SAERMA: Stacked Autoencoders Rule Mining Algorithm for the Interpretation of Epistatic Interactions in GWAS of Extreme Obesity

By

Casimiro Adays Curbelo Montañez, B.Eng., MSc

A thesis submitted in partial fulfilment of the requirements of Liverpool John Moores University for the degree of Doctor of Philosophy

May 2019

DECLARATION

I hereby declare that this work is a product of my own and all else is appropriately referenced. Work derived through collaboration and assistance has been acknowledged in the text, while a list of references is given in the bibliography.

ACKNOWLEDGEMENT

I would first like to express my sincere gratitude to my director of studies Dr Paul Fergus for all his valuable advice and guidance through the course of this PhD. This accomplishment would not have been possible without him. In addition, I am also grateful to Dr Carl Chalmers for supporting me with the experiments conducted and for his counsels during the delicate times of my PhD journey. Your life can always change for the better when you have the right people around. In this sense, I can say that I have been very lucky.

A very special mention also goes to my PhD colleagues Basma, Jade and Ross who also shared this enriching experience in LJMU with me. Thank you for all your support and the knowledge exchanged.

I would also like to acknowledge the Faculty Research Administration and the IT School Resources teams from the Faculty of Engineering and Technology for their aid and support when it was needed.

Last but not least, I must express my very profound gratitude to my parents Casimiro and Remedios, my sisters Almudena and María, and of course, to my partner Helen for providing me with unconditional support and incessant encouragement throughout my years of study and during the process of completing and writing this thesis. Love and support from my family both in Spain and in the UK always fed my soul in difficult times. For this reason, this thesis is dedicated to them.

"Ningún mar en calma hizo experto a un marinero"

- Credit unknown-

ABSTRACT

One of the most important challenges in the analysis of high-throughput genetic data is the development of efficient computational and statistical methods to identify statistically significant single nucleotide polymorphisms (SNPs). Genome-wide association studies (GWAS), are the state-of-the-art in identifying genetic variants for complex disorders, such as obesity. However, GWAS use single-locus analysis where each SNP is independently tested for association with some phenotype. The limitation of genetic variants identified by GWAS is its inability to explain the underlying genetic variation in complex diseases. Consequently, alternative approaches are required that are capable of modelling the intricate relationships between SNPs and phenotypes.

The approach presented in this thesis extends GWAS and explores the use of deep learning stacked autoencoders (SAE) and association rule mining (ARM) to identify epistatic interactions between SNPs. This is achieved using a case-control dataset containing 2,193 observations (962 cases and 1,231 controls) each with 594,034 genetic markers.

A statistical filtering strategy is adopted to reduce the large number of SNPs to a more manageable set suitable for machine learning tasks. Several experiments have been conducted to explore epistasis among the filtered subset (2,465 SNPs) and are compared with results obtained using the industry standard logistic regression via a Generalised Linear Model (GLM). These include a multi-layer feedforward artificial neural network (MLP), SAE and a combination approach using SAE and ARM. Functional enrichment analysis is

adopted to biologically validate association rules mined by the proposed method.

Baseline classification results are initially conducted using standard logistic regression (GLM) with SNPs input derived from several P-value thresholds ($1x10^{-5}$, $1x10^{-4}$, $1x10^{-5}$, and $1x10^{-2}$). The second experiment is carried out using an MLP trained using the same input features. In subsequent experiments, epistasis is investigated using SAE to extract nonlinear SNP-SNP interactions and pre-train a fully connected MLP layer. Features are extracted using four single layer autoencoders (AEs) stacked (containing 2,000-1,000-500-50 hidden units respectively). The initial results show that it is possible to gain an AUC = 85% (SE = 78% and SP = 80%) using 50 hidden neurons. The findings are encouraging; however, it is not possible to identify which information from the 2,465 SNPs is retained in the final AE layer (50 nodes) to initialise and train the MLP. Consequently, ARM is introduced to extend the SAE approach to provide interpretability regarding what SNPs more closely influence the phenotype and the interactions that exist between them. Interestingness measures, support, confidence, lift and chi-square test (χ^2) are utilised to rank and determine correlated rules, under a support-dependence framework. The SNPs from the top rules (top 300, 200, 100 and 50 rules) are used with a SAE and fully connected MLP to measure their discriminant capacity in distinguishing between case-control observations. Graph-based visualization methods are utilised to show the interactions between SNPs as identified by the top rules. While classifier performance metrics are utilised to assess classifier performance.

The SNPs contained in the set of top rules are used as input to different SAEs configurations to compress the features (retain only the salient information) through progressively smaller hidden layers. The final hidden layer is then used to initialise the learnable parameters of a fully connected MLP before it is fine-tuned for classification tasks. The results show that it is possible to achieve an AUC = 77%, SE = 77% and SP = 68%. More importantly, in parallel, it is possible to explore which of the 2,465 SNPs and their epistatic interactions are most strongly associated with obesity. This provides a significant novel contribution to the field of computational biology and is the first study of its kind to combine deep learning epistatic analysis using SAEs and ARM to classify case-control observations and provide an interpretation of the final trained classification model. The level of accuracy required is fully tuneable, i.e. it is possible to increase/decrease the results obtained by the SNPs in classification tasks by increasing/decreasing the rule mining support and confidence parameters, defined in the rule generation stage.

Additional experiments were conducted as a proof of concept to support the use of a statistical filtering approach to reduce the dimensionality of the data before investigating epistasis. Gene set enrichment analysis was utilised via the *i*-GSEA4GWAS web tool. Enriched gene sets were then used as input features for classification experiments using an MLP and their performance reported. Although this approach is based on biological knowledge, that is, genetic variants are filtered based on biological pathways, classification results did not outperform those achieved by SAERMA. This, thus, justifies the use of statistical filtering within the proposed algorithm.

I, therefore, claim the approach posited in this thesis is foundational in character and is the first study of its kind that combines GWAS quality control and logistic regression with association rule mining and deep learning stacked autoencoders to study epistatic interactions between SNPs in polygenic obesity GWAS.

LIST OF PUBLICATIONS

1. P. Fergus, **C. A. Curbelo Montañez**, B. Abdulaimma, P. Lisboa, C. Chalmers, and B. Pineles, “Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, Jan. 2018.

2. **C. A. Curbelo Montañez**, P. Fergus, C. Chalmers, and J. Hind, “Analysis of Extremely Obese Individuals Using Deep Learning Stacked Autoencoders and Genome-Wide Genetic Data,” in *15th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics* (CIBB), Caparica, Portugal, pp. 1–13, Sep. 2018.

3. **C. A. Curbelo Montañez**, P. Fergus, A. C. Montañez, A. Hussain, D. Al-Jumeily, and C. Chalmers, “Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs,” in *2018 International Joint Conference on Neural Networks (IJCNN)*, 2018, p. 8.

4. **C. A. Curbelo Montañez**, P. Fergus, A. Hussain, D. Al-Jumeily, M. T. Dorak, and R. Abdullah, “Evaluation of Phenotype Classification Methods for Obesity Using Direct to Consumer Genetic Data,” in *Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017*, D.-S. Huang, K.-H. Jo, and J. C. Figueroa-García, Eds. Liverpool: Springer International Publishing, 2017, pp. 350–362.

5. **C. A. Curbelo Montañez** *et al.*, “Machine learning approaches for the prediction of obesity using publicly available genetic profiles,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, pp. 2743–2750.

6. J. Hind, A. Hussain, D. Al-jumeily, B. Abdulaimma, **C. A. Curbelo Montanez**, and P. Lisboa, “A robust method for the interpretation of genomic data,” in *2017 International Joint Conference on Neural Networks (IJCNN)*, 2017, vol. 2017–May, pp. 3385–3390.

7. **C. A. Curbelo Montañez**, P. Fergus, A. Hussain, D. Al-Jumeily, B. Abdulaimma, and H. Al-Askar, “A Genetic Analytics Approach for Risk Variant Identification to Support Intervention Strategies for People Susceptible to Polygenic Obesity and Overweight,” in *Intelligent Computing Theories and Application: 12th International Conference, ICIC 2016, Lanzhou, China, August 2-5, 2016, Proceedings, Part I*, D.-S. Huang, V. Bevilacqua, and P. Premaratne, Eds. Cham: Springer International Publishing, 2016, pp. 808–819.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1. EXTENDED ABSTRACT AND SCOPE OF THE THESIS

1.1 Preamble

According to the World Health Organization (WHO), the occurrence of obesity and overweight worldwide doubled between 1980 and 2014 (World Health Organization 2014). In 2016 more than 1.9 billion adults were overweight and 650 million were obese (World Health Organization 2018). The condition was initially recognized as a disease in 1948 by the WHO (James 2008) and since then its prevalence has continued to increase making it a global phenomenon and one of the main contributors to poor health. It is considered one of the most difficult clinical and public health challenges worldwide (Yang et al. 2007; Hruby & Hu 2015). Obesity is a major risk and the leading cause for many other diseases such as Type 2 Diabetes (T2D), cardiovascular disease, premature death, hypertension, osteoarthritis, stroke and certain cancers (Walley et al. 2006; Bhaskaran et al. 2014; Yang et al. 2007). Consequently, it is high on the political agenda of many countries (Vallgårda et al. 2015).

1.2 Polygenic Obesity

The predisposition to obesity in humans is referred to as polygenic obesity and is considered a complex and multifactorial disease caused by interactions between genetic, behavioural and environmental factors. While obesity tends to exist within families, the way it is inherited does not correspond to known patterns. Numerous studies have shown that an individual's predisposition to

obesity is more similar among genetically related individuals than those that are not. The phenotypes associated with obesity exhibit significant additive heritability (h^2, the proportion of the variability of a trait that is attributable to genetic factors) (Min et al. 2013). In the case of Body Mass Index (BMI), family and twin studies have shown that between 40 and 70 percent of the inter-individual variation in obesity can be attributed to genetic differences in the population (Wardle et al. 2008; Zaitlen et al. 2013; El-Sayed Moustafa & Froguel 2013). The remaining percentage is associated with other factors, such as lifestyle and environmental.

1.3 Genome Wide Association Studies

Understanding polygenic obesity is complex and it does not exhibit a typical Mendelian pattern of transmission. There is evidence derived from Genome-wide Association Studies (GWAS) that suggest single nucleotide polymorphisms (SNPs) in certain genes are associated with obesity risk factors and BMI. Examples of these SNPs include those associated with fat mass and obesity (FTO) and the melanocortin 4 receptor genes (MC4R) (Wang et al. 2011; Xi et al. 2011; Corella et al. 2012; Loos 2012). Additional studies have reported that certain genes, including those mentioned, have a strong link with energy consumption in the nervous system when the hypothalamus part of the brain is stimulated (Willer et al. 2009).

Although GWAS have successfully discovered numerous genetic loci affecting complex diseases, the molecular pathways connecting genetic variants to complex traits remain poorly understood. This is mainly due to a large proportion of disease-associated signals (SNPs) being in non-coding regions of

the genome, which introduce the necessity of additional means for interpreting and validate GWAS results (Anon 2010). Performing gene set enrichment analysis (GSEA), pathways analysis or incorporating biological information such as expression quantitative trait loci (eQTL) have contributed to provide functional interpretation of many trait-associated SNPs in a biological context (discussed later in this thesis). This has opened opportunities for characterising functional sequence variation while improving understanding of basic processes of gene regulation and interpretation of GWAS. Therefore, an essential task to systematically disentangle the molecular mechanisms underlying complex diseases, is via the identification of complex interplays among multiple genes in a genome-wide context, using functional genomics (Mattson & Liang 2017).

Additionally, it has been suggested that common forms of diseases are not the result of single gene changes affecting a single outcome. Instead, complex diseases are most likely the result of complex relationships between gene networks which are modelled by complex genetic and environmental interactions. Hence, identifying interactions between two or more genes affecting disease susceptibility, namely epistasis, will help to provide a better understanding of diseases such as common obesity (De et al. 2015). This is regarded as a much more intuitive approach given that complex diseases cannot be reduced to single univariate SNP-phenotype interactions. Epistasis can be conducted from a biological and statistical viewpoint (Jiang et al. 2011). While biological epistasis investigates physical interactions occurring at molecular level, statistical epistasis represents the effect of the interactions between multiple genetic variants on the phenotype that cannot be estimated by individual loci exhibiting weak marginal effects. Both aspects of epistasis need

to be considered in order to provide a complete evaluation of the results. Once statistical epistasis has been identified, the biological implication of the interactions can be investigated using, for example, pathway analysis.

Advances in Human Genomics have provided significant opportunities in genetic studies and research has suggested that it might be possible to quantify an individual's susceptibility to obesity from an early age and thus, manage risks as individuals progress through life (Loos 2012). Given that sequencing the human genome is possible, and new genotyping and sequencing technologies are available, it is possible to analyse whole genetic sequences and detect diseases and associated traits (Pirmohamed 2011). Initiatives such as the 100,000 Genomes Project[1], conducted by Genomics England, aim to sequence 100,000 genomes from 70,000 NHS patients to provide treatments for those with rare diseases and cancer (Griffin et al. 2017). The information will be used to create a genomic medicine service for the NHS to enable new scientific discovery and provide medical insights. Therefore, combining personalised medicine with genetic information and integrating it into medical care and person specific risk assessments will help us to mitigate the long-term effects of obesity and its associated co-morbidities (Mardis 2008). This is being made possible through advances in bioinformatics (Sung 2012; Samish et al. 2014), data science (Higdon et al. 2013; Rudin et al. 2014) and advanced machine learning algorithms (Deo 2015).

[1] http://www.genomicsengland.co.uk/

1.4 Computational Biology

The availability of advanced analytical tools and a deeper understanding of the biology of genomes are necessary if we are to decipher, interpret, and optimize the clinical utility of variation in the human genome. To successfully identify genetic features for disease and health prediction using a genome-wide approach, several significant challenges must be addressed (Moore & White 2007). Traditional statistical methods such as logistic regression have shown limited power in modelling high-order nonlinear interactions between genetic variants (Gilbert-Diamond & Moore 2011). Hence, better data mining and machine learning approaches are required to statistically model relationships between DNA sequence variations and disease predisposition. Additionally, the high dimensionality present in genomic data makes it computationally difficult to exhaustively evaluate all SNP combinations. This is indeed, a well-known computational challenge referred to as the "curse of dimensionality" in the field of computer science (Altman & Krzywinski 2018). Therefore, filtering genetic variants or features plays an important role in genomic studies.

Performing SNP selection based on arbitrary significance threshold (i.e. some predefined P-value) can help to reduce computational complexity by calculating a test statistic for each marker separately and evaluating all possible interactions in the filtered subset (Hoh et al. 2000; Marchini et al. 2005). In this approach, the data is processed statistically to assess the quality or relevance of each SNP with an associated phenotype, which can then be evaluated using classification techniques. Although not ideal, this approach allows reducing

genome-wide datasets to a more manageable dimension, which eases the study of epistasis without any prior knowledge about the disease under investigation.

Finally, it is important to interpret gene-to-gene interactions in the context of human biology before any results can be translated into specific recommendations and treatment strategies. However, making etiological inferences from computational models has been considered the most relevant but difficult challenge (Moore & Williams 2005). In this thesis, results from the proposed method (SAERMA) are validated via functional analysis. This allows, for example, to prove that the interactions identified by association rules represent true epistasis in case items in the rules are mapped to a biological pathway.

1.5 Scope of Research

The research question is to investigate whether complex interactions between SNPs (epistasis) can explain obesity predisposition in humans. Following traditional GWAS quality control and association analysis, the most significant SNPs are selected and used in subsequent analysis to investigate epistasis. Stacked autoencoders are implemented as a feature extraction technique to capture epistatic interactions between SNPs based on variants identified through association rule mining (ARM). When these two techniques are combined in this configuration, it is possible to control the classification results produced in the final fully connected MLP layer of the stacked autoencoder by manipulating the interestingness measures, support and confidence, in the rule generation process. This direct correlation between the SAE and the ARM provides an interpretation of the proposed architecture. Additionally, pathway analysis

based on the variants within the rules identified by ARM was used as a biological validation of epistasis. Candidate variants and their interactions identified by this approach provide new evidence, widening the potential of genetically supported early diagnosis and prevention of obesity and obesity-related conditions.

1.6 Research Aims and Objectives

In this thesis, obesity aetiology through the effective use of bioinformatics and machine learning algorithms is investigated. Through the course of this research, it is expected to redefine the established upper bound to support early prediction of individuals at risk of becoming obese, demonstrating that artificial intelligence paradigms can bring a fundamental shift in capability to this field. The research aims to investigate:

1. Individual genetic variants (SNPs) or groups of variants associated with obesity susceptibility using a genetic dataset. This is conducted using best practice quality control and association analysis via PLINK to ensure data quality prior to epistatic analysis.

2. Statistical filtering using GWAS to select SNPs for epistatic analysis based on modified P-value thresholds.

3. Interactions between SNPs (epistasis) and machine learning modelling to classify obesity from case-control observations. To explore a novel approach to perform and validate epistasis (SAERMA).

4. Model interpretation using association rule mining. To approximate a set of SNPs that represents the best features extracted by the SAE.

5. Post-analytical interrogation of the results. To report information on overlapping or nearest genes using open source tools.

The proposed methodology could ultimately contribute to the development of new strategies to mitigate the effects of obesity and related comorbidities, providing the research community with candidate genetic variants that can be used in further studies of obesity.

1.7 Novel Contribution

Genetic variant interactions or epistasis discovery has become a subject of active development in the fields of statistics, machine learning and biology. The challenge undertaken in this research is, therefore, to examine how intelligent systems combined with standard bioinformatics approaches (association analysis and quality control) can be applied to identify complex interactions between genetic variants while increasing the effectiveness and efficiency of obesity risk prediction. A novel strategy to detect epistatic interactions in obesity is proposed and described as SAERMA: Stacked Autoencoders Rule Mining Algorithm. Each of the novel contributions claimed in this document are discussed in turn in the following subsections.

1.7.1 Stacked Autoencoders

Stacked autoencoders is a powerful unsupervised feature extraction technique based on a deep learning architecture which has been vaguely explored in the bioinformatics field, with exception to high dimensional gene expression profiles (Danaee et al. 2017) and organ detection in heterogeneous magnetic resonance imaging (MRI) (Hoo-Chang Shin et al. 2013). In this thesis, using SAE as a feature extraction technique combined with GWAS for epistasis is the first study of its kind in the investigation of obesity as a complex disease.

- Stacked autoencoders have been used as a proven alternative to traditional epistatic analysis approaches and to the best of our knowledge is the first comprehensive study of its kind (Fergus, Curbelo et al. 2018; Curbelo, Fergus, Chalmers, et al. 2018).

1.7.2 Association rule mining

The *Apriori* algorithm is an unsupervised machine learning algorithm commonly utilised to identify patterns in large datasets. Although it was originally designed to solve problems in domains such as market basket analysis (transactional databases), its use has been extended to the area of bioinformatics. Particularly, the *Apriori* algorithm has been utilised to extract frequent itemsets to investigate epistasis in case-control data. However, this technique has not been previously used to generate rules in GWAS for subsequent SAE feature extraction and fine-tuned classification modelling in polygenic obesity studies.

- In this thesis, the *Apriori* algorithm is utilised, but extended beyond other works to generate rules from GWAS, which are later combined with SAEs to learn epistatic interactions between SNPs.

1.7.3 SAERMA

SAERMA combines GWAS, ARM and SAE, to provide a tight correlation between SNPs in generated rules (compressed by an SAE to capture epistasis), and a fully connected MLP classification model (weights initialised using the final layer in SAE) that is fine-tuned to classify case-control observations. While GWAS analysis filters the SNPs for dimensionality reduction, changing the interest measures support and confidence in ARM directly affects the

classification results produced. Using an application specific set of performance metric thresholds during model training and test set validation, it is possible to use the ARM model to interpret the deep classification model structures and approximate what SNPs are important and the epistatic interactions between them using the association rules.

1.7.4 Discover tool for new genetic candidate variants

SAERMA is a first step discovery method that allows bioinformaticians to create network plots from association rules configured to provide the best classification results when distinguishing between case-control observations. The most significant rules represent SNP interactions formed by SNPs that often co-occur, with an indication of the direction of the rule. Experts in the field of genetics and medicine can use the outcome of these interactions as candidate variants to facilitate clinical management and better therapies.

To the best of our knowledge this is the first time deep learning SAE and ARM have been combined and used with obesity GWAS data to classify case-control observations with machine learning model interpretation.

1.8 Structure of the thesis

Following this Introduction, Chapter 2 provides background information on obesity as a condition and its occurrence. The chapter also provides an overview of the genetic aspects underlying obesity, which forms the foundation for this research. To complement this, the analytical aspects of genetic analysis are described, with emphasis on how GWAS is conducted and the common data format as required by the bioinformatics tool PLINK. Several key findings made

by GWAS are also presented. Furthermore, the chapter describes epistasis and introduces current approaches within the scope of this research. The discussion investigates how SNP-to-SNP interactions are identified via statistical approaches which typically require dimensionality reduction stages, also described in this chapter. A general overview of epistasis is followed by an introduction to promising and less explored techniques considered in studies on epistasis. These include, association rule mining, neural networks, deep learning and stacked autoencoders. A description of functional analysis available for GWAS hits interpretation are discussed before the concept of systems medicine is introduced as the theoretical framework of this thesis, to close this chapter.

Chapter 3 describes the proposed SAERMA methodology. A complete description of the pipeline that comprises SAERMA is presented. The approach is decomposed into several stages, namely genomic analysis (QC and association analysis), epistasis (ARM and SAE), and classification (MLP). The first stage considers two different aspects: (1) the identification and elimination of variants and/or individuals that introduce bias or erroneous data into the analysis, and (2) statistical filtering for dimensionality reduction using common GWAS techniques. The second stage identifies SNP-to-SNP interactions by, combining ARM and SAEs. This analytical process is evaluated using an MLP trained using the identified SNPs and the interactions between them, to classify case-control instances. The architecture of the proposed network is also presented in this chapter.

In Chapter 4, the results are reported using several experiments and a real case-control dataset obtained from dbGaP. The results obtained in the QC and

association analysis stages are presented first as the gold standard approach in GWAS and as a statistical filtering strategy. Next, baseline logistic regression (GLM) classification results using the filtered SNPs (different P-value threshold) are also reported to demonstrate the use of industry standard models before carrying out experiments with more advanced machine learning techniques. The baseline results with GLM is followed by the results obtained from an MLP trained on the SNPs filtered by different P-value thresholds. This experiment is extended to include an SAE and a final fully connected MLP to investigate epistasis. Although using SAEs is novel and powerful, a final experiment is constructed to provide model interpretability, an inherent limitation of SAEs. This is achieved by combining ARM and SAE models, to generate rules for deep feature learning using the SAE. The correlation between the two models provides a base understanding for the classification results that captures important SNPs and the epistatic relationships that exist between them.

Chapter 5 discusses the results obtained in each of the experiments conducted in this thesis. Experimental evidence shows that SAEs perform well in identifying non-linear interactions between SNPs and provide good classification performance with a substantially reduced number of input features. Adding ARM to the proposed methodology decreases the results, however it provides an interpretation of the model, which is not the case for any of the other experiments presented in this thesis.

In order to validate the rules identified by SAERMA from a biological point of view, functional analysis was conducted and relevant pathways mapping items in the rule were reported in Chapter 6. This chapter represents a proof of

concept analysis to demonstrate that the identified rules are an indication of true epistasis and not mere chance.

Finally, the thesis is concluded in Chapter 7. This chapter provides an overview of what was presented in the thesis and summarises the research findings. While the thesis made several novel contributions and the results were encouraging there is still room for improvement. Consequently, further directions are also included as future work. In particular, it highlights how SNP filtering is less than ideal as many important but less influential features that may hold key information needed to fully understand polygenic obesity could be eliminated.

Overall though, the methodology in this thesis to the best of our knowledge is a world first. The approach provides researchers with a new tool in the quest to better understand epistatic interactions between SNPs in GWAS data.

Chapter 2. INTRODUCTION

2.1 Introduction

This thesis is a multidisciplinary project. Hence, an overview of several concepts from different areas of study is provided in this chapter.

The chapter is organised into eleven main sections. The chapter begins by describing the obesity problem and the effect it has on society. This is followed by a discussion on the genetic perspectives of obesity and an explanation of the core terminology used. The state-of-the-art methods for the analysis of genome-wide data is presented before an overview of the findings achieved by GWAS are discussed. The chapter continues introducing the term epistasis and multivariate approaches for the investigation of obesity epistasis are discussed. Finally, a summary of the chapter is provided.

