myeloid cell differentiation		6	4
Organic hydroxy compound transport	5	4	
Mammary gland development	5	4	
Positive regulation of ion transport	5	4	
Regulation of glucose import	5	4	
Epithelial cell development	4	5	
Regulation of cell cycle arrest	4	5	
Negative regulation of apoptotic signaling pathway	4	5	
Glycogen metabolic process	5	3	
Cellular glucan metabolic process	5	3	
Glucan metabolic process	5	3	

Continued on next page

Table 12 continued from previous page

Name	Known Risk Genes	Central Genes	Significant Genes
Cellular polysaccharide metabolic process	5	3	
Liver development	4	4	
Glial cell differentiation	4		4
Response to starvation	4	4	
Regulation of angiogenesis	4	4	
Hepaticobiliary system development	4	4	
Regulation of wound healing	3	5	
Negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway		5	3
Neuron fate commitment	4		3
Response to estradiol	3	4	
Oligodendrocyte differentiation	3		4
Positive regulation of calcium ion transport	3	4	
Negative regulation of hemopoiesis		4	3
Protein import into nucleus, translocation	3	3	
Smoothened signaling pathway	3	3	
Positive regulation of DNA binding	3	3	