2.2 Obesity Epidemic

Obesity prevalence has been increasing for several decades and has now reached epidemic proportions (James 2008; Mehlhorn 2010; Hall 2018; Lobstein et al. 2004). This has had a significant impact on morbidity and mortality rates (Flegal et al. 2007; Hirko et al. 2015). According to the World Health Organization (WHO), in 2014, overweight and obesity prevalence worldwide was more than two times higher than in 1980 (World Health Organization 2014). The occurrence of obesity has been typically associated with high-income countries but nowadays, it is also a rising problem in low and middle-income countries (Li, Zhao, Luan, Ekelund, et al. 2010; Gortmaker et al. 2011). If current trends

continue, obesity prevalence will fail to meet the global non-communicable disease (NCD) targets (World Health Organization. Department for Prevention of Noncommunicable Diseases 2017) by 2025, with global prevalence projected to be slightly higher in women than in men (NCD Risk Factor Collaboration (NCD-RisC) 2016).

According to the WHO, overweight and obesity is the fifth leading risk factor for mortality, resulting in approximately 2.8 million deaths globally every year (WHO 2009). Obesity is a known contributor of numerous complications (Segula 2014) which include T2D, cardiovascular disease, and certain types of cancer (Borrell & Samuel 2014; Lifestyles statistics team. Health and Social Care Information Centre 2014; Renehan et al. 2008). Furthermore, overweight and obesity can significantly limit life expectancy (Peeters et al. 2003). In fact, the effects of obesity are so grave that it reduces life expectancy on average by 3 years – in cases of severe obesity this can vary between 8 and 10 years (Bello et al. 2013; Borrell & Samuel 2014). Consequently, governments and organizations from the public and private sector all have a role to play in contributing to obesity prevention (Hilton et al. 2012).

The United Kingdom (UK) is currently ranked as one of the most obese nations in Europe where obesity rates have nearly doubled between 1993 and 2016 (National Clinical Guideline Centre 2014; Carl Baker 2018). Data extracted from the Health Survey for England (HSE) in 2015 revealed that 41% of men and 31% of women were overweight whilst 27% of adults (men and women) were obese, with a body mass index (BMI) equal to 30 kg/m^2 or higher (Department of Health 2016). The morbid obesity figures, although lower,

indicate a greater prevalence in women compared to men (4% for woman and 2% for men). However, in cases of overweight and obesity, statistics tend to be higher in men than in women, 65.7% and 57.1% respectively. Although attention is usually centred on overweight and obesity, it can be noted that morbid obesity cases in adults, in England, increased from 0.8% to 2.9% between 2005 and 2015. A 2016 survey revealed that for every one hundred adults in England, more than a quarter (twenty six adults) were obese, of which three were morbidly obese. Obese and overweight comprised a total of 61.4% of all adults with 2.9% being morbidly obese. In Figure 2-1, an increment in the prevalence of obese adults, both men and woman, can be observed between 1975 and 2014, with especial interest in North America and Europe, Oceania and Latin America and the Caribbean. The figure depicts the prevalence of obesity in adults aged 18 years and older in different regions, measured as the percentage of adults (both male and female) with a BMI $\geq$ 30 kg/m^2.

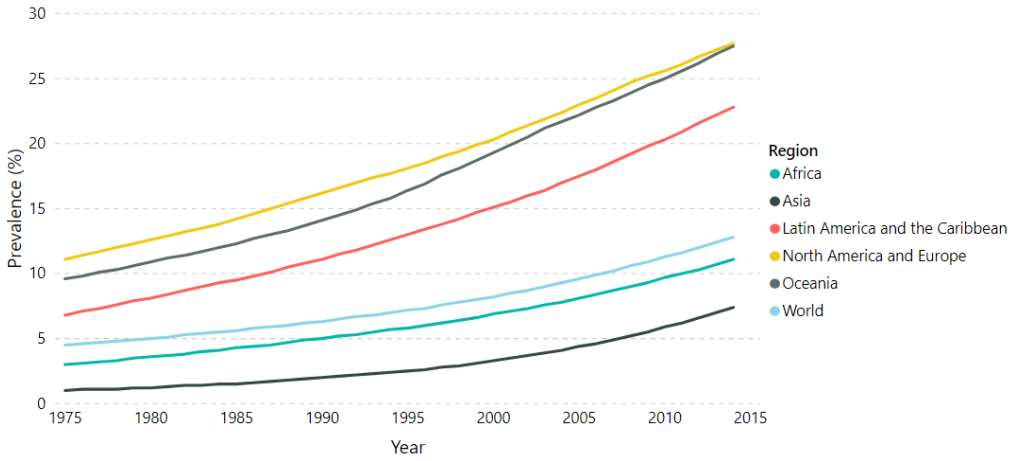


Figure 2-1: Prevalence of obesity in adults by region between 1975 and 2014

Source: UN Food and Agricultural Organization/WHO (Ritchie & Roser 2019).

The annual direct cost of obesity and its consequences to the National Health Service (NHS) was approximately £5.1 billion between 2006 and 2007 (Scarborough et al. 2011). Although calculations of the exact economic cost of obesity is difficult to ascertain, prominent reports, such the one issued by the UK Government's Foresight Programme in 2007, estimated that the NHS costs attributed to overweight and obesity was forecasted to rise to £6.3 billion in 2015, £8.3 billion in 2025 and £9.7 billion in 2050 (Butland et al. 2007). Furthermore, it has also been reported that by 2030, there will be 11 million more obese adults in the UK. This represents an estimated combined medical cost for treatment of associated diseases at £1.9-2 billion per year (National Clinical Guideline Centre 2014).

In general, global obesity rates tend to be higher in adults than in children. However, childhood overweight and obesity is also a major public health problem in economically developed countries as well as in urbanised populations (Shawky & Sadik 2012; Wang & Lobstein 2006). As derived from Figure 2-2, it can be observed that childhood obesity is a serious problem particularly in the United States. The graph shows the overweight trend in a number of countries around the world.

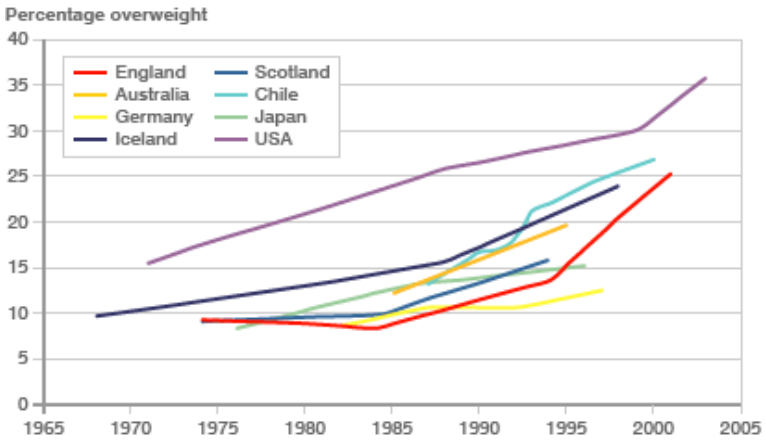


Figure 2-2: Snapshot of the increasing number of overweight children in several countries

Source: Government Office for Science.

Childhood overweight and obesity prevalence increased from 4.2% to 6.7% between 1990 and 2010 worldwide and estimated projections for 2020 predict a rise from 6.7% to 9.1% (de Onis et al. 2010). In England, data extracted from the HSE in 2012 revealed obesity prevalence levels increasing from 11% and 12% to 17% and 16% in boys and girls respectively between 1995 and 2011, reaching peaks of 18% and 19% among boys and girls around 2005 (The NHS Information Centre Lifestyle Statistics 2011). More recent estimations reveal that approximately one in five children in Reception are classified as obese or overweight, while one in three children are identified as obese or overweight in Year 6 (Statistics Team NHS Digital 2017). Obese children are more likely to become obese adults, with associated health problems and consequential costs to the NHS (Whitaker et al. 1997).

The dramatic rise of obesity

An energy imbalance between caloric intake and caloric expenditure leads to an increase in body weight. In other words, obesity arises in excessive absorption

and storage of energy constituents such as carbohydrates and fats, when energy intake surpasses energy expenditure (Pang et al. 2014).

One of the most frequently used quantitative measures of adiposity is the BMI. Body fatness measures can be classified as direct and indirect (Lobstein et al. 2004). BMI is an indirect measure of body fatness although it is not the only one; measures such as skin-fold thickness, waist circumference or waist-to-hip ratio (WHR) are also examples of indirect measures. Conversely, examples of direct measures of body fatness are underwater weighing, magnetic resonance imaging (MRI), computerized tomography (CT), Dual-Energy X-ray Absorptiometry (DEXA) or bioelectrical impedance analysis (BIA) among others. People with a very muscular physique (i.e. athletes) can have a high BMI but not necessarily excess fat. In these cases, BMI cannot be used to diagnose obesity conclusively. Despite this limitation, BMI is a convenient population-level indicator of whether someone is in healthy weight, overweight or obese since it is a cheap non-invasive index of relative adiposity that is widely accepted (Value et al. 2016).

BMI is formally defined as weight in kilograms (kg) divided by the square of the height in metres (m^2), as indicated in Equation 2-1:

$$\mathrm{BMI} = \frac{\mathrm{kg}}{\mathrm{m}^2} \qquad 2\text{-}1$$

The BMI classification for the general population recognised by the WHO and other organisations is shown in Table 2-1 (Bjorntorp et al. 2000; Jensen et al. 2014; U.S. Department of Health and Human Services 1998). This

classification is the same for both sexes, but it may differ in different populations because of differences in body proportions.

Classification		BMI (kg/m2)	Risk of comorbidities
Underweight		< 18.5	Low
Normal weight		18.5 - 24.9	Average
Overweight		25 – 29.9	Increased
Obese	Class I	30.0 - 34.9	Moderate
	Class II	35.0 - 39.9	Severe
	Class III	≥ 40	Very severe

Table 2-1: BMI classification for adults according to the WHO

Source: World Health Organisation

Therefore using (Table 2-1), someone is considered clinically overweight or obese if his/her BMI is higher or equal to 25 kg/m^2 and 30 kg/m^2 respectively. Since the risk of comorbidities increases with BMI values of 25 kg/m^2 or higher, BMI levels in adults should be sustained within the range 18.5−24.9 kg/m^2 to maintain optimal health. BMI cases of obesity class III, are termed morbidly obese (BMI ≥ 40 kg/m^2) and, are at the highest risk of morbidity and mortality (World Health Organization 2014).

Physical activity and homeostatic metabolic processes are mechanisms through which energy is consumed in the human body (Pang et al. 2014), although several other factors play a significant role in the development of obesity. Among these factors, genetic predisposition, physical activity, caloric intake (diet) and socioeconomic factors are included (Smith & Smith 2016). While an imbalance between energy intake and expenditure drives obesity, its aetiology is complex and multifactorial. Thus, providing a conclusive

explanation of the causes behind the obesity epidemic is not an easy task (Hall 2018).

The ubiquitous availability of low-cost hypercaloric food combined with an increasingly sedentary lifestyle and other environmental factors have played a fundamental role in the development of the obesity epidemic. This is true since the prevalence of obesity has drastically risen while changes in the genes are unlikely to have happened so rapidly. Surprisingly, not every individual exposed to such environments, also known as obesogenic environments (Jones et al. 2007), becomes obese. As stated earlier, the aetiology of obesity is multifactorial, indicating that lifestyle and environmental factors may interact with multiple genes, thus causing this disorder. This is further supported by twin, adoption, and family studies which found that variation in BMI was largely due to heritable genetic differences, with heritability (Min et al. 2013) (the proportion of the variability of a trait/phenotype that is attributable to additive genetic factors among individuals in a given population) estimates in adults ranging between 40 and 70 percent (Maes et al. 1997; Walley et al. 2006; Vogler et al. 1995; Schousboe et al. 2003; Malis et al. 2005). Nevertheless, due to differences in study types, populations, and the age group targeted, these estimates have broadly fluctuated across studies. Hence, it is believed that obesity risk is higher among those individuals genetically predisposed to gain weight and who are exposed to obesogenic environments where gene-environment interactions occur.

Current management strategies

Managing the obesity epidemic can be costly as obesity places a significant financial burden on the healthcare system in most countries (Withrow & Alter 2011). However, overweight, obesity and their related comorbidities are preventable. In this sense, governments and communities are best placed to promote healthy diets and regular physical activity. Intervention options for obesity may include non-surgical and surgical treatment. Non-surgical treatment often involves calorie reduction via dietary changes, behaviour modification, physical activity, and when necessary, pharmacotherapy or psychological support. Pharmaco-therapeutic and bariatric surgery approaches are examples of ways to deal with obesity (Gautron et al. 2015). Although if the genetic predisposition to this disease is identified, health services from many countries will save money, using interventions at a much earlier stage in life. Greater awareness of the causes of obesity would benefit the planning and development of international collaborations and programs to solve this growing public health crisis. Consequently, early detection and prevention strategies are more suitable options for all affected countries that differ from traditional, physician centred, diagnosis and treatment models (Kirk & Penney 2013).

2.3 Genetics of Obesity

Over the past century, obesity has been comprehended differently, with different theories supporting its aetiology. It was at the end of the 1980s and during the early 1990s when the first twins and adoption studies revealed genetic factors with robust implications in body weight regulation (Jou 2014).

Twins studies allowed the relative contributions of genetics to be unravelled as opposed to the environment, for a variety of human traits. In these studies, monozygotic (MZ or identical) and dizygotic (DZ or fraternal) twin pairs are compared to evaluate the impact genetic and environmental influences have on specific traits under investigation. For a given trait or condition in one of the twins, the idea is to find what the likelihood is that the other twin possesses the same trait or condition. Hence, it is said that genetic differences are present if MZ twins show more similarities on a given trait compared to DZ twins. In contrast, environmental factors are more likely to influence the trait if MZ and DZ twins share a trait equally.

In 1986, Albert J. Stunkard et al. conducted one of the most convincing twin studies in obesity, where they proved that an individual's weight could be governed by the individual's parentage (Stunkard et al. 1986). The authors investigated data from 540 adult Danish adoptees and their biological and adoptive parents, to measure the relationship between BMI of both parents and adoptees. Their findings revealed that adoptees' BMI were similar to those of their biological parents instead of their adoptive parents, even though adoptees shared the environment with adoptive parents. Consequently, the authors concluded that for most adoptees, obesity was inherited from their biological parents, suggesting that the family environment on its own has no clear effect in the development of obesity.

In 1990, a group of researchers including Albert J. Stunkard, conducted an adoption study using a Swedish twin registry to investigate genetic weight regulation further (Stunkard et al. 1990). Twins that were raised by their

biological parents and twins who were raised by an adoptive family were compared. The authors confirmed, one more time, that for identical twins their weight was practically identical to their biological parents regardless of the environment where they grew up. More recently, a meta-analysis of the heritability of human traits conducted using information from fifty years of twin studies, provided a good overview of the relative contribution genes and the environment have on specific traits (Polderman et al. 2015). The results revealed that all the investigated traits were heritable.

In addition to twin studies, hypothesises such as the 'thrifty gene' proposed by Neel in the 1960s, suggested that in populations that experience frequent periods of starvation, genes that predispose them to obesity had a selective advantage (Neel 1962). This, combined with today's obesogenic environments, might cause a disproportion of body weight for those who carry these genes. However, this theory has been considered controversial (Speakman 2007). A more plausible theory called 'predator release hypothesis' argues that the lack of predation risk in modern societies might justify the distribution of obese individuals in the population, since negative pressure on genetic variants no longer impacts modern day living in developed countries (Speakman 2007).

Family, twin and adoption studies as well as natural selection hypothesis have provided solid evidence to justify moderate to high heritability of BMI. Although twin studies suggest that certain disorders or traits have a genetic component, this does not provide information about the gene or genes involved and their location.

2.3.1 Central Dogma of Molecular Biology

To better understand the context of the data analysed in this thesis, a brief introduction to the central dogma of molecular biology and the basic genetic concepts utilised in this thesis is provided. For a more in depth discussion of the topic, the reader can refer to (Watson et al. 2014; Alberts et al. 2014; Alberts et al. 2015).

The DNA or Deoxyribonucleic Acid is a double-stranded macromolecule formed by two long polynucleotide chains (DNA strands) running in opposite directions, which enclose all the biological instructions to build and maintain an organism (Alberts et al. 2015). Fundamentally, DNA encodes a sequence of four building blocks named nucleotides, abbreviated as A (Adenine), G (Guanine), T (Thymine) and C (Cytosine) which, combined, specify most of the amino acid sequences of proteins. In Figure 2-3, a DNA molecule composed of two antiparallel strands and its building blocks is depicted, where the arrowheads at the end of each strand represents the polarities. These nucleotides can be seen as letters in a four-letter alphabet which allows for the spelling of biological messages. Hence, differences between organisms are due to differences in the nucleotide sequences in their respective DNA molecules, which represent different biological messages.

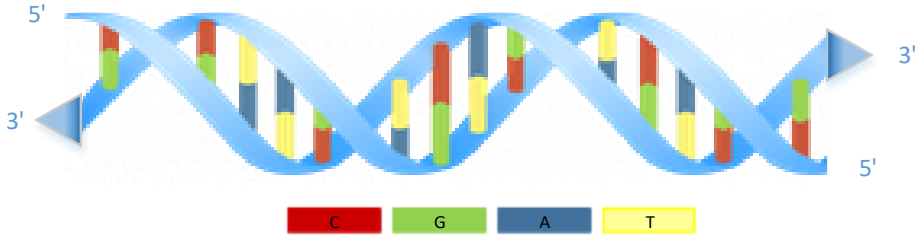


Figure 2-3: DNA chain internal structure

A very large number of different genetic messages can be produced (virtually infinite) with only four nucleotides (4^N, for N nucleotides in the sequence).

The long chain of DNA is separated into smaller biological segments or functional units known as genes, that contribute to manifestation of phenotypes (Alberts et al. 2014). In fact, the most relevant function of DNA is to carry those genes which contain information to specify all the proteins and RNA molecules that constitute an organism. The whole human genome, that is the totality of genetic information belonging to our species, is divided into a number of genes and distributed over 23 different pairs of chromosomes, where the position of a gene within the genome is called locus. Following the publication of the full human DNA sequence (Lander et al. 2001), the knowledge about gene distribution along each chromosome was improved. In Figure 2-4, an example of how genes are organised on a human chromosome is depicted (Alberts et al. 2015). The example is based on one of the smallest chromosomes, chromosome 22, which represents about 1.5 per cent of the human genome. Figure 2-4 (2) represents a segment of chromosome 22 where known and predicted genes are indicated in brown and rose respectively, whereas an expanded section of Figure 2-4 (2) with 3 genes is shown in Figure 2-4 (3). Finally, Figure 2-4 (4) shows an exon-intron arrangement of a representative gene, where exons (rose) are sections of DNA (or RNA) that code for proteins whilst introns (grey) are non-coding sections of an RNA transcript or the DNA encoding it.

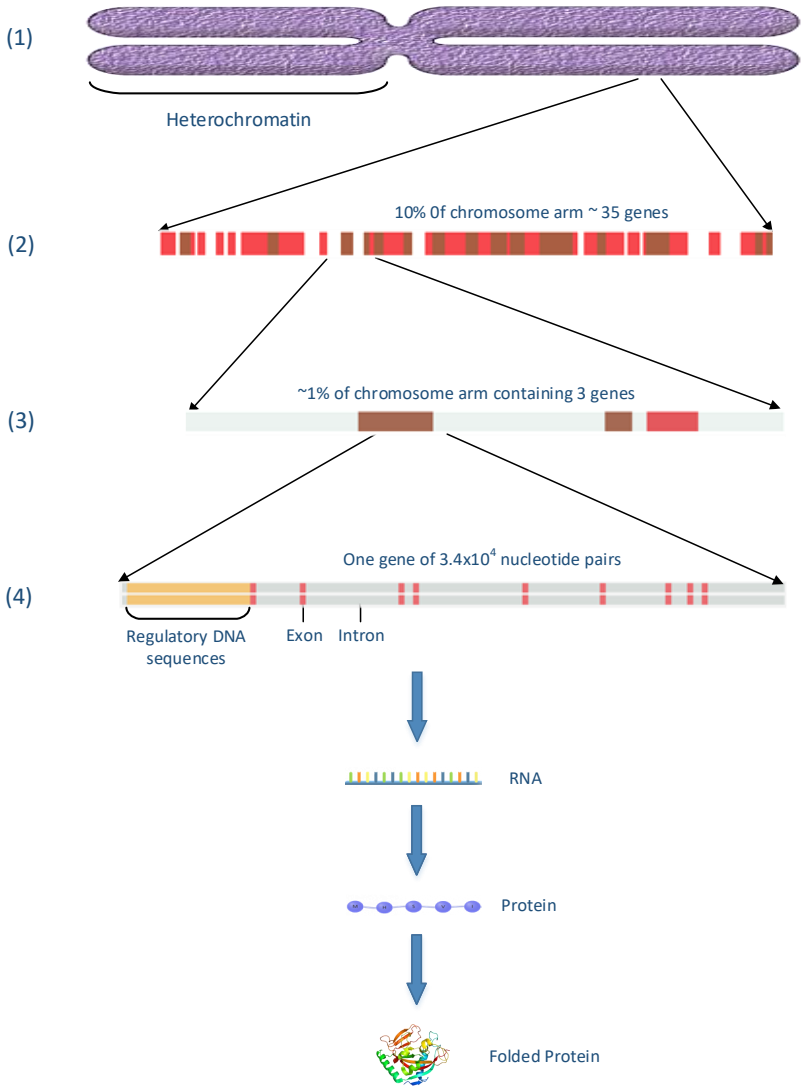


Figure 2-4: Example of organisation of genes on a human chromosome

Coined in the early days of modern biology by Francis Crick, the central dogma has been accepted as the biological pathway where information flows from gene to protein in a unidirectional way (Alberts et al. 2014). It focuses on how proteins are synthesized from DNA, according to the flows of information from DNA to Ribonucleic Acid (RNA) and RNA to protein as shown in Figure 2-5. Genes encode amino acid chains, the building blocks of proteins, which have specific functions in the organism (Alberts et al. 2015). From top to bottom, the arrow encircling DNA in Figure 2-5 means that DNA governs its own replication; the process represented by the arrow between DNA and RNA

(transcription) indicates that the synthesis of RNA is controlled by a DNA template that rewrite the DNA sequence in a similar RNA alphabet (in eukaryotes cells, RNA becomes messenger RNA); whereas in the translation process, where proteins are synthetized, the messenger RNA (mRNA) is decoded to specify the amino acid sequence of a polypeptide. In this biological flow of genetic information, the possibility of RNA to be determined by proteins or DNA to be made on RNA templates has been neglected. Not all genes specify proteins, instead they can provide the instructions to build functional RNA molecules that play roles in translation (i.e. transfer RNAs and ribosomal RNAs) (Watson et al. 2014).

"The 'Central Dogma' is the process by which the instructions in DNA are converted into a functional product." [2]

[2] http://www.yourgenome.org

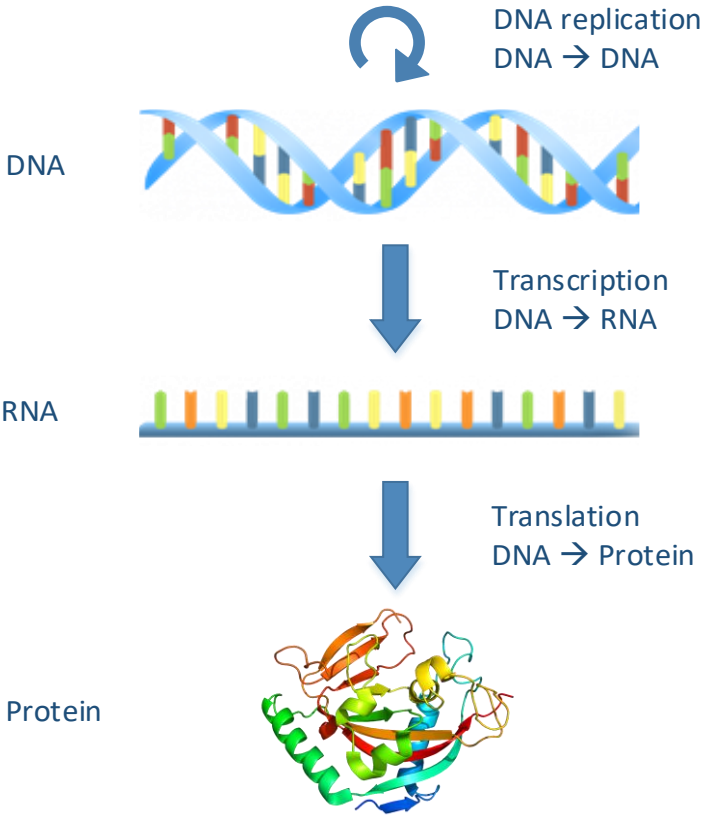


Figure 2-5: Illustration of the central dogma of molecular biology

Although the overall structure of the central dogma (DNA→ RNA→Protein) has remained intact over the past century, this fundamental biological law as an absolute principle has been questioned as it seems to be more complex than previously thought (Koonin 2012). In addition to serving as intermediate carriers of genetic information, RNA molecules are responsible for many critical tasks in the cell. The obsolete protein-centred version of the central dogma treated genomic regions transcribed into non-coding RNAs (ncRNAs) as '*junk*' with no biological meaning (Ling et al. 2015). Although many human transcripts are not translated into protein, many of these are functional. When DNA is transcribed into both coding and non-coding RNA, a subsequent translation of the coding RNA is produced into protein while a concurrent regulation of these steps is controlled by non-coding RNA. Figure 2-6 depicts a most up to date version of the central dogma taking the role of ncRNAs into

consideration (Marques et al. 2015). Therefore, there is a special interest in how ncRNA transcripts modulate gene expression and their role as epigenetic modifiers.

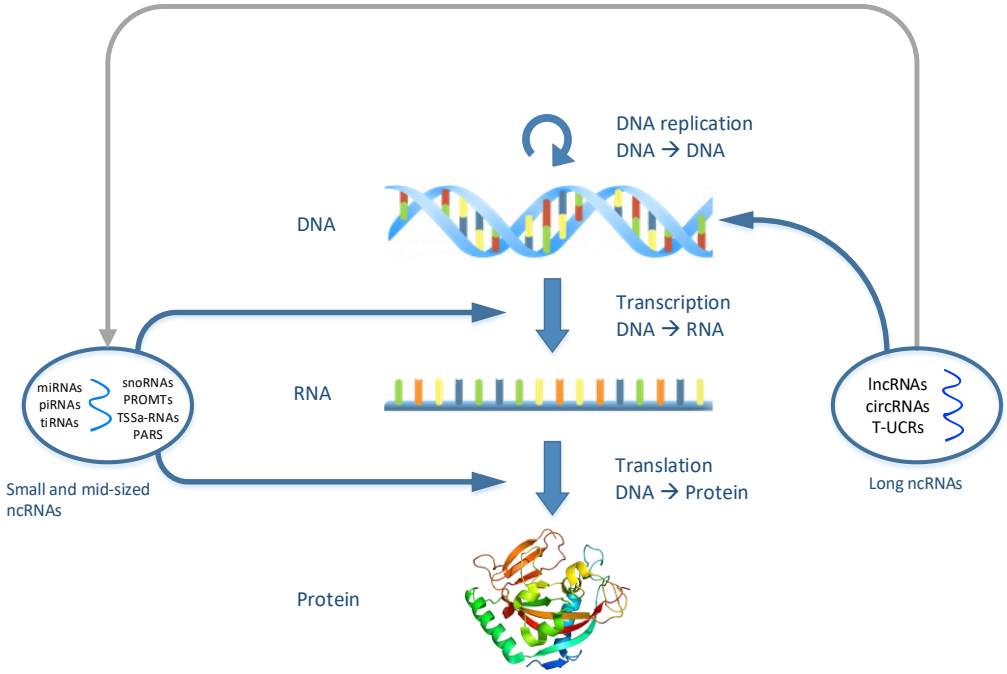


Figure 2-6: Role of ncRNA in a more up to date version of the central dogma

The regulation of gene expression in the absence of a change in the underlying nucleotide sequence is known as epigenetics (Waterland & Michels 2007). Epigenetic changes are typically reversible, are associated to chemical modifications to DNA and can alter the way transcription of genes is controlled. This has the potential to promote pathologies by deactivating specific genes or by aberrantly expressing others. Classical epigenetics mechanisms are governed by DNA methylation or posttranslational modification of histones although gene expression can vary based on the function of RNA molecules as well as their interactions with DNA and/or proteins. While DNA methylation correlates with transcriptional suppression, histone modifications activate or repress gene

expression depending on modification type (acetylation, phosphorylation, methylation and others) and locus (Gasperskaja & Kučinskas 2017).

To date, a class of small ncRNAs termed microRNAs (miRNAs) are the most widely studied (He & Hannon 2004), although it is believed that many other contribute to the development of many human disorders (Esteller 2011). These can be contextualised according to their varying sizes of non-coding transcripts: short ncRNAs (miRNA and piRNAs) below 40 nucleotides (nt) in length, mid-sized ncRNAs (snoRNAs) between 60-300 nt, and long ncRNAs (lncRNAs) that are at least 200 nt in length. Of these, long ncRNAs have attracted much attention as they are large in number and due to their functional relevance in complex disorders (Ling et al. 2015). These type of ncRNA wield changes in gene expression throughout different mechanisms highlighted by Peschansky & Wahlestedt (2014), necessary for appropriately targeting of histone modifying complexes or play a role in DNA methylation.

Non-hypothesis-driven studies based on large scale genotyping from population-based samples (i.e. GWAS), commonly used for disease and trait gene association, have provided valuable information for investigating the genetic architecture of human disease. As discussed later in this thesis, such approaches do not provide a clear molecular link between genetic markers and the phenotype under investigation, owing to the fact that most of these disease-related genetic variants are located in intergenic or non-coding regions of the genome. Therefore, it is of critical importance to consider the genetic context for a more comprehensive functional annotation in the investigation of complex diseases.

This has the real potential to find better ways to treat and prevent diseases through early detection and prevention strategies, personalized drugs and tailored therapies. The completion of the human genome has made a significant contribution in achieving this, principally because researchers can now understand how species function and how phenotypes and diseases are made.

2.3.2 Glossary

Several relevant technical terms will be listed and defined in this section. Uses of these terms in the remainder of this thesis are described in Watson et al., (2014) and listed below.

- *DNA:* deoxyribonucleic acid (DNA) is a double-stranded macromolecule which consists of two long polynucleotide chains composed of four types of nucleotide bases.
- *Gene:* segment of the DNA chain that includes the nucleotides needed to encode the amino acid sequence of a protein. It is the fundamental unit of heredity.
- *Chromosome:* long linear DNA molecule associated with proteins. Its most important function is to carry genes.
- *Locus/Loci (plural):* position of a gene in the genome.
- *Genetic marker:* defines a genomic region, i.e. a segment of DNA that varies among individuals.
- *Allele:* one of the two or more versions of a given gene that can exist at a single locus.
- *Trait:* an attribute of a phenotype.

- *Phenotype:* individual's physical appearance or biochemical characteristic.
- *Genotype:* genetic makeup of an organism. Actual alleles of an individual.
- *Homozygous:* having two identical alleles on a pair of chromosomes at a given locus.
- *Heterozygous:* having two different alleles on a pair of chromosomes at a given locus.
- *Minor allele:* it refers to the second most frequent allele.
- *Amino acids:* building blocks for proteins.
- *Linkage Disequilibrium:* the occurrence of some genes together, more often than would be expected by chance.
- *Protein:* specific sequence of amino acids, responsible for body structures.
- *Association:* Statistically significant correlation between a biological/genetic marker and a disease or phenotype.

The definitions provided in this section are intended to aid the reader to understand the context of this thesis. However, a full genetic background is not provided since it is not the main topic of this multidisciplinary research project.

2.3.3 Human Genetic Variations

Approximately 99.9% of the base pairs (nucleotides) in the human genome are identical between any two individuals (Gonzaga-Jauregui et al. 2012). This level of similarity defines us as species. Hundreds of complex phenotypic traits contribute to our appearance and behaviour, as well as to our predisposition to certain diseases (differences in the remaining 0.1 percent hold important evidences about the causes). Complex phenotypes are thought to be

characterized by a combination of hereditary factors in the form of genetic variants, as well as environmental influences. During the past several decades, the main challenge has been to determine which genetic variants are behind inherited phenotypic components.

Human genetic variants are usually classified as common, low-frequency or rare variants, depending on whether the minor allele frequency (MAF – discussed below) in a given population is higher than 5%, between 1-5% or less than 1% respectively (Bomba et al. 2017). Common variants tend to have a very weak effect on the phenotype while low-frequency and rare variants have small to modest effects. Based on this classification, common variants are typically referred to as polymorphisms whilst rare variants are termed mutations (Karki et al. 2015). Furthermore, depending on their nucleotide composition, genetic variants are separated into two classes: single nucleotide variants (SNVs) and structural variants (SVs) (Kidd et al. 2008).

For the purpose of this PhD, only single nucleotide variants are considered.

2.3.4 Single Nucleotide Polymorphism - SNP

Single nucleotide polymorphisms or SNPs (Gray 2000; Dunnen & Antonarakis 2000) are the most common type of genetic variation among humans, and have become the genetic marker of choice in the genetic mapping of complex traits, including obesity and diabetes, to name only a few (Dudoit & van der Laan 2008). DNA sequences are constituted by a chain of four building blocks or nucleotides as discussed earlier: A, G, C, and T. The human genome contains over 3 billion base pairs (nucleotides). The sequence of these chemical bases

determines the biological instructions contained in DNA, for building and maintaining an organism. A genetic variation is classified as an SNP when at least 1% of a given population does not carry the same nucleotide at a specific position in the DNA sequence. Hence, each SNP represents a variation in a single building block or nucleotide in the DNA sequence. Consider the following example depicted in Figure 2-7, where two sequenced DNA segments from different individuals include a variation in a single base (nucleotide). Since the two DNA fragments ACTTG**C**GA and ACTTG**T**GA contain a difference at the same locus, it is likely that there is an SNP at this position of the genome. Furthermore, SNPs can be identified in both coding and non-coding regions of the DNA, for example, in a regulatory region or between genes respectively. SNPs within a gene or in a regulatory region near a gene may have functional significance with a direct role in disease.

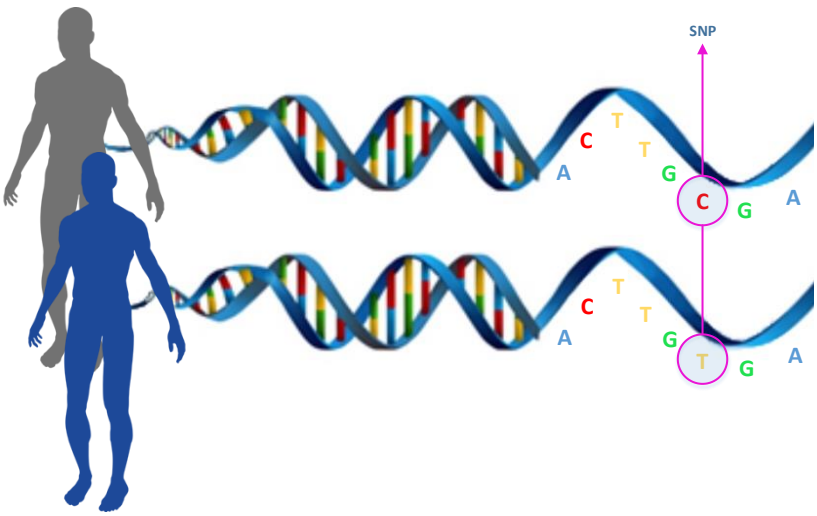


Figure 2-7: Example of a SNP representation

Within a population, SNPs are generally diallelic (i.e. have two alleles at a site within a loci), e.g. C or G. Note that, while locus is used to refer to the location of a gene in the genome, site refers to the location of alleles of a SNP

within a loci or within the region of a locus (Elston et al. 2012). Furthermore, the SNP frequency is given in terms of the MAF (the less common allele), so that an SNP with minor allele (C) frequency equal to 0.30 means that 30% of the population has the C allele while the 70% of the population carries the, more common, major allele.

SNP data has been used in different study approaches aiming to elucidate the underlying causes of common and rare diseases and is proving to be very important in the study of human health. This has contributed to the investigation and identification of genes with significant roles in obesity aetiology.

2.3.5 Genetic architecture of Obesity (Types)

Up to now, research has provided substantial evidence about the heritability of obesity through twin and adoption studies, as discussed earlier in this chapter. However, molecular approaches have contributed to the discovery of the first human genes and syndromes associated with obesity (Mutch & Clément 2006).

Based on its aetiology, human obesity has been historically classified into monogenic, syndromic and polygenic obesity (common obesity) (Cummings & Schwartz 2003). Next, each type of obesity is briefly described.

2.3.5.1 Monogenic obesity (Non-syndromic)

Single gene alterations can lead to monogenic obesity (Farooqi 2008). Rodent mouse experiments have made it possible to identify most of the monogenic forms of obesity in humans, with many genes involved in the regulation of appetite via the leptin-melanocortin pathway (Farooqi 2008). The first evidence of a single-gene causing severe obesity in humans was the Leptin (LEP) gene,

reported in the late 1990s by C. Montague, I. Farooqi, J. Whitehead et al. (Montague et al. 1997). The authors examined two severely obese children of a highly consanguineous family, where leptin deficiencies where found by sequencing the LEP gene, indicating a strong influence in energy balance. The leptin protein acts in the hypothalamus part of the brain, which controls eating behaviour and plays a crucial role in regulating body weight by inhibiting food intake and stimulating energy expenditure. Extreme obesity treatment has been successful in cases of leptin insufficiency in children, where injections of recombinant human leptin has led to a reduction in body weight and fat mass (Farooqi et al. 2002). In addition to LEP and its receptor (LEPR), other mutations in the melanocortin 4 receptor (MC4R), pro-opiomelanocortin (POMC) and proprotein convertase subtilisin/kexin-type 1 (PCSK1) genes have been shown to cause monogenic forms of obesity (Nordang et al. 2017). Findings from studies on monogenic disorders leading to human obesity have been reviewed and summarised elsewhere (Muñoz Yáñez et al. 2017; O'Rahilly 2009; Chung 2012).

These mutations have led to cases of obesity observed from a very early stage in childhood and have very strong biological effects. However, these monogenic disorders are rare and therefore insufficient for justifying current levels of obesity in the population.

2.3.5.2 Syndromic obesity

Syndromic forms of obesity, also known as pleiotropic syndromes, have also provided additional insights into the mechanisms causing obesity (Milani et al. 2014). These relatively rare forms of obesity, also caused by discrete genetic

defects or chromosomal abnormalities, are additionally accompanied by neurological disease (i.e., mental retardation and/or intellectual disability), dysmorphic features and developmental abnormalities. A syndrome is defined as a cluster of signs and symptoms persistently appearing together. Several syndromic forms of obesity have been recognised (Shawky & Sadik 2012; O'Rahilly 2009; Chung 2012). Among the most well-known syndromes where obesity is one of the main phenotypes (clinical feature), Prade-Willi syndrome (PWS) and Bardet-Biedl syndrome (BBS) can be highlighted (Farooqi 2008).

2.3.5.3 Polygenic obesity "Common obesity"

Early obesity studies were predominantly driven by research into monogenic and syndromic obesity (Walley et al. 2006). The knowledge provided by this research has contributed enormously to the understanding of the physiology underlying appetite and feeding behaviour facilitating, in some rare cases, the treatment of affected individuals.

Polygenic (common) obesity represents the third subgroup which affects most obese cases in the general population (Herrera & Lindgren 2010). While monogenic obesity is produced by mutations in a single gene with a major effect on the development of severe obesity, polygenic forms of obesity are determined by the cumulative effect of environmental factors and multiple common genetic variants or SNPs, each with modest effects on the phenotype (Li, Zhao, Luan, Luben, et al. 2010).

The idea that common obesity is likely influenced by genetic variation that is also common in the general population (disease-predisposing alleles with relatively high frequencies), is supported by the common disease-common

variant (CD-CV) hypothesis (Reich & Lander 2001). Alternative hypothesis, such as common disease-rare variant (CD-RV) argue that the main contributors to genetic susceptibility underlying common diseases are rare variants with relatively high penetrance (disease-predisposing alleles with lower frequencies in the population) (Schork et al. 2009). These hypotheses might not be mutually exclusive, so a combination of both rare and common variants can predispose individuals to complex or common diseases. Figure 2-8 extracted from McCarthy et al., (2008); Manolio et al., (2009) represents a visual map where the rare and common variants are conceptualised based on allele frequency and effect size (the strength of genetic effect). Mendelian disorders with highly penetrant alleles (the likelihood of manifesting a trait given a specific genotype or combination of genotypes is high) are extremely rare and have large effect sizes (top left in Figure 2-8) whereas variants identified in common diseases tend to have small effect sizes and higher frequency (lower right in Figure 2-8).

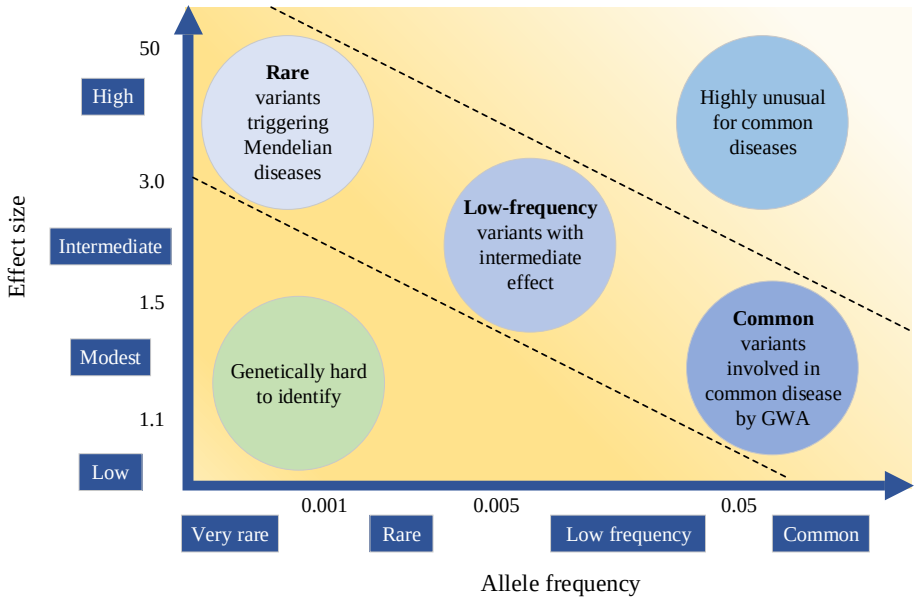


Figure 2-8: Visualisation of allele frequency vs effect size variant definition spectrum

Common obesity has been chosen in this PhD as the main phenotype of interest. Hence, the rest of this thesis will be based on this subgroup only, with particular interest in the study of common variants. In the following sections, the main approaches utilised to discover SNPs associated with polygenic obesity will be introduced.

2.3.6 Identifying genetic loci for common obesity

In 2004, the National Human Genome Research Institute (NHGRI) started a funding program aimed at reducing the cost of human genome sequencing to $1,000 or less (Schloss 2008). Figure 2-9 shows how the cost of sequencing a human genome has been significantly reduced since 2001. This is primarily due to advances in sequencing techniques (Wetterstrand KA 2018) where prices between 2001 and 2007 are for Sanger sequencing, whereas from 2008 onwards, costs are based on next-generation sequencing (NGS). A disruption in the human genome sequencing cost was observed around 2007 after the first individual human diploid sequence obtained by Sanger sequencing, followed by the first complete individual human genome sequenced with new revolutionary genomic tools (i.e. NGS) (Zhang et al. 2011).

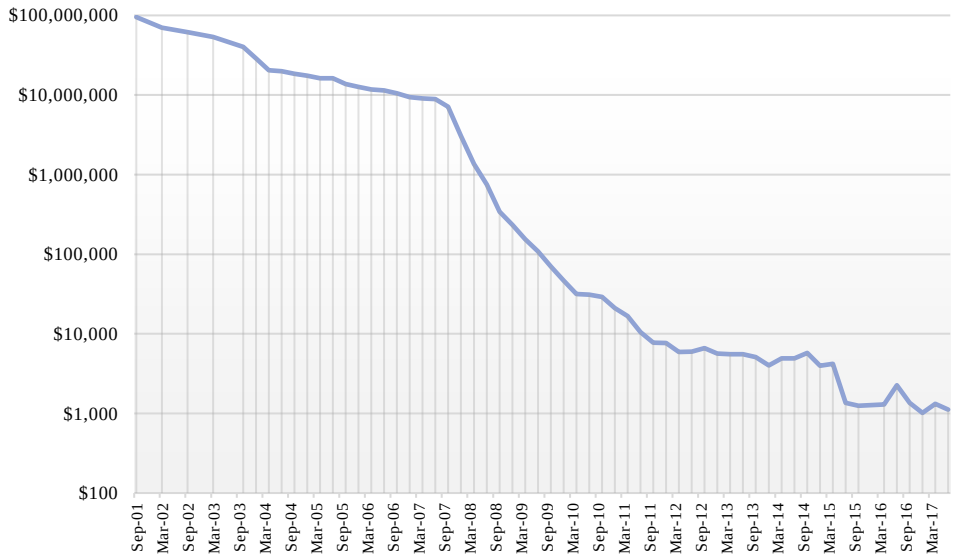


Figure 2-9: Cost of human genome sequencing according to the NHGRI

This has resulted in a plethora of sequencing methods designed to determine the order of nucleotide bases that make up a molecule of DNA (Grada & Weinbrecht 2013). The first generation of sequencing, known as Sanger sequencing, was developed by Frederick Sanger and his team in the 1970s and it has been widely used for nearly 30 years due to its reliability. This led to biology's first large-scale project, the Human Genome Project (Green et al. 2015; Lander et al. 2001). The Human Genome Project has transformed biology by providing a platform and tools to decipher genes in a reliable and reproducible manner. Nevertheless, Sanger approaches are limited in terms of scalability, time and resources, resulting in faster, higher throughput and cheaper technology.

Sanger sequencing has now been replaced with second-generation sequencing also known as NGS technologies which provide high throughput and cheaper DNA sequencing alternatives at unprecedented speeds (Shendure & Ji 2008; Rabbani et al. 2014). They permit large scale studies to be conducted

and for faster sequencing of entire genomes, which have revolutionised the study of genomics and molecular biology. One of the first NGS platforms to become commercially available as a product was the 454 system (454 pyrosequencing), although other platforms such as Solexa/Illumina sequencing and Sequencing by Oligo Ligation Detection (SOLiD) were also available (Shendure & Ji 2008). Some examples of well-known sequencing companies are Illumina, Roche, and Life Technologies among others. Illumina is currently one of the most prominent NGS platforms, offering the highest throughput and lowest per-base cost (Van Dijk et al. 2014). Sequencing of the first human genome, co-published in Nature (Lander et al. 2001) and Science (Venter et al. 2001) in 2001, took 15 years and approximately 3 billion US dollars. In contrast, Illumina with the HiSeqX™ machines released in 2014, sequenced more than 45 human genomes per day at a cost of ~ $1000 per genome (Van Dijk et al. 2014). This has shown significant advancement over the last 20 years or so, reaching the $1000 milestone set by the NHGRI.

Although NGS continues to be the preferred approach, third-generation sequencing technologies, based on single molecule detection and real-time sequencing, are gaining significant interest (Schadt et al. 2010). It is reported that third-generation sequencing technologies will provide significantly longer sequence read lengths over a much shorter period of time and with lower costs. Examples of third generation sequencing technologies are provided by Pacific Biosciences' (PacBio) and Oxford Nanopore Technologies, although other approaches do exist or are under current development (Schadt et al. 2010). Oxford Nanopore Technologies released a commercially available portable sequencer, the MinIONTM nanopore sequencer, which enables single molecule

sequencing in real-time. Nanopore argue that performing whole-genome sequencing (WGS) in humans is hypothetically possible using a single portable MinIONTM sequencer, although several challenges are still present, including high error rates in sequencing reads (Jain et al. 2018).

These advances being made in genetic knowledge and genotyping technology, along with sequencing cost reduction, has contributed to the development of powerful tools and approaches for gene discovery. Linkage and candidate gene studies have been successful in identifying loci associated with rare single gene disorders (i.e. monogenic or syndromic obesity), but less successful in identifying genetic variants that affect common diseases. However, more modern, methodologies such as genome-wide association studies (GWAS, discussed below) have been used to study the genetic basis of both monogenic and common obesity with more success (Bailey-Wilson & Wilson 2011). Most common GWAS have used microarray-based techniques for the identification of disease associations in the genome, interrogating between 500,000 and over one million SNPs per individual (Bush & Moore 2012). However, this number is far from the 3.2 billion nucleotides present in the human genome. Alternatively, whole-exome sequencing (WES) and whole-genome sequencing (WGS), have recently arisen to overcome coverage limitations of GWAS chips (Schwarze et al. 2018).

Array-based Technologies

Despite important reductions being made in the cost of WGS, the cost of sequencing a single individual using this technique remains a constraint. However, in the past years, efforts have concentrated on targeting regions of

interest in the genome, using selective DNA enrichment techniques, to improve efficiency and overall cost. Such strategies allowed, in 2011, a reduction of the overall cost of sequencing of a single individual to approximately $10,000 in comparison to $100,000 using whole-genome sequencing (Zhang et al. 2011).

DNA microarrays represent a high-throughput and cost effective automatically genotyping assay, extensively used in gene expression, transcription factor binding and genotyping analysis (Bumgarner 2013). Microarray procedures rely on parallel quantitative measurement of various sequences in a complex mixture, where tagged nucleic acid molecules in solution hybridize to complementary sequences fixed on a solid substrate (Cummings 2000). Several microarray approaches have been developed, although the most commonly used are oligonucleotide and complementary DNA (cDNA) microarrays (Pereira et al. 2015).

While microarrays have proved to be useful in many applications, they present some limitations. For example, microarrays provide an indirect measure of the relative concentration of different DNA or RNA, are difficult to design so different genes are not bind to the same probe on the array and, they are limited to detect sequences that the array was originally configured to detect (Bumgarner 2013). Despite these limitations, DNA microarray technologies have matured over the past decades becoming the tool of choice for numerous studies, as they are reliable and well-reproducible techniques when appropriately used (Sánchez-Pla 2014).

SNP arrays emerge as a type of DNA microarrays used to detect SNPs within populations where, for each identified polymorphism, the array contains the

possible variations at the specific site (Sánchez-Pla 2014). Identification of SNPs using SNP arrays varies between companies, although the most common ones are those used by Affymetrix and Illumina arrays, allele discrimination by hybridization and, allele-specific extension and ligation to a bar-code oligonucleotide hybridized to a universal array respectively. The variability of available platforms and their specific configurations (allele calling and file formats) complicates the integration of information from various sources, especially for researches with limited bioinformatics skills (Louhelainen 2016). As previously mentioned, microarray-based technologies can genotype over one million SNPs concurrently. An example of some arrays for the human genome available from Affymetrix and Illumina are listed in Table 2-2, which has been extended from (Lamy et al. 2011; Illumina 2010).

Affymetrix	**#Arrays**	**#SNPs**
GeneChip Human Mapping 10K 2.0 Array	1	10,204
GeneChip Human Mapping 100K Set	2	116,204
GeneChip Human Mapping 500K Array Set	2	500,568
Genome-Wide Human SNP Array 5.0	1	500,568*
Genome-Wide Human SNP Array 6.0	1	906,600**
Illumina	**#Samples**	**#Markers**
HumanCytoSNP-12 DNA Analysis BeadChip	12	299,140
Human660W-Quad v1 DNA Analysis BeadChip	4	657,366
HumanOmniExpress BeadChip	12	730,525
Human1M-Duo DNA Analysis BeadChip	2	1,199,187
HumanOmni1-Quad BeadChip	4	1,140,419***
HumanOmni1S-8 BeadChip	8	1,185,076***
HumanOmni2.5-Quad BeadChip	4	2,450,000***
HumanOmni2.5-8 BeadChip	8	2,379,855***
Omni5 BeadChip	4	4,301,331***

*Additional 420,000 non-polymorphic probes for copy number analysis.
** Additional 946,000 non-polymorphic probes for copy number analysis.
*** Probes for CNVs are also included.

Table 2-2: Example of microarray products offered by Affymetrix and Illumina

In this thesis, genetic data used for the experiments was genotyped using Illumina HumanOmniExpress BeadChip (OmniExpress) which provides excellent power for common-variant GWAS and high sample throughput (Illumina 2010).

Public Repositories and International research

Several international research projects and public data repositories have been created to share the data and discoveries achieved using high-throughput sequencing technologies. This has led, to an explosive growth in individual genome sequencing data (Duan et al. 2016). Tens of millions of DNA variants (SNPs) have been identified in different populations. This is revolutionising our understanding of the relationships that exist between genomic variation and phenotypes, which is considered one of the main aims in biology and medicine.

International projects such as The 1000 Genomes Project (Durbin et al. 2010) and the International HapMap Project (Gibbs et al. 2003) have made it possible to investigate complex and multifactorial disorders using GWAS, which has permitted the creation of widespread catalogues for human genetic variation. The 1000 Genomes Project is a comprehensive public reference database of human genetic variation (SNPs and SVs) across multiple populations to help improve our understanding of the genetic contribution to human phenotypes. This was achieved by sequencing the genomes of at least 1,000 people (today approximately 2,000 individuals or more). The identification of SNPs, in numerous populations, has contributed to our understanding of rare and common variations and how they are distributed in the genome. This has advanced our understanding of disease biology. These variants can be used, for

example, for genetic imputation in GWAS and other human genetic studies (Gibbs et al. 2015). In a similar way, the International HapMap Project is a collaborative initiative designed to identify and catalogue common genetic variation (primarily SNPs) in different population ancestries including African, Asian and European. It reports information about where genetic variations occur across the genome and describes correlations between variants (Linkage Disequilibrium) and how they are distributed among individuals and their populations.

The European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) and the National Human Genome Research Institute (NHGRI) Catalog, funded by the NHGRI in 2008, has collected data from the literature since the first published GWAS, in compliance with eligibility criteria (MacArthur et al. 2017). This catalogue, also known as the GWAS Catalog, is a manually curated and publicly available database of SNP-trait association data discovered via genetic association studies, which is collaboratively produced and maintained by the NHGRI and EMBL-EBI. Information contained in the catalogue can be used by experts in several fields such as bioinformatics and biology among others, to baseline and conduct investigations, to better understand disease aetiologies and develop novel therapies (Welter et al. 2014).

Several institutions have invested heavily in data collection to gather clinical and genetic data within different domains. This has resulted in significant amounts of big data (Marx 2013) and today organisations, such as the National Institute of Health (NIH), which sponsors the Database of Genomes and Phenotypes (dbGaP), are making this data available to interested parties, subject

to data access agreements (Tryka et al. 2014). Researchers are required to submit an application for approval to get access to individual-level phenotype and genotype data.

In the private sector, genetic screening services can be delivered directly to consumers. Individuals provide a saliva sample to a Direct-to-Consumer Genetic Testing (DTCGT) company and obtain genetic information without any healthcare provider involvement (Su 2013). Many of these DTCGT services use SNP identification to determine ancestry and genetic markers associated with specific diseases with the objective of informing clients about their health and how to change behaviours to improve it (Su 2013). Consumers of these services often share their personal genetic data with non-profit organisations such as the Personal Genome Project (PGP) (Ball et al. 2014).

The PGP was created to promote the availability and use of personal health and genome data to accelerate the understanding of genetic variation in humans (Ball et al. 2012). While many object to privacy, confidentiality and anonymity issues, the PGP believes that sharing such data is fundamentally important to advance science and society. This is a view endorsed by members of the public who understand the risks and share their personal information. The PGP was initiated by the Harvard Personal Genome Project, which now hosts publicly shared genomic and health data from thousands of participants. In 2005 information on 10 fully identified individuals was available; today, more than 5,000 participants have shared their DNA with PGP[3].

[3] https://pgp.med.harvard.edu/about

2.3.7 Types of Studies

Polygenic obesity studies are based on the analysis of genetic variations in DNA (i.e. SNPs) located within or near genes. Identifying the genes that contribute to diseases has been one of the primary goals in human and medical genetics. Consequently, this has led to studies focused on linkage, association and candidate gene approaches (Bailey-Wilson & Wilson 2011). Such studies are conducted using family members, also known as family studies (linkage analysis), or unrelated individuals (case-control studies) to determine possible associations between a gene's allelic variation and disease traits. This section reviews traditional and modern approaches utilised to identify genetic variants associated with obesity.

2.3.7.1 Linkage Analysis

Linkage analysis is a well-stablished hypothesis-free approach conducted in related individuals and used to identify genome regions predisposed to disease (Dawn Teare & Barrett 2005). They are part of a larger process termed reverse genetics, as it starts with the trait under investigation and uses linkage analysis as well as other analytic approaches to map the predisposing genes (Cantor 2013). In a broad sense, in linkage analysis, genotypes markers are tested in a study of pedigrees samples where statistically significant markers showing linkage (by exceeding a predefined threshold) pinpoint the gene to the chromosome segment where the markers reside.

These types of studies have successfully identified highly penetrant genetic variants of large effect (very high odd ratios) in humans, which are responsible for many Mendelian diseases. However, the same level of success has not been

achieved in studies aimed at identifying genetic variants with small effect size in common diseases (Bailey-Wilson & Wilson 2011). Multigenerational pedigrees of affected individuals is often hard to collect which limits sample size and therefore the power of the study. Evidence indicating linkage with obesity related phenotypes in humans, including BMI, waist circumference and obesity, has been reported in (Snyder et al. 2004). Significant evidence of linkage on different chromosomes and genomic regions was identified in the different studies discussed, indicating, in some cases, evidence of parent-specific linkage including linkage of paternally and maternally derived alleles.

2.3.7.2 Candidate Gene Studies

Candidate gene approaches have been used since the early 1990s to explore sequence variation in relatively small-scale studies with a small number of case-control observations. Genes are selected based on previous information on the biology or pathophysiology of the disease. Therefore, candidate genes require a detailed understanding on disease aetiology. This hypothesis driven approach relies on the discovery of associations between a variant with or within the candidate gene and traits, such as obesity.

Candidate genes considered in obesity studies for BMI variance, are selected based on their roles in central or peripheral pathways responsible for energy intake and expenditure (Hinney & Giuranna 2018). Early candidate gene approaches were conducted by comparing at least one carefully selected variant located at candidate genes in cases of unrelated obese patients relative to non-obese patients in the control group (Herrera & Lindgren 2010). However, most of the genes identified in this way lacked support in replication studies across

different independent databases. This was due to the presence of false positive results derived from multiple testing in typically small and underpowered datasets. Additionally, knowledge limitations about the molecular mechanisms of common obesity in the early days of candidate gene studies made it more complicated to choose ideal candidate genes. More recent candidate gene experiments that utilise large cohorts or combined data (meta-analysis) have been performed to increase power in studies. By doing this, strong associations with variants in different genes were identified, including MC4R, PCSK1, adrenergic β3 receptor (ADRB3), endocannabinoid receptor 1 (CNR1) and brain derived neurotropic factor (BDNF) (Vimaleswaran & Loos 2010).

Nonetheless, linkage and candidate gene studies have had limited success in identifying genetic variants predisposing individuals to obesity and other comorbidities, such as type 2 diabetes. However, the advent of GWAS has revolutionised the field by accelerating and improving the detection of variants with small effect sizes that influence common traits and diseases.

2.3.7.3 Genome-Wide Association Studies (GWAS).

Genome-wide association studies (GWAS) have been used in obesity research to identify obesity related loci. GWAS are more cost effective, have greater resolution and do not require pedigree data in comparison to linkage studies. Candidate gene and GWAS have been the two major approaches utilised to detect genes implicated in body weight regulation (Hinney & Giuranna 2018). Chip-based microarray technology (Gunderson et al. 2005) has made GWAS possible, providing an unbiased approach where millions of SNPs throughout the genome can be tested for associations with a phenotype. In situations where

single GWAS are underpowered, statistical results from different independent GWAS can be combined in meta-analysis to increase the power of the study and to reduce false-positive discoveries (Evangelou & Ioannidis 2013).

Over the past decade, at least 2,400 GWAS have been conducted (Nakka et al. 2016) and the NHGRI Catalog contains over 3,300 GWAS publications and almost 60,000 unique SNP-trait associations[4]. More than 200 common genetic variants from the central nervous system, food and sensing digestion, lipid metabolism and many other biological pathways have been associated with polygenic obesity and body weight regulation (Pigeyre et al. 2016). A list of polygenic loci associated with obesity and other fat distribution traits have been provided by Pigeyre et al., (2016); Hinney & Giuranna, (2018).

GWAS aims to reveal variants at genomic loci that are associated with complex traits in the population. In these studies, a large number of genetic variants (normally 500,000 or more SNPs) are tested for associations with the phenotype (i.e. disease trait) of interest, such as obesity, diabetes or coronary artery disease among others (Visscher et al. 2012; Hardy & Singleton 2009; Fall & Ingelsson 2014; Burton et al. 2007). Strong associations do not necessarily indicate that the SNPs are causal themselves but are likely to be in linkage disequilibrium (LD) with the influential SNP, a phenomenon called indirect association (Ramachandrappa & Farooqi 2011).

The availability of chip-based microarray technology for assaying potentially millions of SNPs has favoured the introduction of GWAS. However,

[4] https://www.ebi.ac.uk/gwas/

this technology is evolving fast as discussed previously and new technologies are being introduced. New NGS methods provide a snapshot of all the DNA sequence in the genome. Powerful and unbiased approaches such as whole-genome sequencing (WGS) and whole-exome sequencing (WES) are becoming more accessible and this will allow for the detection of genetic variation within an individual. Nonetheless, sequencing the whole genome is expensive. Therefore, researchers tend to focus their efforts on protein-coding regions of the human genome known as the exome, especially in the study of rare-disease causing genes. It is estimated that 85% of disease-causing mutations with large effects are harboured by protein-coding genes, which constitute only ~1% of the human genome (Majewski et al. 2011).

2.4 Computational Analysis in GWAS

In GWAS, the phenotype under investigation can be either qualitative (often binary case/control) or quantitative (continuous). In case-control studies, binary disease traits such as obesity are investigated to identify genetic variants associated with this trait, where 0 represents controls and 1 represents cases. In contrast, the goal of quantitative trait association studies is to identify genetic factors associated with continuous traits like BMI. The most common approach in GWAS is case-control analysis, where cases refer to a cohort affected by the disease while control refers to a cohort unaffected by the disease. Conversely, quantitative phenotypes improve the power to detect genetic effects which can be more interpretable. Despite these differences, both types of phenotypes can be used to conduct successful studies (Bush & Moore 2012).

Single SNP analysis is the most commonly used approach in GWAS (Shi & Weinberg 2011). Statistical tests applied in GWAS depend on whether the phenotype is qualitative or quantitative, although in both cases each SNP is independently tested for association with the phenotype. Common approaches used to analyse quantitative traits are generalised linear models (GLM) or Analysis of Variance (ANOVA) (He et al. 2016). In case-control studies where categorical phenotypes/traits are used (i.e. obese/non-obese), chi-squared (χ^2) or contingency table-based tests in addition to logistic regression are generally adopted to test each SNP for association with the phenotype (Zeng et al. 2015).

Single nucleotide polymorphisms are commonly biallelic, assuming alleles *A* (major allele) and *a* (minor allele) and possible genotypes *AA*, *Aa* and *aa* (Lewis 2002) (See Figure 2-10).

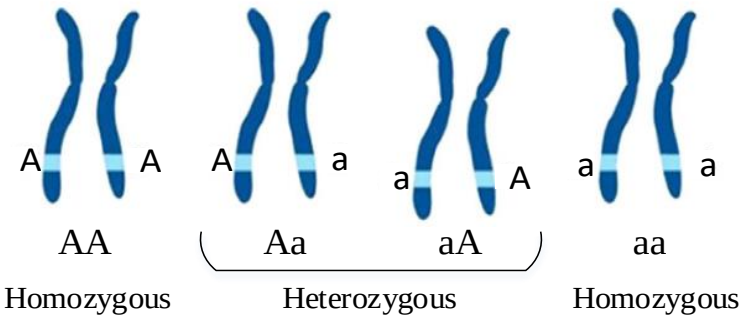


Figure 2-10: Homozygous and heterozygous values of a SNP

In GWAS, association tests can be conducted by comparing allele (allelic association test) or genotype (genotypic association test) frequency between cases and controls, where each SNP can be represented as a contingency table of counts of disease status using either allele or genotype count. In allelic association testing, disease risk increases or decreases linearly based on the number of risk alleles (minor allele versus major allele). Furthermore,

associations between one allele of the SNP and the phenotype are tested. This is in stark contrast to genotypic association testing where, genotypes and phenotype are tested. Therefore, the presence of an SNP allele may increase disease risk if there is an increased frequency of an SNP allele or genotype in cases compared with controls (Lewis 2002). Additionally, genotypic association testing can be organised in different models to encode the data for association testing: dominant, recessive, multiplicative or additive; each of them with different genetic effects in the data. Although several models are available, the additive model is often the preferred approach since in most genetic association studies the underlying genetic model is unknown, and it is easier to interpret (Sebastiani & Solovieff 2010a; Clarke et al. 2011; McCarthy et al. 2008). Considering two alleles for an SNP (*A* and *a*), the additive model assumes a linear increase of risk for each copy of the *a* allele so that the risk for *Aa/aA* is γ and the risk for *aa* is 2γ, where γ is a genetic penetrance parameter ($\gamma > 1$). A more detailed explanation for the different standard disease models available can be found in (Clarke et al. 2011), where the disease penetrance for genotypes *AA, Aa* and *aa* and associated relative risks for the *AA* and *aa* genotypes was described as shown in Table 2-3.

	Penetrance			**Relative Risk**	
Model	***AA***	***Aa***	***aa***	***Aa***	***aa***
Additive	*0*	γ	2γ	γ	2γ
Multiplicative	*0*	γ	γ^2	γ	γ^2
Recessive	*0*	*0*	γ	*1*	γ
Dominant	*0*	γ	γ	γ	γ

Table 2-3: Disease penetrance and relative risk for different genetic models (*a* is the risk allele)

A typical GWAS setup is illustrated in Table 2-4.

Study design	**Case-control study:** obese cases and non-obese controls.
Variables	Genetic markers: Biallelic SNPs with alleles ***A*** and ***a***. Possible genotypes ***AA***, ***Aa*** and ***aa*.**
Platform	Affymetrix: GeneChip Human Mapping100K Set GeneChip Human Mapping500K Set Genome-wide Human SNP Array 5.0 Genome-wide Human SNP Array 6.0 Illumina: HumanOmniExpress-12v1.0 Sentrix HumanHap300 Genotyping BeadChip Sentrix HumanHap550 Genotyping BeadChip Sentrix HumanHap650Y Genotyping BeadChip Human1M DNA Analysis BeadChip
Quality Control	Crucial step in GWAS. Removal of individuals and SNPs with unreliable data leading to spurious results.
Genetic Model	Selection of genetic model (Clarke et al. 2011): • Dominant • Recessive • Multiplicative • Additive

Table 2-4: Basic GWAS scenario example. Based on (Ziegler et al. 2008)

2.4.1 Multiple testing

In genetic experiments, such as those conducted in GWAS, a large number of hypothesis tests are performed. This implies that a large number of variants are expected to be deemed significant by chance. This problem, known as multiple comparison or multiple testing, results from statistical analysis that involves

multiple simultaneous statistical tests (Sedgwick 2014). For example, if 100 statistical tests are conducted, all with null hypothesis actually true, it is expected around 5 of the tests to be significant at a standard P-value cut-off of 0.05 ($\alpha = 0.05$) just due to chance. The probability of making at least one false positive (Type I error) when *m* tests are performed, also termed family-wise error rate (FWER), is given by Equation 2-2.

$$FWER = 1 - (1 - \alpha)^m \qquad 2\text{-}2$$

In the above example for 100 tests, Equation 2-2 indicates that, if the tests are statistically independent from each other, the probability of at least one incorrect rejection is 99.4%. Hence, estimating the significant thresholds, that is the proportion of false positives tolerated by a researcher, in studies involving a large number of genetic variants such as GWA studies is an important task that needs to be controlled for multiple testing (Clarke et al. 2011).

Associations between investigated traits and SNPs are classed as having genome-wide significance when they have a P-value $\leq 5x10^{-08}$ (Panagiotou & Ioannidis 2012; Fadista et al. 2016; Dudbridge & Gusnanto 2008). Conversely, SNPs showing suggestive associations have P-values $< 10^{-5}$ and are more likely to include additional true positive signals (SNPs). In these instances further analysis is required (Zhang et al. 2016; Below et al. 2016; Deloukas et al. 2013).

Significant values (α) can be adjusted using Bonferroni correction, a highly conservative method designed to minimize type I errors in multiple testing studies (Dudbridge & Gusnanto 2008). Bonferroni adjusts the probability of rejecting the null hypothesis when it is true by the number of statistical tests

performed from $\alpha = 0.05$ to $\alpha = (0.05/n)$, with *n* being the number of SNPs tested. The value of the confidence threshold for a single test α is equal to 0.05 as historically used in many studies (Noble 2009). Hence, Bonferroni adjustment considers a result significant only if the corresponding P-value $\leq$ 0.05/n. Although this procedure is intended to reduce the chance of false-positive findings (Type I errors), it is too conservative. This is true since many SNPs are correlated due to LD which means they are not independent. This leads to failures identifying true results, since a large number of potentially true associated SNPs are omitted. Bonferroni correction is the simplest and most widely used approach to correct for multiple testing (Noble 2009) although not the only one; Benjamini-Hochberg false discovery rate (FDR) and permutation testing are less conservative approaches to deal with the multiple testing problem in GWAS (Dudbridge & Gusnanto 2008).

2.4.2 Quality Control in GWAS

Before conducting GWAS, certain considerations must be taken into account to avoid systematic bias (Teo 2008). Among these, data quality per individual and SNP, also known as quality control (QC), relatedness among samples, genetic outliers or population stratification need to be performed.

Data quality-control (QC) is a key step taken prior to GWAS analysis (Clayton et al. 2005; McCarthy et al. 2008). QC is applied to individuals and genetic markers (SNPs) although the order of this process depends on the GWAS characteristics (Weale 2010). For example, standard QC protocols conducted in GWAS have been detailed in (Anderson et al. 2010; Laurie et al. 2010; Turner et al. 2011), where the authors recommend conducting QC on

individuals first and then SNPs to maximise the number of genetic markers for subsequent analysis. The effect of applying QC on the data typically leads to a reduction in both the number of SNPs and individuals.

The QC on individuals is typically divided into five steps: 1) sex inconsistencies, 2) high missing genotype rates or call rate, 3) high heterozygosity rates, 4) relatedness or duplicate individuals, and 5) population outliers. QC on genetic variants usually involves three steps: check 1) SNPs with elevated missing genotype rates, 2) SNPs showing substantial deviation from Hardy-Weinberg equilibrium (HWE) and 3) minor allele frequency. These steps are briefly described below and explained in detail in a number of works (Anderson et al. 2010; Laurie et al. 2010; Turner et al. 2011; Gondro et al. 2013).

2.4.2.1 Individual level quality control

2.4.2.1.1 Sex inconsistencies

In many studies, sex is often reported by subjects. Therefore, one of the first QC steps applied is to check for sex inconsistencies, where the reported sex of each individual is compared against the sex predicted based on genotype data from the X chromosome. Males and females tend to cluster differently based on X and Y chromosome intensities, where individuals annotated as males show greater Y intensity compared to those marked as females (Laurie et al. 2010).

2.4.2.1.2 Individuals with missing genotype

Individuals with low call rates (the proportion of SNPs with missing genotypes for a given individual) should be removed from further analysis since it may be an indication of poor DNA quality. That is, individuals with high rates of

missing genotype data in the typed SNPs. Recommended thresholds to prune individuals based on missing genotype data range between 95% and 99% although it can vary depending on the study (Namjou et al. 2015; Wang et al. 2011; Willer et al. 2009).

2.4.2.1.3 Heterozygosity

Another good quality indicator is the distribution of mean heterozygosity through all individuals which may indicate DNA sample contamination or inbreeding (Teo 2008). Heterozygosity occurs when an individual has different alleles at a locus whereas homozygosity implies subjects are carrying the same alleles. High heterozygosity rates within an individual may indicate poor sample quality whereas low heterozygosity levels may indicate inbreeding.

Heterozygosity rate can be inspected by estimating the mean (m) and standard deviation (SD) of the heterozygosity of all individuals and then pruning those outside the bounds $m \pm 3$ SD. The distribution of mean heterozygosity across all individuals is computed as the ratio of the number of heterozygote genotype calls (N-O) to the number of non-missing genotypes (N), where O is the observed number of homozygous genotypes for a given sample:

$$\frac{N-O}{N} \qquad \text{2-3}$$

2.4.2.1.4 Relatedness or duplicate individuals

In population-based studies limited to unrelated individuals, such as GWAS, closely related subjects should be removed due to possible correlation structure that can lead to the introduction of false positives and/or false negative results. Duplicate individuals are treated as an extreme case of relatedness. The

identification of duplicate or related individuals is typically carried out by calculating the identity by descent (IBD) metric (Weale 2010). This metric looks at the average proportion of common alleles between two individuals. Using the IBD metric individuals can be classified as duplicated or monozygotic twins, first degree relatives, second degree relatives and third degree relatives for IBD values 1, 0.5, 0.25 and 0.125 respectively. It has been recommended to remove individuals based on an empirical threshold IBD higher than 0.1875 which is set between second and third degree relatives, whilst an IBD > 0.98 identifies duplicate samples (Weale 2010). It is a common practice prior to relatedness or duplicate analysis, to apply dimensionality reduction by pruning highly correlated SNPs in regions of extended linkage disequilibrium (LD) to, for example, improve computational efficiency (Burton et al. 2007).

2.4.2.1.5 Population stratification (Divergent Ancestry)

The presence of individuals with different ethnic backgrounds (multiple populations) in the study can also lead to systematic bias (type I and type II errors) in GWAS, since allele frequency differences between cases and controls occur due to ancestral differences as opposed to effects on disease risk (Cardon & Palmer 2003). For example, if a disease is more common in a population in one region of the world than another, genetic differences between the two populations will look like they are associated with disease. This phenomenon is known as population stratification and it is an important confounder in association analysis.

Several methods have been proposed to detect and account for population stratification, including genomic control (GC) and principal component analysis

(PCA) based methods among others (Sebastiani & Solovieff 2010b; Wu et al. 2011). However, the most commonly used method is PCA (Price et al. 2006). PCA produces several principal components (uncorrelated variables) from a data matrix of observations across several potentially correlated variables (observations are individuals whereas potentially correlated variables are SNPs). This is conducted using genotype data from known ancestry populations (i.e. Europe, Asia and Africa) from HapMap (Gibbs et al. 2003), which are used to cluster GWAS individuals into distinct ancestry groups. Generally, the top principal components capture the population substructure due to ancestral differences in the GWAS data. Using PCA to identify outliers and hidden population structure can be performed using freely available software such as EIGENSTRAT (Price et al. 2006).

2.4.2.2 Genetic variant level quality control

Once individual level QC is completed, individuals can be pruned, and marker level QC can be performed. Genetic variants (SNPs) are commonly filtered and used for subsequent analyses if their minor allele frequency is greater than 1-5%, Hardy Weinberg equilibrium P-values are lower than 10^{-4}, and call rate is at least 95% across all samples (Grundberg et al. 2010). Of course, these parameters can vary from study to study.

2.4.2.2.1 SNPs with missing genotype

An important QC step conducted at genetic marker level is to inspect the proportion of missing genotypes per SNP, which is a complement of individual missingness explained above. This happens when genotypes are not assigned to SNPs in the genotyping process for many individuals. Therefore, SNPs missing

a high proportion of genotypes are excluded (SNPs with low call rates). Problematic SNPs with low call rates are often removed using a recommended 95% threshold which is equivalent to 5% missing genotype, although this may vary from study to study (Zhang et al. 2016; Turner et al. 2011). In small sample settings, more stringent thresholds can be used (Reed et al. 2015).

2.4.2.2.2 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) is a genetic principle utilised as a QC measure for the identification of systematic genotyping errors in unrelated samples (Wittke-Thompson et al. 2005). It describes the relationships between allele and genotype frequencies in a specific population. Allele and genotype frequencies should remain stable between generations in large, randomly mating, homogeneous populations. HWE assumes that any deviations from stable relationships between allele-genotype frequencies can be considered a problem in the genotyping process. To check that an SNP genotype distribution follows HWE, statistical tests can be used. Therefore, probabilities of the genotypes at a biallelic genetic variant in HWE are $(1-q)^2$, $2q(1-q)$ and q^2 for *aa*, *Aa* and *AA* respectively, for a given MAF *q*. Under the above assumption, these probabilities should remain stable over generations. When using HWE in QC processes, it is recommended to check deviation only in the control set (in case-control studies) and disregard SNPs from further analysis if the deviation test produces, for example, a P-value $< 10^{-5}$ although this value has varied significantly between studies (Anderson et al. 2010). This is conducted typically in controls as deviations in cases could be an indication of true genetic association with the disease risk.

2.4.2.2.3 Minor allele frequency

Minor allele frequency (MAF) is a common genetic variant data filtering step since statistical power is extremely low for rare SNPs (Morris & Zeggini 2010). Therefore, if the MAF is low, SNPs are typically removed. Common values for the MAF filter criterion varies between 1% and 5% depending on the sample size and study design (Himes et al. 2013; Org et al. 2009).

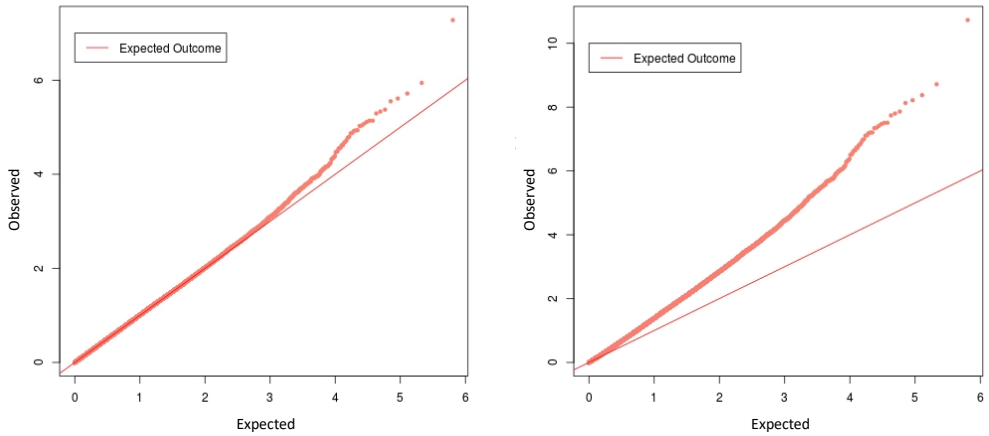
2.4.3 Association Analysis

Statistical tests to establish an association between each genetic marker and the phenotype under investigation can be conducted using several approaches as discussed earlier. Logistic regression is the most popular one used in GWAS (Zhang et al. 2016; Lewis et al. 2010; Whitaker et al. 1997; Gormley et al. 2016; Bao et al. 2016), which supports the use of covariates. In this approach, the logarithm of the odds of disease is the response variable, with linear (additive) combinations of the exploratory variables (genotype variables or any covariates) applied to the model as predictors.

The primary outcome of this statistical analysis is a list of SNPs, their corresponding position in the chromosome, and a P-value which indicates the statistical significance of the association (Turner 2014). In cases where the null hypothesis (H_0) for no association is true, large P-values for significance are expected from the association test. Whereas, small P-values mean the hypothesis must be rejected, subject to a genome-wide significant threshold (Fadista et al. 2016). In statistical hypothesis testing, type I errors (false positives) are produced when the null hypothesis is rejected when in fact it is

true. Contrariwise, type II errors (false negatives) are produced when the null hypothesis is accepted but it is actually false.

QQ-plots are a common statistical tool to demonstrate that confounders such as population structure are not present in the study (Rentería et al. 2013). In Figure 2-11 examples of QQ-plot are depicted.



a) b)

Figure 2-11: Examples of QQ-plot showing expected vs. observed [-log10(P-value)] values

The x-axis shows the expected distribution of -log10(P-values) under the null hypothesis of no association. The y-axis shows the observed -log10(P-values) in the association analysis. P-values are generally transformed by -log10 so that the smallest values near zero become the larger values and are thus easier to see. Each dot in the plot represents an observed -log10(P-values) calculated for the SNP. The default lines (expected outcome) show where x = y. A solid line in the QQ-plot matching x = y until it starts deviating at the upper-right end of the plot, represents strongly associated SNPs as shown in Figure 2-11 a). Conversely, any early deviation from the x = y line may indicate a consistent

difference between cases and controls throughout the genome, suggesting bias such as population structure (See Figure 2-11, b).

When conducting association analysis, it is also possible to apply genomic control, which assumes that the statistical test is inflated by a constant inflation factor λ, to evaluate if population structure still exists (Clarke et al. 2011). The inflation factor (λ), measures the degree of deviation from the y = x in the QQ-plot. In a homogeneous population, λ should be equal to one, although, empirically, a $\lambda < 1.05$ is considered acceptable (Zeng et al. 2015).

Manhattan plots are also used as a visual tool in association studies to visualise the P-values of association (Zeng et al. 2015). The x-axis presents the SNPs, in chromosome order and visualised using alternating colour ranges. The y-axis reports the -log10(P-value) of each SNP association as shown in Figure 2-12. The red and blue lines correspond to the significance and suggestive thresholds respectively. Hence, the smallest P-values suggest potential disease-related SNPs and typically need to reach one of the thresholds mentioned to be considered for subsequent analysis. Therefore, given that the smallest P-values produce the strongest associations, the -log10 of these P-values will have the highest height in the Manhattan plot.

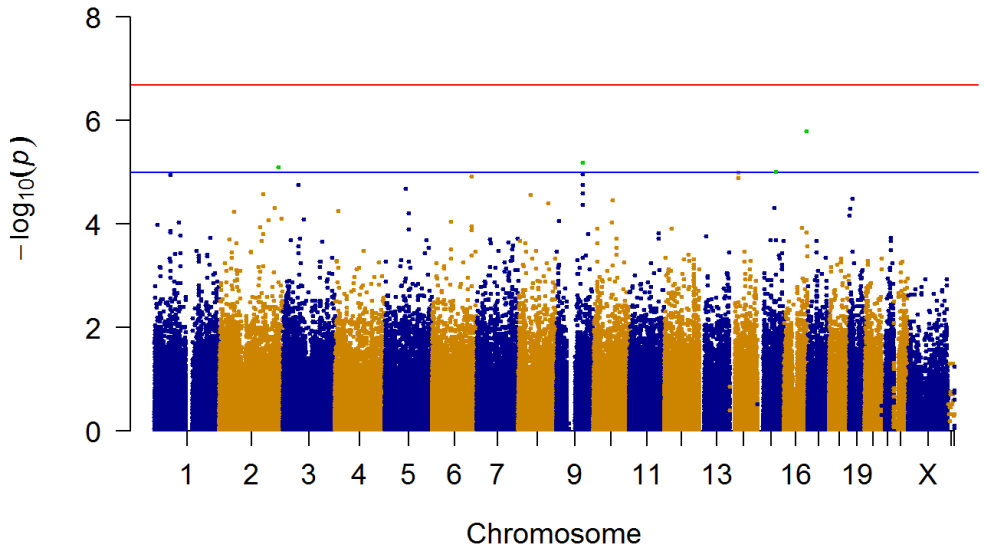


Figure 2-12: Example of Manhattan Plot. The x-axis indicates location while y-axis displays the significance of the association

2.4.4 Analytic tools for GWAS

All QC procedures and association tests conducted in GWAS, can be performed using different bioinformatics tools (Sebastiani & Solovieff 2010a). However, many researchers commonly use the well-established and computationally efficient software program for analysing genotypic data, PLINK (Purcell et al. 2007). PLINK provides a comprehensive and well-documented collection of commands and tools to conduct several data management and QC tasks, including tools for analysis, i.e. standard association analysis, IBD and sex checks. The software can be freely installed on Windows, Mac OS X and Unix machines (such as Linux) since it is open source. The main commands used in PLINK to conduct QC procedures and those used in this thesis can be found in (Purcell 2009; Purcell & Chang 2018). PLINK allows associations between SNPs and a binary outcome to be tested using the options --assoc or --logistic which perform a χ^2 test of association or logistic regression, respectively.

QC and association analysis can also be conducted using standard statistical software such as R (GNU Project n.d.), which additionally allows us to visualise the results. R is an open-source software environment for statistical computing and visualisation under a GNU-GPL licence. R compiles and runs on Windows, Mac OS X, and numerous UNIX platforms (such as Linux). R packages such as "qqman" provide a convenient and flexible way to generate Q-Q and Manhattan plots from PLINK results (Turner 2014).

2.4.4.1 Data Format

One of the most common data formats utilised when conducting GWAS in PLINK, is the linkage or pedigree file format PED/MAP (Turner et al. 2011). This white-space (space or tab) delimited text file format consists of two files: 1) PED files with extension **.ped* that describes individuals and associated genetic data. It contains one row per individual with column names family ID, individual ID, paternal ID, maternal ID, sex, phenotype and the genotypes (two columns per genotype; one for each allele). 2) The MAP file (**.map*) which contains information about the genetic markers (SNPs). The genotype columns in the **.ped* file are associated with SNPs in the **.map* file. The available columns are chromosome, SNP identifier (rs#), genetic distance and physical position.

Reading *.ped and *.map files can be time consuming. Thus, a compressed and significantly more efficient form of the pedigree file format is typically recommended to speed up analysis (Rentería et al. 2013). This compressed file format is termed the binary file (PLINK file formats BED/BIM/FAM) and is composed of three sub-files: 1) compressed binary file (*.bed) containing

genotype information, 2) a text file (*.fam) with information about the individuals (first six columns of the *.ped file) and, 3) a text file (*.bim) that contains information about the SNPs (chromosome, marker ID, genetic distance, physical position, allele 1, allele 2). In Figure 2-13, an overview of these commonly used PLINK formats is depicted.

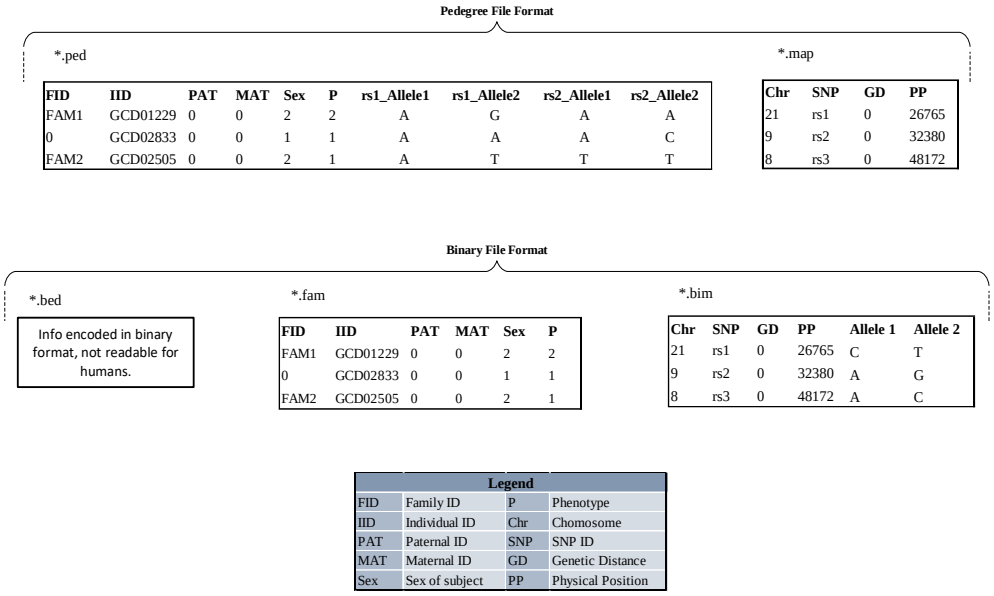


Figure 2-13: Overview of common PLINK formats utilised in GWAS

PLINK also allows us to generate a *.raw file from the binary files, which can then be loaded into the R environment to conduct further experiments. Genotype data is recoded based on the number of minor alleles. A text file is generated, using the function --recodeA (additive component file), with several columns and one row per individual. The first six columns are equivalent to the FAM file discussed above: FID, IID, PAT, MAT, SEX and P. These columns are followed by one extra column per variant with column names in the form: [Variant ID]_[counted allele], where variant ID is the SNP ID and counted allele is the

minor allele. Therefore, given *a* as a minor allele for a locus, the number of alleles is 0 if genotype is *AA*, 1 if genotype is *Aa/aA* and 2 if genotype is *aa*.

In Figure 2-14, an example of the output file (*.raw) produced by the --recodeA command in PLINK is shown. The example contains data about three samples and two SNPs per sample recoded in terms of additive components.

FID	IID	PAT	MAT	Sex	P	rs1_G	rs2_T
FAM1	GCD02632	0	0	1	2	0	2
FAM2	GCD03035	0	0	2	2	1	0
FAM3	GCD01227	0	0	1	1	1	0

Figure 2-14: Example of *.raw file produced in PLINK with the command --recodeA

2.5 Key findings in GWAS for obesity

Through the course of an initial literature search, a series of studies on obesity related GWAS have to date been found. Although the number of examples is not representative of the application domain as a whole, the initial results obtained represent a useful start.

The study of obesity has benefited from a series of discoveries derived from at least four waves of large scale high-density GWAS to date (Vimaleswaran & Loos 2010; Loos 2012). These GWAS have identified common variants in the fat mass and obesity-associated gene (FTO) - the first gene associated with polygenic obesity. This was confirmed in various age groups and ancestry populations (Loos & Yeo 2014). Extensive evidence about the effect FTO has on body weight regulation in humans and rodent models has been reported. The evidence has been carefully reviewed by Speakman (Speakman 2015) who highlighted some of the major findings. In addition to the FTO gene, common

variants near the MC4R gene have also been associated with obesity risk (Loos et al. 2008). More recent meta-analysis of ~340,000 individuals reported 97 Genome-wide Significance (GWS) polymorphisms associated with BMI, of which 56 were novel (Locke et al. 2015). The Venn-diagram in Figure 2-15 extended from (Loos 2012), depicts obesity susceptibility genes discovered in different waves of GWAS and meta-analysis for waist-to-hip ratio, waist circumference, body fat percentage, extreme and early onset obesity, BMI and visceral adiposity.

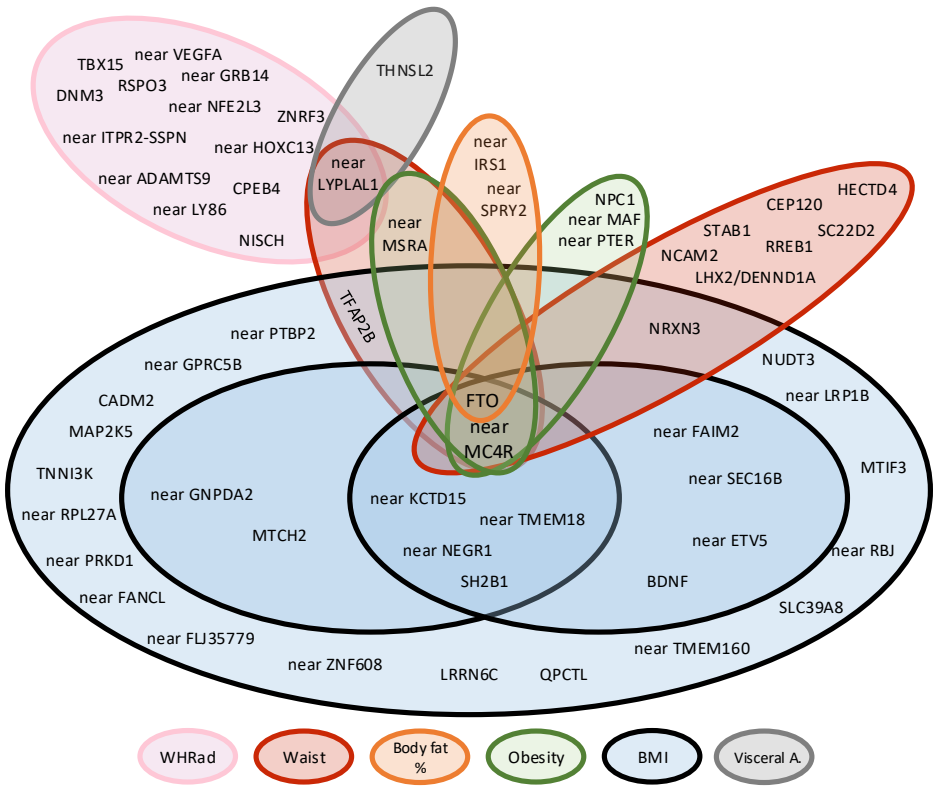


Figure 2-15: Obesity-susceptibility genes discovered in different waves of GWAS

For a more comprehensive view of genome-wide studies of obesity, the most relevant studies where these genes were discovered are summarised in Table 2-5.

Study Design	Ethnicity	Traits	Replicated genes	References
Type 2 diabetes case-series, case- series and population-based series. Adults and children	White Europeans from the UK, Italy	BMI and obesity	FTO	(Frayling et al. 2007)
Family studies. Adults.	European Americans, Hispanic Americans, and African Americans	BMI, hip and weight	FTO	(Scuteri et al. 2007)
Population-based studies, case series, obesity case-control studies, family study. Adults and children.	White Europeans from the UK, Italy, Germany, USA, Sweden and Finland	BMI and obesity	FTO , near MC4R	(Loos et al. 2008)
Population-based studies. Adults.	Indian Asians and white Europeans from the UK	BMI, waist circumference and waist-hip ratio, weight, and other metabolic traits	near MC4R	(Chambers et al. 2008)
Population-based studies, case-series, and obesity case-control studies. Adults and children.	Whites from the UK, USA, Sweden, Finland, Italy, and The Netherlands	BMI and obesity	FTO , near MC4R, NEGR1, near TMEM18, near KCTD15, SH2B1, near GNPDA2, MTCH2	(Willer et al. 2009)
GIANT meta-analysis and population-based study. Adults.	Whites from the UK, USA, Sweden, Finland, Italy, The Netherlands and Denmark.	BMI and weight	FTO , near MC4R, NEGR1, near TMEM18, near KCTD15, SH2B1, SEC16B, near ETV5 & DGKG, BDNF, near BCDIN3D & FAIM2	(Thorleifsson et al. 2009)
Population-based studies, case series, and obesity case-control studies. Adults and children.	Whites from the UK, USA, Sweden, Finland, Italy, The Netherlands, Australia, Estonia, Germany, France, Norway	BMI and obesity	FTO , near MC4R, NEGR1, near TMEM18, near KCTD15, SH2B1, near GNPDA2, MTCH2, SEC16B, near ETV5 & DGKG, BDNF, near BCDIN3D & FAIM2, TFAP2B, NRXN3, near RBJ & POMC, near GPRC5B, MAP2K5, QPCTL & near GIPR, TNNI3K, SLC39A8, near FLJ35779, LRRN6C, near TMEM160, near FANCL, CADM2, near PRKD1, near LRP1B, near PTBP2, MTIF3, near	(Speliotes et al. 2010)

			ZNF608, near RPL27A & TUB, NUDT3	
Population-based studies, case series, and obesity case-control studies. Adults.	Whites from the UK, USA, Sweden, Finland, Germany and The Netherlands	Waist circumference and WHR	FTO , near MC4R, TFAP2B, near MSRA, near LYPLAL1	(Lindgren et al. 2009)
GIANT meta-analysis. Adults.	Whites from the UK, USA, Sweden, Finland, Italy, Switzerland, Germany, and Iceland	Waist circumference	FTO , near MC4R, NRXN3	(Heard-Costa et al. 2009)
Population-based studies, case series, and obesity case-control studies. Adults.	Whites from the UK, USA, Sweden, Finland, Italy, The Netherlands, Australia, Estonia, Germany, and France	WHR adjusted for BMI	RSPO3, near VEGFA, TBX15, near NFE2L3, near GRB14, near LYPLAL1, DNM3, near DNM3 & SSPN, near LY86, near HOXC13, near ADAMTS9, ZNRF3, NISCH, CPEB4	(Heid 2005)
Obesity family study with at least one extremely obese child or adolescent. Adults and children.	White Europeans from Germany	Extreme obesity	FTO	(Hinney et al. 2007)
Four case-control studies and two population-based studies. Adults and children.	White Europeans from France, Germany, Finland, and Switzerland	Obesity and BMI	FTO, near MC4R, NPC1, near MAF, near PTER	(Meyre et al. 2009)
Two case-control studies and seven population-based studies. Adults and children.	White Europeans from Germany and USA	Obesity and BMI	FTO, near MC4R, near MSRA	(Scherag et al. 2010)
Large scale population-based study. Adults.	People living in Norwich, UK (The EPIC-Norfolk study)	Obesity and BMI	FTO, near TMEM18, MTCH2, SH2B1	(Li, Zhao, Luan, Ekelund, et al. 2010)
Meta-analysis. Population and family-based versus case-control. Adults and children.	White, African American, Asian, Hispanic from Europe, North America and Asia	BMI, waist circumference and Body fat percentage	FTO	(Kilpeläinen et al. 2011)

Table 2-5: Summary of large-scale high-density genome-wide association studies for obesity related traits

Wang et al. (2011) reported that they found strong links between the FTO gene and obesity as well as other important findings through a GWAS, where 16 genome-wide significance signals were found within the FTO gene. The authors used 520 cases and 540 control subjects of non-Hispanic Caucasian ancestry and performed a GWAS on obesity as a binary trait. Obese cases, families and never-overweight controls were evaluated to perform association analysis on obesity and multiple quantitative phenotype measures. When comparing their results with respect to previously reported associations with obesity related traits, the authors highlighted the strong effect size of the genes FTO and MC4R. The results justify the identification of these genes in GWAS for BMI in previous studies. Therefore, the results suggest that FTO and MC4R are the main two genes that have a direct effect on obesity. New candidate genes for obesity-related traits were also identified, with special interest on the association of the Neurexin 3 (NRXN3) gene with body fat distribution in extremely obese individuals. The results of the study revealed that FTO and MC4R might only be the two main genes for common obesity variants in populations of European ancestry. Despite the association of NRXN3 with body fat distribution, this gene has been associated with many other traits. Identifying the specific causal SNPs may be complicated as NRXN3 is an extremely large gene, composed of ~1.5 Mb (Million bases). To conduct the association test between SNP genotypes and specific phenotypes of interest, the authors applied standard linear regression using PLINK (Purcell et al. 2007).

In Xi et al., (2011) the authors investigated whether sedentary behaviour and physical activity contribute to an association between SNPs and obesity risk in Chinese children from Beijing. The authors selected, from recent publications

and known databases, 6 SNP candidates associated with obesity risk among white populations. In this study, 1,229 obese and 1,619 normal-weight children identified as cases and controls respectively, were selected. The BMI for each child was calculated and a blood sample was collected for genotyping. A validated questionnaire was used to determine their sedentary behaviour and physical activity level. A multiplicative genetic model was used to compare children with risk alleles and children with non-risk alleles. Multivariate unconditional logistic regression models were used in the research to show the association between the 6 selected SNPs and obesity risk. The results were modulated by sedentary behaviour and physical activity, thereby serving as a possible prevention strategy. However, the authors concluded that more studies are required to further identify gene-environment interactions in childhood obesity.

In Ahmad et al., (2013) the authors replicated the findings produced in Li et al., (2010) which showed that the performance of physical activity outweighed the genetic risks of 12 loci responsible for weight gain or loss in individuals (obesogenic loci). These 12 loci were identified in previous GWAS, where they were strongly associated with increased BMI. The number of participants was 111,421 of European ancestry. In this study, only physical activity data was collected using self-administered questionnaires. The authors used general linear models to test the association of a Genetic Risk Score (GRS) with BMI, and logistic regression to verify genetic associations with obesity.

In the study conducted in Zhu et al., (2014) the authors investigated whether loci related to BMI were associated with traits linked to adiposity, and obesity

in Chinese Hans (largest Chinese ethnic group). They studied whether these associations were modified by performing physical activity, similar to the studies conducted by Xi et al., (2011);Ahmad et al., (2013); and Li et al., (2010), although with a considerably lower number of samples. The main objective was to replicate recent large-scale GWAS, predominantly, in populations of European descent and multiple loci associated with BMI. The authors focused on 36 of the 60 obesity-related well-established SNPs, according to their research. These 36 loci were identified in GWAS for BMI in both European and East Asian ancestry populations. However, only 28 SNPs were genotyped because 8 SNPs were monomorphic in Chinese Hans. Individually, 26 of the 28 SNPs showed a strong link with BMI, and the association of four loci reached nominal significance. The observation suggested that obesity-susceptibility loci on BMI tend to be lower in Han Chinese than in European ancestry. Physical activity attenuated genetic predisposition to increase BMI in Han Chinese.

Several studies have used multiple candidate genetic variants from GWAS to test their predictive capacity in complex diseases (Manolio 2010). In the case of obesity, SNPs associated with BMI and obesity have been combined into a genetic risk score (GRS) (Cooke Bailey & Igo 2016) which represents the number of risk alleles of the candidate SNPs selected. Hence, the genetic susceptibility of becoming obese increases with high GRS values. GRS has provided a measure of genetic predisposition to obesity in several studies (Belsky et al. 2013; Hung et al. 2015; Locke et al. 2015; Morandi et al. 2012) utilising models with 97, 56, 32 or less BMI/obesity associated SNPs and other predictors such as age and sex.

Despite all the findings provided by GWAS, identified SNPs only explain a small fraction of the total variation in BMI, although evidence from twin and family studies have shown higher heritability as discussed earlier in this chapter. This mismatch is a phenomenon known as “missing heritability” (Manolio et al. 2009). Furthermore, previously identified SNPs predisposing to obesity have shown poor predictive capability when compared with traditional obesity risk factors such as family history and childhood obesity (Loos 2012). Nonetheless, some of these loci are currently being used in direct to consumer (DTC) personal genomic profiles to estimate the risk of obesity in the lives of individuals (Loos 2012).

The discovered risk variants predisposing subjects to overweight and obesity identified in GWAS only explain a modest proportion of the genetic basis of obesity. This phenomenon is observed not only in obesity but in practically every complex disease analysed by GWAS, with some exceptions such as age-related macular degeneration (AMD) (Klein 2005). Thus, several explanations for ‘missing heritability’ have been proposed (Eichler et al. 2010). These explanations include unaccounted effect of structural variants (deletions, duplications and inversions), the presence of rare variants not identified by previous GWAS or linkage approaches but likely to be detectable by NGS, and the effect of gene-gene interactions as well as gene-environment interactions. It has been suggested by many researchers that the contribution of many genes and mutations, each of them with individual small effects, results in low detection power in most studies conducted, but with a larger collective effect on the phenotype (Visscher et al. 2012). Gene-gene interactions or epistasis is one of the most plausible hypotheses for the ‘missing heritability’ phenomenon

(Manolio et al. 2009; Eichler et al. 2010) and will be the central genetic aspect investigated further in this thesis.

2.6 Functional GWAS

So far in this thesis, the importance of GWAS when identifying genetic variants that are associated with human complex traits has been highlighted, although limitations about this approach when explaining the missing heritability remains an issue. In addition to the aforementioned constraint, the effect of genetic variants identified through GWAS in the genes or DNA functional elements remain largely unknown, so that determining possible causal variants is still challenging. This is true since a large proportion of GWAS hits (~90%) are located in intergenic or non-coding regions (Guo et al. 2018). Therefore, it is believed that most risk variants identified by GWAS regulate the expression of genes.

This lack of certainty about the effect of the SNPs on the causal variants makes it necessary to incorporate additional information for the interpretation and empirical validation of GWAS results. The past few years have witnessed the development of important contributions, particularly in the context of genome variation. In this sense, gene set enrichment analysis (Subramanian et al. 2005), pathway analysis (García-Campos et al. 2015) and, integration of different types of biological information such as expression quantitative trait loci (eQTL) (Westra & Franke 2014) have been used to provide functional interpretation of many trait-associated SNPs in a biological context. This has opened opportunities for characterising functional sequence variation while improving understanding of basic processes of gene regulation and

interpretation of GWAS. Therefore, an essential task to systematically disentangle the molecular mechanisms underlying complex diseases, is via the identification of complex interplays among multiple genes in a genome-wide context, using functional enrichment analysis and functional annotation tools.

2.6.1 Gene Ontology, Enrichment Analysis, and Pathway Analysis

Gene ontology (GO) refers to a controlled vocabulary term used to describe the characteristics of genes based on their function and location, intended to unify gene attributes across all species (Ashburner et al. 2000). The aim is to provide a systematic description of biological features of genes to facilitate integration, retrieval and computation of data, to be used by the community in gene annotation. GO describes gene products in three structured ontologies: biological process (BP), cellular component (CC), and molecular function (MF); which refer to the biological target of the gene or gene product, biochemical activity of gene product and location in the cell targeted by the gene product respectively (Ashburner et al. 2000).

GO can be used for gene annotation processes where GO terms are assigned to gene products. Although this process provides detailed information of a gene product, GO terms not necessarily provide detailed insights into the mechanisms of expression changes for a particular disease (Zhou et al. 2017). To further explore the potential molecular basis of the disease under investigation, enrichment analysis can be used. Enrichment analysis rely on the fact that, for a given study, the chance of a gene set to be selected increases if its underlying biological process is abnormal. To do that, statistical methods are used to identify significantly enriched genes (Huang et al. 2009). Furthermore,

pathway analysis extend enrichment analysis by looking into what pathways the enrichment genes are involved with, minimising, therefore, the complexity of the analysis while providing a good approach for experimental validation (Yoon et al. 2018).

Pathway-based approaches target a predetermined gene set (aggregation of genes or SNPs) contained in a functional unit as defined by prior biological knowledge databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa & Goto 2000) and the Reactome Pathway Knowledbase (Joshi-Tope 2004) among others. Several methods to investigate pathway analysis can be used, including over-representation analysis (ORA), gene set analysis, statistical methods such as PCA or regression, and topology-based analysis. In these methods, discussed by Jin et al. (2014) in more details, information from multiple genetic loci is combined with known pathways to evaluate the association with a phenotype. More specifically, these bioinformatics methods detect enriched or over-represented gene sets (from a list of selected genes) that are functionally related according to current biological knowledge. This relationship is typically stablished based on Gene Ontology (GO) terms, pathways or a common link like a disease (Dennis et al. 2003).

Over the past decade, a plethora of software and web tools utilising the abovementioned databases to conduct functional-based analysis on microarray and GWAS data have been developed. Examples of these tools are DAVID (Dennis et al. 2003), Reactome (Fabregat, Jupe, et al. 2018), INRICH (P. H. Lee et al. 2012) and *i*-GSE4GWAS (Zhang et al. 2010) although many others have

been listed by (Jin et al. 2014). Although pathway-based analysis has served as a valuable tool for revealing functionalities in complex diseases, such approaches depend on pre-existing knowledge from gene or pathway databases. Consequently, knowledge-guide analysis should be interpreted with caution as our knowledge about existing genes or biological pathways remains incomplete (Khatri et al. 2012).

2.6.2 Expression Quantitative Trait Loci (eQTL)

Another approach utilised to understand the mechanisms underlying GWAS hits is to test whether GWAS signals are enriched with eQTL in specific tissues (Nicolae et al. 2010). In other words, eQTLs mapping methods are used to investigate the effects of SNPs on gene expression levels, proximally or distally to the gene. Therefore, genetic variants that explain a fraction of the genetic variance of a gene expression phenotype are known as eQTLs. When the tissue of expression (i.e. adipose tissue) is relevant to the disease/trait under investigation (i.e. obesity), incorporating eQTL analysis with GWAS can be used to discover genes and pathways that, when altered, are likely to cause the disease, providing, thus, disease candidate genes (Nica & Dermitzakis 2008). Project initiatives aiming at mapping regulatory annotations and connections in disease-relevant tissues such as the Encyclopedia of DNA Elements (ENCODE) (ENCODE Project Consortium 2012), Genotype-Tissue Expression (GTEx) (Lonsdale et al. 2013) and the NIH Roadmap Epigenomics Consortium (Kundaje et al. 2015), have enormously contributed to interpret non-coding variants responsible for most GWAS risk alleles identified so far. Moreover, integrative analysis tools such as Sherlock (He et al. 2013), PrediXcan

(Gamazon et al. 2015) and other similar approaches based on probabilistic assessment (Wen et al. 2017), are available to provide co-localisation of eQTL signals with GWAS results.

Depending on whether the effect of eQTLs are local or distant, regulatory variants can be classified into *cis-eQTLs* or *trans-eQTLs* respectively (Edwards et al. 2013). Most of identified eQTL are cis-acting, indicating that most of the regulatory control occur in the vicinity of genes (within 1Mb), have relatively large effect sizes and can usually be detected with small sample sizes. Conversely, *trans-eQTL* regulate genes located further away or even in a different chromosome and their effect size is usually small, so large sample sizes are necessary to detect them. Despite *trans-eQTL* individual small effects, their collective importance in the variation of gene expression has been reported to be relevant to explain heritability of gene expression (He et al. 2013).

The identification of target genes of regulatory variants (-cis and -trans) plays an important role to understand the processes by which SNPs act. However, the identification of eQTL to predict target genes typically provides indirect evidences of an association, making experimental approaches necessary to confirm their mechanistic relevance (Edwards et al. 2013). Unlike rare mutations, achieving a definitive proof of causality for an association is questionable (Chakravarti et al. 2013). This theory is further supported by the fact that causal variants are not necessarily single SNPs acting in isolation but a combination of them, so epistasis may be required to better explain complex diseases.

2.7 Epistasis in Complex Diseases

In the bioinformatics field, numerous efforts have been made to understand how genetic changes in the DNA give rise to molecular effects that cause diseases and phenotypes (Fernald et al. 2011). This has resulted in an exponential growth in our knowledge over the past decade of genetic variants associated with common diseases and traits as discussed earlier in this chapter. GWAS have been useful in this endeavour and have become the state-of-the-art technique for achieving this. However, the limitation of GWAS is its inability to explain the intricate relationships between SNPs and associated phenotypes. Interactions between genetic variants increase the computational complexity required to process them and generate large models and search spaces. This leads to what is known as the "curse of dimensionality" (De los Campos et al. 2010; Altman & Krzywinski 2018).

Research in machine learning and data mining is underway to try and overcome these challenges (Niel et al. 2015). Machine learning uses algorithms to 'learn' features or patterns in training data to solve problems and enable predictions about outcomes in unseen data (Deo 2015; Domingos 2012). More importantly, specific types of machine learning algorithms are capable of detecting the non-linear interactions in genome-wide datasets, which is not easy to achieve using traditional statistical methods (McKinney et al. 2006). Intelligent systems can therefore, process genetic data at a much deeper level to allow rich information structures to be leveraged to help improve phenotype-genotype relationship mappings.

The application of machine learning in bioinformatics has risen in popularity in studies focused on disease prediction, epistasis, and diagnosis and survival analysis (Kourou et al. 2015; Cole et al. 2017). Although there is no perfect method to detect epistatic interactions, it has been suggested that the integration of several machine learning methods could form an efficient framework (Koo et al. 2013). This could ultimately result in more efficiency and better interpretation of genetic and machine learning models. Additionally, once associations between SNPs and the phenotype have been established, classification analysis can be conducted using machine learning techniques to test the predictive capabilities of identified interactions. This will provide a tool to validate and measure the relative importance of genetic feature combinations (Kruppa et al. 2012). Building on these ideas, this chapter discusses what epistasis is and provides a detailed account of current research works on epistatic analysis.

2.7.1 Epistasis

Advances made in GWAS have served to improve our knowledge and understanding of disease genetics. As previously discussed, GWAS are based on single-loci analysis where each SNP is independently tested for association with the phenotype of interest, without considering the interactions that take place between loci. This is regarded as a significant limitation in GWAS, particularly when studying complex disorders that rely on an understanding of gene-gene and gene-environment interactions (Moore & Williams 2009). In BMI and obesity GWAS, gene-gene interactions have received little attention (Wei et al. 2012). It has been suggested that to explain the hidden genetic

variation (missing heritability) in GWAS it is necessary to examine epistasis alongside single SNP-phenotype interactions (Wei et al. 2014; Maher 2008). This approach assumes that genes do not work independently but create "gene networks" with major effects on the tested phenotype. Hence, identifying epistatic interactions can help to understand biological mechanisms and predict complex traits from genotype data.

Combinatorial effects between SNPs/genes are termed epistatic interactions or epistasis (Phillips 1998; Phillips 2008). Different perspectives exist: biological (or functional epistasis) and statistical epistasis (Niel et al. 2015). Biological epistasis involves physical interactions at the molecular level between two or more proteins (and other biological components) whilst statistical epistasis measures the average effect of allele substitution in a given population (genetic variations instead of biological molecules). When using computational methods for the detection of epistatic interactions, Fisher's (1918) definition of statistical epistasis is considered (Niel et al. 2015; Phillips 2008). It refers to the deviation of combinations of alleles at different loci regarding their total contribution to the phenotype.

Conducting experiments for biological epistasis is challenging, more expensive and the interpretation of the interactions is frequently less obvious. However, unlike statistical epistasis, it provides evidence for physical interacting molecules. Statistical epistasis on the other hand, is generally based on genetic variations where the associations do not provide evidence about corresponding physically interacting molecules. Despite the aforementioned limitations, statistical epistasis provides a suitable strategy to discover new

pathways previously ignored. This provides a foundation for new discoveries and testable hypotheses (Ebbert et al. 2015). Hence, by identifying epistasis, gene functions can be recognized, pathways can be identified, and potential drug targets can be discovered.

2.7.2 Epistatic approaches

Epistatic analysis is however computationally and statistically challenging. This is in part, due to the high dimensionality of the data. Investigating all combinations between SNPs in genome-wide studies is computationally very difficult since the number of tests and time necessary to perform exhaustive search increases exponentially with the order of interactions considered (Uppu et al. 2017). While the overall complexity is linear with the number of individuals in the study population, it becomes exponential when the order of the interactions increases. In Figure 2-16 an example of the number of possible pair-wise and three-way interactions between SNPs is provided, where the number of interactions grows exponentially. Consider half a million SNPs (n = 500,000), testing all the combinations between two variant interactions (k = 2) produces 124,999,750,000 possible pairwise interactions whilst combinations between three variants interactions (k = 3) produces 2.08×10^{16} possible three-way interactions as derived from Equation 2-4 (Cole et al. 2017).

$$nC_k = \frac{n!}{(n-k)!k!} \qquad 2\text{-}4$$

The number of combinations to be tested for models that consider more than pair-wise combinations would lead to computational burden (Ritchie 2015).

Therefore, epistatic analysis has primarily been restricted to two locus interactions. In addition to computational and statistical challenges, epistasis also poses challenges to generalisability of genotype-phenotype results, which limit replication and meta-analysis studies (Cole et al. 2017).

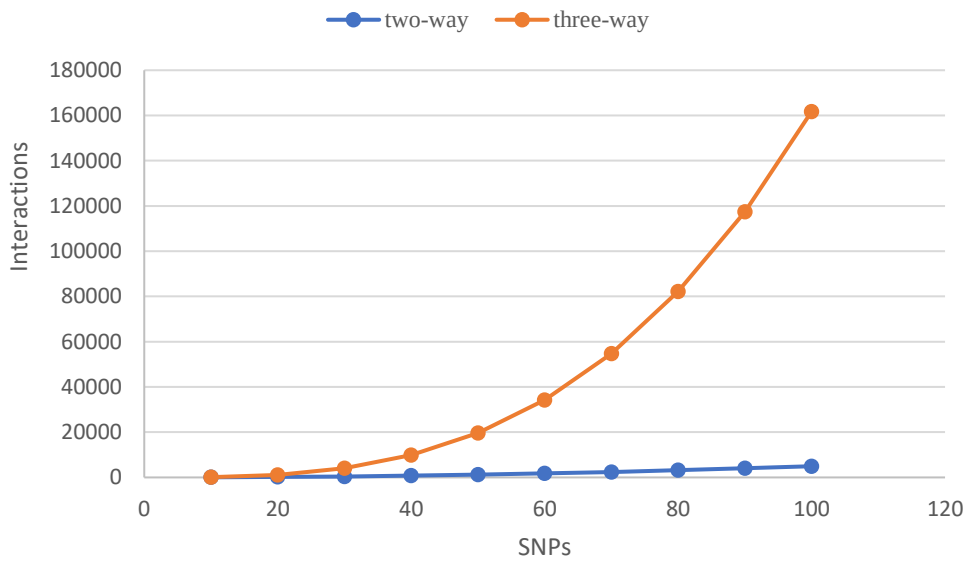


Figure 2-16: Number of possible two and three-way interactions to test for epistasis

Dimensionality reduction in epistasis

Different strategies have been proposed to alleviate computational constraints by restricting epistasis to small subsets of candidate markers. These include statistical filtering (filter approach), intrinsic filtering (wrapper approach) and extrinsic filtering based on biological knowledge (Niel et al. 2015). These techniques help to modify the data representation space, facilitating the detection of non-linear interactions among all remaining variables.

Performing SNP selection based on arbitrary significance threshold (i.e. some predefined P-value) can help to reduce computational complexity by calculating a test statistic for each marker separately and evaluating all possible

interactions in the filtered subset (Hoh et al. 2000; Marchini et al. 2005). In this approach, the data is processed statistically to assess the quality or relevance of each SNP with an associated phenotype, which can then be evaluated using classification techniques.

The second strategy uses intrinsic knowledge to extract SNPs from the dataset and thus reduce the search space for epistatic analysis. Subsets of features are iteratively selected for classification using either a deterministic or stochastic algorithm. Popular dimensionality reduction algorithms such as Relief and variations of it such as ReliefF (Moore & White 2007) learn informative features from the dataset without any a priori knowledge. The algorithms use a nearest neighbour approach to assess the quality of SNPs according to how well they distinguish individuals sharing the same phenotype. The principal difference between statistical filtering and intrinsic filtering is that the classifier plays no role when selecting which features to consider.

Finally, the third filtering approach relies on extrinsic biological knowledge from external databases to filter SNPs relevant to the phenotype of interest and then evaluate multi-SNP combinations. Expert knowledge about protein-protein interactions, Gene Oncology (GO) or common disease pathways from public databases may be used as biological filters. This will facilitate the reduction of SNPs to a list of variants located in genes that encode for proteins involved in relevant interactions. A limitation in this approach, however, is the reliance on pre-existing knowledge from literature and databases, which can prevent novel interactions between genes being discovered.

Statistical approaches for epistasis

Several statistical approaches have been proposed for the detection of epistasis (Howard et al. 2014; Niel et al. 2015). While regression analysis is extensively used to model pairwise SNP interactions, several machine learning approaches such as multifactor dimensionality reduction (MDR) (Gola et al. 2016), support vector machine (SVM) (Chen et al. 2008), neural networks (NNs) (Günther et al. 2009) and Random Forest (RF) (Lunetta et al. 2004), among others have been frequently used in several studies. Many of these methods are parametric and rely on large scale properties to accurately estimate the parameters in the model (Gilbert-Diamond & Moore 2011), whilst nonparametric methods have raised interest due to their ability to generate predictions and reveal the relative importance of genetic feature combinations (Howard et al. 2014). Among these, MDR and logistic regression are the most common non-parametric and parametric approaches currently used respectively (Ebbert et al. 2015). Parametric approaches find it difficult to detect epistasis in the absence of the main effects in the disease, whereas nonparametric methods are unsuccessful when the main effects are present (Günther et al. 2009). Thus, there is no single approach suitable for all type of epistasis.

In the past decade, machine learning approaches such as MDR have been specifically designed to detect epistasis (Gola et al. 2016; Ritchie et al. 2001). MDR is a nonparametric feature extraction algorithm with an important research contribution in the detection of epistasis. MDR methods commonly explore relationships between binary phenotypes and a combination of genotypes among a set of genetic variants (SNPs). It performs exhaustive search

among the SNP interactions and transforms the combinations into new one-dimensional multifactorial classes. MDR exhaustively considers every possible combination of SNPs to a predefined depth. However, there is an exponential correlation between the number of SNPs considered and the computational complexity as previously described. Consequently, studies using this approach are constrained to several hundred SNPs to manage computational overheads. Therefore, successful epistatic analysis currently depends on filtering approaches.

In other approaches, regression has been utilised to model pairwise epistasis in PLINK (Purcell et al. 2007). However, due to the small sample size compared with genome-wide data sizes (small *n* large *p*), parameter estimation is costly and it introduces large standard errors, making it difficult to handle genome-wide datasets (Ritchie et al. 2001). Alternative regression methods such as the least absolute shrinkage and selection operator (LASSO) or smoothly clipped absolute deviation (SCAD) have also been utilised to detect SNP-SNP interactions. Nonetheless, these techniques also suffered from an elevated false positive rate and are constrained to pairwise epistasis analysis (Niel et al. 2015).

LAMPLINK is an extension of PLINK with options to identify high order epistasis (Terada et al. 2016). It performs case-control analysis for genome-wide data using Fisher's exact test or chi-squared test. The goal is to find statistically significant combinations associated with a phenotype under investigation. High order interactions are detected using the Limitless Arity Multiple-testing Procedure (LAMP) (Terada et al. 2013). Although this approach can detect statistically valid high-order interactions from a reasonably high number of

SNPs (tens of thousands), it only supports dominant and recessive models (i.e. additive methods are not considered) which limits solving the problem of missing heritability as stated by Terada et al., (2016). Furthermore, statistical models such as regression models are not contemplated in LAMPLINK (only chi-square and Fisher's exact tests are included) although the authors do report that it will be included in future work as well as extending the number of genetic models to be considered (Terada et al. 2016).

Machine learning algorithms such as NNs, RF and Cellular Automata (CAs), can be integrated within feature extraction methods like MDR to detect epistasis (McKinney et al. 2006). In this sense, the discovery of epistasis can be conducted by applying a multi-step framework combining different parametric and non-parametric techniques as proposed by Moore et al. (Moore et al. 2006). The multi-step approach can be summarised in four steps: 1) Filter a subset of interesting SNPs from genome-wide data, 2) model epistasis, 3) use the attributes capturing epistasis for classification and, 4) facilitate interpretation of the ML models.

Depending on how information is extracted from the data, ML models can be divided into supervised or unsupervised learning, although semi-supervised learning is also considered (Iniesta et al. 2016). Supervised methods usually need labelled data to search for the optimal model weights. Typically, an algorithm is built using a dataset of candidate predictors or features as input, capable of estimating a specific outcome. Supervised learning includes classification and regression problems. Conversely, unsupervised learning is a data-driven method trained using unlabelled data with no predefined outcome

to predict. Unsupervised learning is frequently used for clustering, feature extraction or dimensionality reduction. It is common sometimes to combine an initial training procedure for NN utilisation to identify the most relevant features and then employ those features for classification via a supervised learning step (Ravi et al. 2017).

2.8 Association rule mining

In the field of data mining, association rule mining (ARM) is an unsupervised learning method used to help find and describe relationships between items (variables) that often co-occur in large datasets (Agrawal et al. 1993). The discovery of association rules depends fundamentally on the discovery of frequent itemsets (sets of items), where association rules are required to satisfy minimum support and confidence constraints at the same time. In addition to the extraction of patterns, the approach relies on how patterns are subsequently ranked and filtered.

Most itemset-based mining methods are a variant of the algorithm *Apriori* (Agrawal & Srikant 1994), which was originally intended to assist in the design of product display layouts in supermarket data mining. The algorithm states that if an itemset is not frequent, subsequent supersets of these items will also not be considered frequent. Rule mining was originally introduced by Agrawal et al. (1993) to explore several aspects of the database mining problem (Agrawal et al. 1993) although it has been broadened to solve problems in other domains such as bioinformatics (Naulaerts et al. 2015). This approach has also been used alongside statistical measures to discover binding cores in protein-DNA binding (Man-Hon Wong et al. 2015), associations between the regulation of gene

expression levels and phenotypic variations in gene expression analysis (Chen et al. 2015), epistasis (Ma et al. 2010; Zhang et al. 2014) and mining electronic health records (Li et al. 2013). ARM overcomes the limitations of machine learning approaches such as SVM and NN where the underlying models are not interpretable. Hence, ARM is more transparent, providing knowledge-based explanative rules and is thus regarded as a "white-box" approach (Naulaerts et al. 2015).

An application of the *Apriori* algorithm, in the context of case-control association studies and epistasis analysis, is AprioriGWAS (Zhang et al. 2014). This tool was applied to AMD data and bipolar disorder (BD) with promising interactions between sensible genes being found. The approach proposed by (Zhang et al. 2014) uses frequent itemset mining (FIM) with *Apriori* to look for genotype patterns with different frequencies in cases and controls. Several parameters are set by the user (i.e. minimal support). Depending on the parameter settings, this influences the performance of the algorithm and affects the candidate search space, speed and power to detect patterns. To assess which patterns should be retained, the authors utilised a technique known as proportion test (Peter Armitage, Geoffrey Berry 2001). This is followed by Pearson's chi-square test to detect interactions between variants. In the experiments carried out with AprioriGWAS, only two-locus interaction models were considered.

The primary benefit for using rule mining is its flexibility. This can help alleviate the curse of dimensionality by shrinking unnecessary dimensions of the feature space, thus generating more compact and significant rules. ARM will be discussed in more details in Chapter 3.

2.9 Multilayer Feedforward Artificial Neural Network

Artificial Neural Networks (ANNs) are a promising approach for dealing with the limitations associated with modelling epistasis (Günther et al. 2009). A key element for their success is their ability to solve supervised and unsupervised problems and to deal with complex non-linear relationships between features. Artificial neural networks are machine learning models that imitate biological neurons in the human brain to conduct function approximation and pattern recognition from a set of samples (Manning et al. 2014). The neurons are arranged into layers and each layer is fully connected with neurons in the next layer. An important aspect of ANNs is that they are model free meaning that no assumptions about the genetic architecture that produce a particular phenotype are made, a property particularly relevant when mining high dimensional data.

Neural networks predict the outcome based on the transformed representations of input features. One of the most frequently applied ANN architectures in bioinformatics is the feedforward ANN (FNN) also known as the multilayer perceptron (MLP) (Chen & Kurgan 2012). The goal of the MLP is to find a function f: X→Y, capable of approximating the values of output variables (Y) dependent on the set of input variables (X). At its most basic level, an MLP has an *input layer*, *hidden layer(s)*, and an *output layer* as depicted in Figure 2-17.

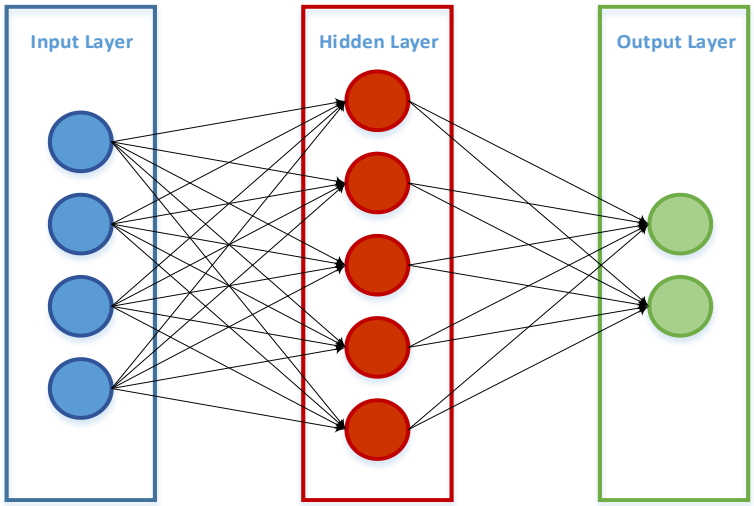


Figure 2-17: Illustration of a single hidden layer NN. The edges connect the output of one node to the input of another

The input layer reads data values from inputs provided by the user (i.e. a suitable representation of SNPs). The hidden layer is constructed from several nodes or neurons that carry out some mathematical operations to derive transformed features and forward them to the next layer in the network (hidden or output). The output layer then uses the transformed features in a model to predict the outcome. The number of neurons in the input and output layers depends on the number of input and output variables required to investigate the problem in hand. Neurons in consecutive layers within the MLP are connected via weighted connections where the values are adaptively updated during the training process. These weights represent the strength of the signal exchanged between two nodes/neurons. This process is referred to as feed-forward processing in ANNs.

The learning process for ANNs can be broadly summarised into three phases (Salari et al. 2014):

- Calculation of the output corresponding to input – feed-forward phase.
- Error calculation and propagation to previous layers – backpropagation phase.
- Adjustment of network weights – adjustment phase using gradient descent.

Supervised learning with an MLP for classification tasks involves a training stage where the network produces some output using inputs and their corresponding target data repeatedly. During each iteration or epoch, the network is presented with a new input sample and the weights of each neuron are adjusted using a learning algorithm that minimises the computed error. This error is then fed back (back propagated) to the network to adjust the weights until a global error is obtained.

MLPs have been widely applied to a range of classifications and regression tasks in research areas such as, accounting and finance, health and medicine, and engineering and manufacturing among others (Paliwal & Kumar 2009). Different ANN architectures have also been developed for use in several bioinformatics scenarios (Manning et al. 2014) such as gene identification/prediction, protein secondary structure prediction, gene interaction and disease diagnoses; and applied to numerous diseases such as cancer (Lisboa & Taktak 2006). However, the use of ANNs in genetic studies

of obesity has not received the same level of interest (Disse et al. 2018; Valavanis et al. 2010). Yet our own research has shown the potential value for this approach in enhancing predictive accuracy in cases of non-linear interactions between features (Curbelo, Fergus, Hussain, Al-Jumeily, Abdulaimma, et al. 2017; Curbelo, Fergus, Hussain, Al-Jumeily, Dorak, et al. 2017; Curbelo, Fergus, Curbelo, et al. 2018; Curbelo, Fergus, Chalmers, et al. 2018).

2.10 Deep Learning

Deep learning (DL) is a type of ANN and one of the most active fields in machine learning. The success of DL can be attributable to the ability of ANNs to approximate nonlinear functions, its high number of parameters and, its flexibility to modify the architecture to adjust to specific problem/domains. DL architectures have shown a marked improvement in image and speech recognition, natural language understanding and most recently, in computational biology (Angermueller et al. 2016). DL networks are characterised by a deep depth of hidden layers and neurons – typically more than two layers.

Several variants of DL are available, including convolutional neural networks (CNNs), which have shown excellent results in computer vision, and recurrent neural networks (RNNs), which perform well in natural language processing. In the field of health informatics, CNNs have had the most impact. Other less frequently explored DL architectures in the field of genetics include stacked autoencoders (SAE) which operate as deep autoencoders for automated feature extraction/learning. SAEs have also been applied to backbone structure

prediction in protein sequences (Lyons et al. 2014) and in magnetic resonance medical image analysis for multiple organ identification in cancer studies (Hoo-Chang Shin et al. 2013).

Deep learning has been used (Eraslan et al. 2016) to select regulatory SNPs with functional impact before association analysis is conducted (DeepWAS). The study focused on variants (SNPs) that alter functional regulatory elements (i.e. elements that control gene expression and DNA methylation) which are identified using a Deep Learning-based algorithmic framework: DeepSEA (Zhou & Troyanskaya 2015). This step is conducted before association analysis since GWAS by itself does not directly provide information about the underlying molecular mechanisms. DeepSEA utilises random information from the 1000 Genomes Project to calculate a threshold parameter (e-value). This value is used to calculate the impact of an SNP on functional reads. The authors used the functional genotypes from case/control GWAS data to associate SNPs with major depressive disorder (MDD). Instead of using standard logistic regression for association analysis, L1 regularised multiple (multi-SNP) logistic regression (LASSO) was utilised.

2.10.1 Autoencoders

Deep feedforward neural networks have been used for unsupervised feature learning and non-linear dimensionality reduction (Hinton & Salakhutdinov 2006). These unsupervised learning architectures learn low-dimensional feature representations from high dimensional unlabelled data, similar to classical PCA, but using non-linear models instead. Autoencoders (AEs) belong to this unsupervised learning class of algorithms and can be used to initialise the

weights of complex models such as neural networks to help improve classification performance (Le 2015).

In bioinformatics, stacked autoencoders have been employed in breast cancer detection using gene expression data (Danaee et al. 2017). In this approach, denoising autoencoders were stacked to extract functional features from high dimensional gene expression profiles. Reduced features were then evaluated using a supervised classification model. Researchers have also conducted experiments where linear representations of gene expressions were obtained using PCA whereas nonlinear relationships were capture using AEs, to enhance cancer diagnosis and classification from gene expression data (Fakoor et al. 2013).

2.10.2 Hyper-parameter optimisation

In ANNs, hyper-parameters are variables that are either set a priori or automatically through an external model-tuning mechanism. These parameters are related to the functions used in feature transformation and class prediction, which include weight initialisation, learning rate, activation function, regularisation, and hidden unit configuration among others (Cook 2016; Candel & LeDell 2018). Identifying a suitable configuration for hyper-parameter tuning requires specific knowledge, intuition, but more importantly it is often based on trial and error testing to attain a good model (Mantovani et al. 2015). Typically, machine learning models are trained using a training set and validated using a holdout or validation set. To ensure that overfitting does not occur a test set is used to evaluate model performance.

Hyper-parameter optimization tries to find an optimal set of hyper-parameters that minimise the generalisation error *E* for a given learning algorithm which, in turn, produces classifiers with good predictive performance (Bergstra & Bengio 2012). However, this can be very challenging when the number of parameters to be tuned is high, such as in ANNs (Bergstra et al. 2011). Methods for optimizing hyper-parameters in machine learning approaches include grid search, Bayesian optimization, random search, and gradient-based optimisation; grid search is a common approach in the literature (Mantovani et al. 2015; Braga et al. 2013).

Overfitting is a common phenomenon, when using supervised ANNs with overly complex structures, that needs to be considered during the training process (Piotrowski & Napiorkowski 2013). It occurs when models memorise training data but do not generalise to new cases. Several techniques are used to prevent overfitting from occurring or, at least reduce it. Among these techniques, decreasing model complexity (i.e. the number of hidden layers and neurons), increasing the size of the training set or using regularization are all valid solutions (Bishop 2006). Although, it has been suggested that using regularisation techniques such as dropout, early stopping and layer-wise pre-training can help to avoid overfitting in deep neural networks (Sheehan & Song 2016). Dropout regularisation is a technique that prevents neurons from co-adapting, which reduces overfitting. This approach has been successful in many domains including object classification, speech recognition or analysis of biology data (Srivastava et al. 2014). Dropout achieves this by randomly selecting a fraction of neurons in each layer and dropping them out of the

training process by setting the neuron values to zero. When performing tests, no neurons are dropped but instead their weights are scale appropriately based on:

$$W_{test}^{(l-1)} = pW^{(l-1)} \quad 2\text{-}5$$

where l is the layer where the neurons are dropped with probability p (i.e. p = 0.5 indicates that 50% of the neurons are dropped at an iteration). Thus, during test, the incoming weights to the layer l are scaled by p according to Equation 2-5. If the validation performance starts to deteriorate, the training process is stopped, and the parameters of the best model in the validation set are chosen. This regularisation process is termed early stopping. Another way to reduce overfitting is to pre-train the layers of the network in a unsupervised manner via autoencoders or restricted Boltzman machines, rather than training the entire network right from the start (Bengio et al. 2007).

Activation functions, such as the rectifier, tanh and maxout (Cook 2016) are typically used in classification tasks - these are defined in Table 2-6.

Function	Formula	Range
Rectifier Linear	$f(\alpha) = \max(0, \alpha)$	$f(\cdot) \in \mathbb{R}_+$
Tanh	$f(\alpha) = \frac{e^{\alpha} - e^{-\alpha}}{e^{\alpha} + e^{-\alpha}}$	$f(\cdot) \in [-1,1]$
Maxout	$f(\alpha_1, \alpha_2) = \max(a_1, a_2)$	$f(\cdot) \in \mathbb{R}_+$

Table 2-6: Available activation functions used in this thesis (Candel & LeDell 2018)

Finding the best activation function depends on the data, since each of them may outperform each other depending on the experimental setting. Therefore, it is recommended to try all options or use a grid search approach to determine which one performs better with the data used for training the models.

Additionally, all these activation functions can be used with dropout regularisation (it controls the rate at which outputs are randomly set to zero) in order to avoid overfitting and produce a more robust model. Table 2-7 lists some examples of available deep learning tuning parameters used in this study.

Tuning Parameter	Description
Activation	Activation function to be used in the network.
Input dropout ratio	Fraction of input neurons to be dropped from training. Helps improve generalisation.
Hidden dropout ratio	Fraction of inputs in each hidden layer to be omitted from training. Helps improve generalisation.
Learning rate	Learning rate at each iteration of gradient descent.
Rate annealing	Is used to reduce the learning rate to avoid getting stuck in local minimum.
Rate decay	It controls the modification of the learning rate across layers.
Stopping metric	Metric to decide whether to stop training early or not.
Early stopping	Stop training if model does not improve for a certain number of scored rounds.
Stopping tolerance	Stop training if the stopping metric has not improved as indicated by this value.
Stopping rounds	Number of epochs before the model stops if it has not improved as indicated by stopping tolerance.
Input_dropout_ratio	A fraction of features in each training row to be removed from training. This can improve generalization.
L1	Lasso regularisation. Sets the value of many weights to 0.
L2	Ridge regularisation. Sets the value of many weights to smaller values.
Hidden layers	Number of hidden layers.
Neurons	Size of hidden layer.
Epochs	Number of iterations over the training set.

Table 2-7: Tuning parameter example used in this study (Candel & LeDell 2018)

Random search has proven to be as good as, or even better than, pure grid search when applied to ANNs, saving computational time (Bergstra & Bengio 2012). This is true since random grid search can effectively search a larger, and often less promising, configuration space (Bergstra & Bengio 2012).

2.11 Systems Medicine

In recent years, the intersection between medical research and practice, and expertise from biology, biostatistics, informatics, mathematic and computational modelling has given rise to an interdisciplinary approach termed *Systems Medicine* (Wolkenhauer 2013; Kramer et al. 2018). As an extension and adaptation of *Systems Biology* (Wolkenhauer et al. 2013), systems medicine emerged in medical research to understand and treat diseases by studying not only the elements of the system, but their interactions (Gomez-Cabrero, Menche, et al. 2014). This concept is speeding up changes in clinical and translational research and healthcare by bringing investigator teams and expertise from different disciplines together. The potential of systems medicine in the study of complex diseases have been confirmed in cases of chronic respiratory diseases, i.e. Chronic Obstructive Pulmonary Disease (COPD) (Gomez-Cabrero, Menche, et al. 2014), cholesterol and glucose regulation (Shu et al. 2016).

In this PhD, the investigation of obesity as a complex disease can be compared with a systems approach where the genetic elements contributing to the understanding of common obesity are identified systematically to, finally, investigate consequences of particular interactions, the emergent patterns or behaviour of the system. To achieve this, it is necessary to provide data

integration from different sources (i.e. sequencing technologies and external databases), and analysis and interpretation of this data through the combination of multiple statistical, computational and mathematical procedures (integrative workflow) in a rational and reproducible way implemented in software tools (Apweiler et al. 2018). Such strategies may facilitate the discovery of combined biomarkers with predictive power of disease (Gomez-Cabrero, Lluch-Ariet, et al. 2014).

Therefore, SAERMA is presented in this thesis as a generic pipeline that helps leverage combined statistical patterns from GWAS, association rule mining and deep learning to identify relevant pathways and key drives in biological systems as similarly conducted in (Shu et al. 2016).

2.12 Chapter Summary

Identifying the genetic cause of obesity is complex. This multifactorial disease is caused by environmental changes, eating behaviours, physical activity and genetic factors. Whereas environmental and lifestyle changes have driven obesity prevalence to epidemic proportions, there is evidence that a substantial genetic component exists (h^2 ~ 40-70%) supported by heritability studies. Although many genes that play a significant role in the development of obesity have already been identified, a large proportion of the heritability is still unexplained, making this a prominent area of research.

As the price of genome-wide genotyping has dropped, the number of studies utilizing GWAS has increased dramatically (Gretarsdottir et al. 2010; Speliotes et al. 2010; Tryka et al. 2014; Kamitsuji et al. 2015; Frazer et al. 2009;

MacArthur et al. 2017). The importance of GWAS is advancing scientific understanding of disease mechanisms and providing starting points and potential opportunities for researchers to improve the development of medical treatments or prevention therapies (Blank & Gutzwiller 2014; Christensen & Murray 2007). Results from GWAS reveal SNPs that serve as candidate biomarkers for genes and these might provide important indicators for the existence of complex diseases in individuals. This approach is based on single-locus analysis where each SNP is independently tested for association with a phenotype of interest, omitting the existence of interactions between loci. Genetic studies of obesity have mainly considered the effect of single variants or sets of variants previously associated with BMI and obesity related traits where the joint effect or epistatic interactions have also been ignored or investigated in less detail.

The amount of data extracted from GWAS opens up new opportunities to establish and develop suitable analytical methods that help to translate knowledge into biological and clinical discoveries. This thesis will build on the significant work already done and concentrate on providing new insights into the genetics aspects of obesity. This field still needs further exploration and new techniques, such as those posited in this thesis, may help to identify new variants or new interactions between them by applying state of the art artificial intelligence techniques, particularly machine learning and the new advances currently being made in deep learning (discussed in more detail below). The discovery of interactions between genetic variants (epistasis) is currently a subject of active development in statistics and machine learning.

Gaining a better understanding about the causes of obesity and related comorbidities, including cardiovascular disease and type 2 diabetes mellitus (T2DM) is one of the main goals in genetic investigations of obesity. In this sense, systems medicine arises as a novel multidisciplinary approach to medicine, which performs a methodological pipeline consisting of quantitative technologies to produce data, information systems for data management as well as methods for analysis and interpretation of the data. Furthermore, the interpretation and translation of such knowledge opens up new opportunities to introduce personalised medicine to obese patients, enabling more specific diagnosis of the causal factors underlying obesity.

When susceptibility to complex traits in diseases is analysed, epistasis continues to emerge as a likely explanation for missing heritability where many genetic factors interact simultaneously. In the case of obesity, these factors not only act independently but they also interact with each other and the environment. The computational burden however in exploring interactions between genetic variants in the sea of data generated in GWAS, combined with small to moderate sample sizes, has prevented epistasis from being the main focus in GWAS analyses.

Therefore, most methods used in studies on complex diseases are based on traditional statistical regression models characterised by univariate tests. Here, the genetic variants that have independent effects on the phenotype are unable to capture complex interactions between multiple variants (Saeys et al. 2007). Nevertheless, the known limitations of traditional methods in situations where non-linearity and high-dimensional settings are an issue, has led to the pursuit

of more complex approaches such as, machine learning, which have proved to be more effective (Salari et al. 2014; Valavanis et al. 2010).

Multivariate methods permit the identification of complex interactions between genetic and non-genetic risk factors that modulate the probability of developing a specific condition and its level of severity. Still, investigating epistasis can be extremely computationally expensive due to the high dimensional space and the high number of models that need to be explored. Autoencoders represents a flexible approach capable of dealing with nonlinearities in the data while allowing them to be stacked to form a deep network.

While several strategies can be used to help reduce computational burden, no one strategy will be optimal in all cases. Different methods have advantages and disadvantages, indicating that they do not follow a "one fits all" criteria. Therefore, in the experiments presented in this thesis, a filter approach based on the statistical evidence of single-SNP effects has been implemented since it is simple, unbiased with respect to the researcher (no previous biological knowledge is required) and it has been shown to have high power. This thesis is built on existing methodologies and combines supervised and unsupervised machine learning methods into a single framework to investigate epistasis in obesity GWAS. Combined, the proposed framework includes ARM, deep learning SAE and an MLP, and is discussed in detail in the following chapter.

Genetic variants identified by GWAS and subsequent analysis do not necessarily represent specific genes but genomic regions. This complicates the task of making direct biological inference from the results of statistical tests, so

additional information is necessary to interpret and empirically validate the results. Therefore, GSEA, pathways analysis and eQTL information can be used to provide functional interpretation of many trait-associated SNPs in a biological context. In this thesis, biological validation of proposed genetic variants identified by ARM are validated using functional analysis via functional annotation tools, to provide a biological interpretation of the mined rules prior to feature extraction and classification analysis.

Chapter 3. METHODS

3.1 Introduction

The proposed method described in this thesis expands GWAS analysis by combining techniques not used before to learn the epistatic interactions between SNPs to model and classify extremely obese and non-obese individuals.

Quality control (QC) and logistic regression (LR) steps typically performed in GWAS are combined with association rule mining (ARM) and deep learning (DL) stacked autoencoders (SAE) to create a novel framework for learning epistatic interactions between SNPs. A multilayer feedforward artificial neural network classifier is initialised using SNP features and the epistatic information learned by a deep learning stacked autoencoder model guided by ARM to classify case-control samples obtained from the eMERGE MyCode dataset. The complete network models the epistatic effects of SNP perturbations while ARM provides model interpretation via network visualisation and rule inspection.

The approach comprises six stages (Figure 3-1 illustrates the complete algorithm):

1. Data file pre-processing.
2. Quality control.
3. Association analysis.
4. Combination of association rule mining and stacked autoencoders.
5. Classification using a multilayer perceptron artificial neural network.
6. Interpretation of the results.

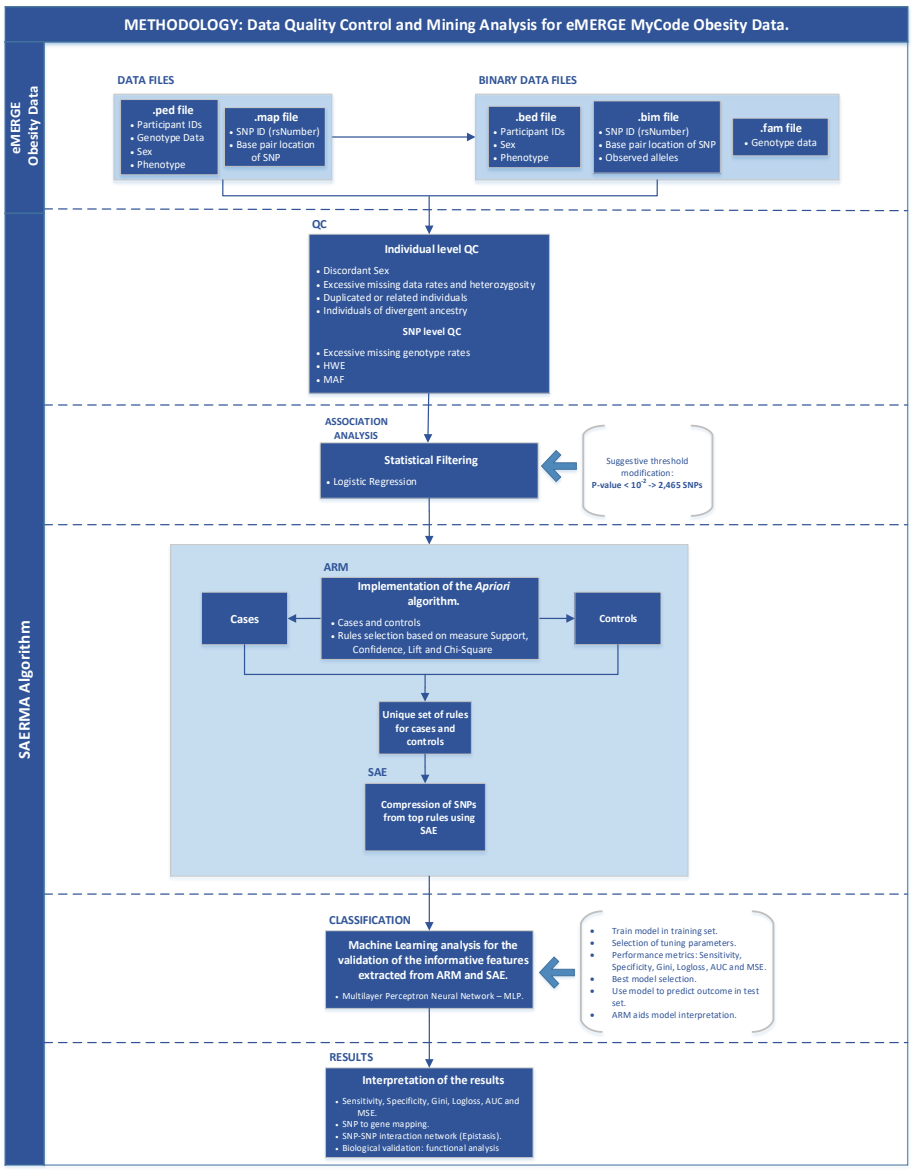


Figure 3-1: Proposed methodology

In general terms (see Figure 3-2), the proposed algorithm investigates the genetic architecture of obesity by taking the interactions between genetic variants into account, using a case-control dataset. To achieve this, several machine learning methods (ARM, SAE and MLP) have been used along with common genetic techniques utilised in GWAS (QC and association analysis). The results obtained by the deep learning architecture, constituted by SAE-MLP, were driven by ARM which allows for an interpretation of the model.

Finally, statistical results from ARM are also validated from a biological point of view via functional analysis using gene set enrichment analysis.

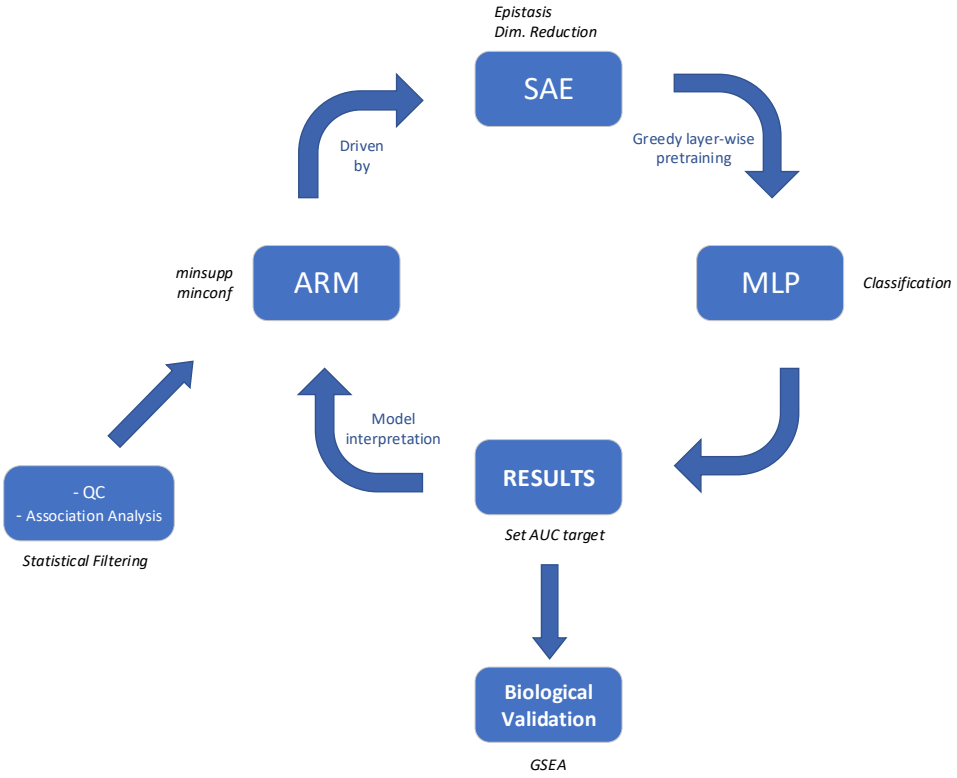


Figure 3-2: Overview of SAERMA

3.2 Data Description

Case-control data utilized in this study were obtained from the database of Genotypes and Phenotypes (dbGaP) (CHEN et al. 2012), which provides open and controlled access to genomic data and phenotypic information. Access to the following datasets was granted following a formal request to dbGaP:

- Control: eMERGE Geisinger eGenomic Medicine (GeM) - MyCode Project Controls (dbGaP study accession phs000381.v1.p1).
- Case: eMERGE Genome-Wide Association Studies of Obesity project (dbGaP study accession phs000408.v1.p1).

Participants are part of the MyCode Community Health Initiative (MyCode) project. This is a sophisticated platform for translational research (Carey et al. 2016). It was created as a central biorepository to collect blood and DNA samples from a representative cohort of patients from the Geisinger Health System (GHS). This is an integrated health care delivery system that provides services to participants that are resident in Pennsylvania. Samples and molecular data generated by MyCode have been used in numerous research studies, including the electronic Medical Records and Genomics (eMERGE) Network. This network represents a collaboration between institutions with biobanks linked to electronic medical records (EMRs) and is supported and funded by The Genomics Workgroup of the NHGRI (McCarty et al. 2011). One of the institutions that supplies anonymised samples to the EMR is the Geisinger Clinic, among others (Gottesman et al. 2013).

Cases and controls provided by dbGaP were extracted from different study cohorts provided by the Geisinger MyCode project. Control patients from *eMERGE Geisinger eGenomic Medicine (GeM) - MyCode Project Controls*, were eligible if they were primary patients of a Geisinger Clinic with non-urgent visits. A subset of 1,231 unique samples were genotyped using Illumina HumanOmniExpress-12 v1.0 arrays and used as population controls for *eMERGE Genome-Wide Association Studies of Obesity project* dataset. All study participants provided written consent prior to study enrolment as part of the MyCode DNA biobank.

Case samples were part of a cohort of primary Caucasian patients from the Geisinger Clinic with extreme obesity who underwent bariatric surgery. A

subset of 962 unique samples with a mean BMI of 49.17 (± 8.83 SD) was genotyped using Illumina HumanOmniExpress-12 v1.0 arrays. The control group (1,231 individuals) includes 488 females and 743 males with a mean age of 66.74 (± 13.95 SD) while cases comprise 788 females and 174 males with a mean age at surgery of 46.42 (± 11.26 SD). The final dataset contains a total of 2,193 participants of which 917 are males and 1,276 are females. The proportion of males and females by phenotype is illustrated in Figure 3-3.

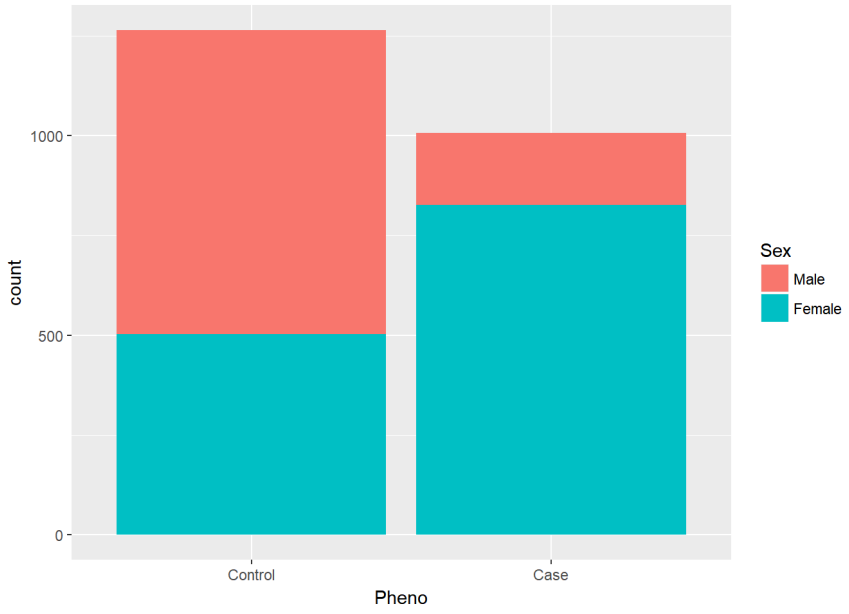


Figure 3-3: Proportion of males and females per phenotype label

Each participant contains 594,034 markers. Furthermore, 99.5% of the participants belong to a white ethnic background (Caucasians) as shown in Table 3-1.

Ethnicity	Case	Control	Total samples
White	960	1,223	2,183
Black or African American	1	5	6
American Indian/Alaska Native	0	1	1
Hispanic or Latino	0	1	1
Native Hawaiian or other Pacific Islands	0	1	1
Unknown	1	0	1

Table 3-1: Case-control samples by population

Genetic data is encoded in the pedigree file format (PED/MAP) although it was converted to binary format (BED/BIM/FAM) for efficiency reasons as discussed earlier in Chapter 2 (Section 2.4.4.1). Genotype data for all individuals (cases and controls) is contained in the *bed* file. The subject-related information (i.e. FID, IID, PID, Sex and phenotype) is stored in the *fam* file. Finally, the *bim* file contains information about all relevant SNPs used in the study.

The case-control sets were merged, and the phenotype was defined before conducting any analysis. Phenotypes (affection status) in both datasets were originally set to missing (-9). Therefore, information for cases (severe cases of obesity) and controls (healthy individuals) was updated accordingly (-9 or 0 missing, 1 unaffected and 2 affected). Experiments in this study have been conducted using the binary files (.bed, .bim, .fam). To make the merge process easier, phenotype values were assigned to cases and controls separately. Both datasets were merged to create the main dataset used for QC and association analysis. Thus, the *fam* files were updated to match the right phenotype in cases and controls. Sex values are coded as 1 = male, 2 = female and other = unknown.

In addition, the case and control datasets were supplied with clinical data. The control dataset (dbGaP study accession phs000381.v1.p1)**,** provides medical background information about healthy patients, including general socio-demographic information, weight, height and BMI, blood pressure data, abdominal aortic aneurism (AAA), peripheral arterial disease (PAD), arterial dissection, bariatric surgery and information about tobacco or alcohol consumption. The case dataset (dbGaP study accession phs000408.v1.p1) provides medical background information relating to obesity prior to bariatric surgery, and includes information about weight, height and BMI, Hba1c, blood pressure data, insulin, glucose levels, and medication use, including biguanides, insulin, sulfonylureas, or insulin sensitizing agents. Data about tobacco and alcohol use, general socio-demographic information and weight measurements following bariatric surgery, are also included.

Dataset balance

Typical GWAS experiments suffer from data imbalance problems where the number of control samples collected from healthy individuals is greater than individuals with the desired trait or phenotype under investigation (Bao et al. 2016; Zhou et al. 2018). Hence, datasets with a large number of controls and a small number of cases, introduce bias in association tests such as logistic, in favour of the dominant group (Owen 2007).

Therefore, dataset balance is an important issue to be taken into consideration before performing analysis. Generally, the ideal situation to have the least biased performance consists in having approximately 50% of the individuals belonging to cases and 50% to controls. In this thesis, the MyCode

dataset does not pose a significant imbalance problem to warrant intervention and as such is not necessary in this PhD (Haibo He & Garcia 2009).

In the remainder of this thesis, cases and controls from the Geisinger MyCode will be denoted as the MyCode dataset that is sufficiently balanced for further study.

3.3 Quality Control (QC)

To conduct association analyses, only those individuals reported to be Caucasian (white) were selected to reduce potential bias due to population stratification (Price et al. 2010). In addition, analysis was also conducted to identify problematic samples and SNPs. Thus, QC of the genotyped data and filtering procedures were performed on individuals and then on markers (SNPs) to maximise the number of remaining SNPs. It should be noted that there is no universally accepted QC threshold for the exclusion criteria. Therefore, all QC steps were performed in accordance with standard QC protocols and guidance from (Anderson et al. 2010; Weale 2010). The protocol written by Anderson et al. has been successfully applied in GWAS elsewhere (Ferrari et al. 2015).

3.3.1 Individual Level QC

3.3.1.1 Identification of individuals with discordant sex information

The first QC step conducted on MyCode dataset was to identify and remove data samples with discordant sex, using genotype data from the X-chromosome. In PLINK, sex inconsistencies were checked using the command --check-sex. After identifying discordant sex information, 3 individuals were removed from the main dataset.

3.3.1.2 Identification of individuals with elevated missing data rates or high heterozygosity

Using PLINK, individuals with elevated missing data rates or high heterozygosity were examined (--missing and --het). The plot in Figure 3-4 was utilised to determine appropriate thresholds. In the figure, missingness and heterozygosity are considered together to determine the filtering thresholds.

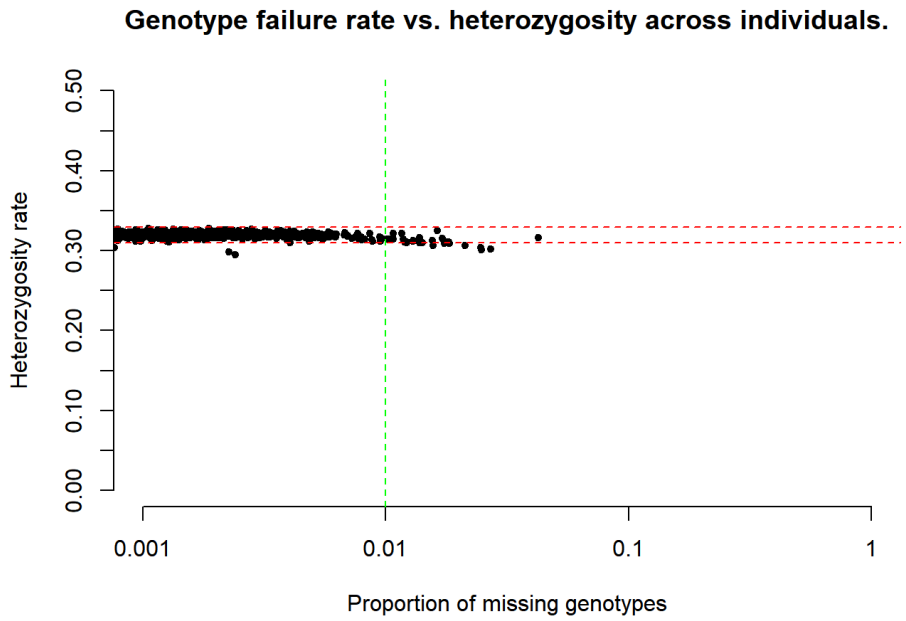


Figure 3-4: Genotype failure rate vs. heterozygosity across all individuals in the study. Dashed lines denote QC thresholds selected

Hence, individuals with a genotype failure rate ≥ 0.01 (vertical line) and heterozygosity rate ± 3 standard deviations from the mean (horizontal lines) were excluded. This resulted in 43 individuals being removed.

3.3.1.3 Identification of duplicated or related individuals

IBD was calculated for each pair of individuals based on the average proportion of alleles shared in common at genotyped SNPs, without considering the sex

chromosomes. This is then used to identify and remove duplicated or related individuals using IBD coefficient estimates (IBD > 0.185).

To reduce computational complexity, SNPs from extended regions of high LD are excluded. This is an important quality assurance step for GWAS analysis, especially when performing IBD estimation or PCA, which will obtain better results if the SNPs are not in LD with each other (Anderson et al. 2010). To remove data redundancy due to LD, each individual chromosome was scanned using a moving window size set to 50 SNPs with a step length of 5 SNPs. Furthermore, LD cut-off was set to 0.2. Again, these experiments were conducted using PLINK. This resulted in 156 individuals being removed due to IBD.

3.3.1.4 Identification of individuals of divergent ancestry

Genotypes are merged with HapMap phase 3 data from three ethnic populations: (Europe) CEU, (Asia) CHB+JPT and (Africa) YRI. Principal component analysis was performed on the case-control data to identify outliers and hidden population structure using EIGENSTRAT. A scatter diagram for the two first principal components is depicted in Figure 3-5 which is sufficient to cluster samples from the three populations.

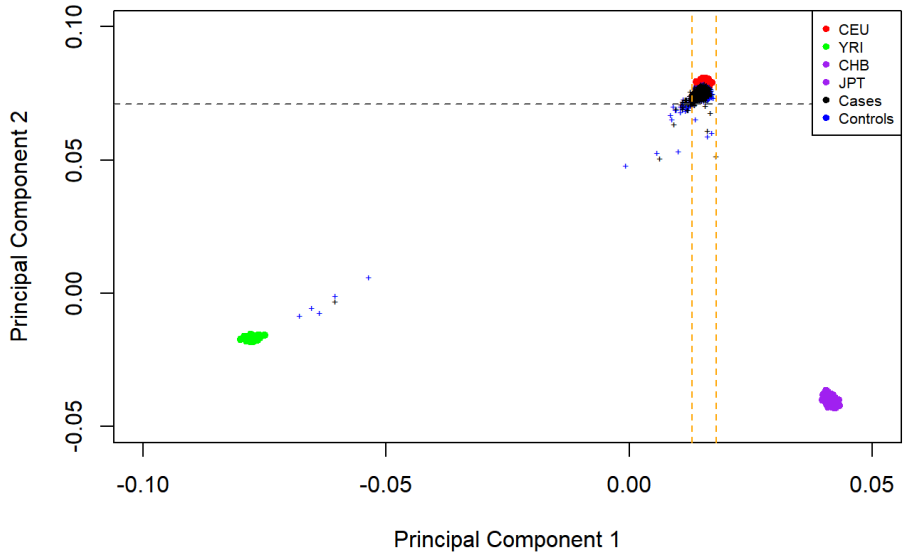


Figure 3-5: Ancestry clustering based on GWAS data

PC1 and PC2 thresholds were derived so that only individuals that match the given ancestral population are included. For populations of European descent this will be CEU HapMap3 individuals. Here, all individuals with a second principal component score less than 0.071 were excluded as shown in Figure 3-5 (dark dash-line). Conversely, PC1 values between 0.013 and 0.018 were selected (orange dashed line). After performing divergent ancestry QC, 93 samples were removed.

3.3.2 SNP Level QC

Following the removal of individuals failing QC, QC applied to genetic variants is performed. This QC step is usually composed of 3 steps including: identification of genetic variants (SNPs) with excessive missing data rates, SNPs with significant deviation from HWE and pruning of SNPs with low MAF.

3.3.2.1 Identify all markers with excessive missing data rates

The missing genotype rate for each marker is calculated using PLINK. The command --*geno* 0.01 was used to exclude genetic variants (SNPs) with genotype missing data rates higher than 0.01, equivalent to a call rate of less than 99%. This resulted in 323,529 genetic variants being removed.

3.3.2.2 Hardy-Weinberg equilibrium

A Hardy-Weinberg Equilibrium P-value of $1x10^{-5}$ was used as a threshold in control subjects to remove SNPs failing this QC step (--*hwe*). However, 0 variants were removed suggesting that there was no deviation from HWE.

3.3.2.3 Minor allele frequency

Genetic markers (SNPs) were excluded from analysis if the MAF was lower than 5%. This resulted in 28,848 SNPs being removed.

3.4 Association Analysis

In this study, association analysis is used to reduce the computationally large number of genetic variants (240,950 SNPs after QC) prior to epistatic analysis and machine learning tasks. Statistical association testing between individual SNPs and the obesity phenotype was conducted under an additive model using logistic regression (Li 2007). A logistic function was used to predict the probability of a case given a genotype class although other genetic models are available (Bush & Moore 2012). Covariates help to explain some of the phenotypic variation, which can help to improve the power of statistical testing in linear models and logistic regression analyses of quantitative traits (Pirinen et al. 2012). However, association analysis in this thesis is used as a statistical

filtering approach to reduce the dimensionality of the data. Therefore, association analysis with logistic regression was conducted without taking into consideration any covariates.

Genotypes are grouped into an additive model. Given *a*, a uniform, linear increase in risk for each copy of the *a* allele is assumed. For example, if the risk is γ for *Aa/aA*, then there is a risk of 2 γ for *aa*. Let *i* be the individuals (*i* = 1, 2, …, n), Y_i the phenotype for individual *i* and X_i the genotype of individual *i* at a particular SNP. Let $Y \in \{0,1\}$ be a binary phenotype for case/control status and $X \in \{0,1,2\}$ be a genotype at the typed SNP, where 0, 1 and 2 represent homozygous major allele *AA*, heterozygous allele *Aa*, and homozygous minor allele *aa* respectively. Let p_i represent the expected value of a phenotype Y_i, given a genotype X_i,

$$p_i = E\left(Y_i \mid X_i\right) \tag{3-1}$$

Logistic regression modelling is therefore defined as (Wang et al. 2016):

$$logit\left(p_i\right) = ln\left[\frac{p_i}{\left(1-p_i\right)}\right], \tag{3-2}$$

however, it can also be given as a linear predictor function:

$$logit\left(p_i\right) \sim \beta_0 + \beta_1 X_i \tag{3-3}$$

PLINK was utilised to test the association between SNPs and obesity as a binary trait, where the default option format for the phenotype was considered: 1 = unaffected, 2 = affected and 0 or -9 to represent a missing phenotype. To

test for association, the option --*logistic* was utilised. Population stratification was assessed, and standard errors were adjusted using the genomic inflation statistic (λ). Quantile-quantile (Q-Q) and Manhattan plots were generated to visualise the GWAS results (Turner 2014).

Utilising logistic regression, while not ideal, enables the number of SNPs with insignificant marginal effects to be reduced to meet the computational needs required for epistatic analysis and machine learning tasks. An advantage of applying this filtering technique is that it greatly reduces the number of combinations that need to be evaluated in subsequent machine learning experiments, thus reducing the chance of overfitting (Moore & Andrews 2015). The remaining SNPs capture the significant linear associations between SNPs and the phenotype.

The results of all association tests with P-values lower than $1x10^{-2}$ were considered in this thesis to allow epistatic interactions to be detected and minimise computational overheads. This threshold was selected based on empirical results obtained in our previous work (Curbelo, Fergus, Hussain, Al-Jumeily, Dorak, et al. 2017; Curbelo, Fergus, Curbelo, et al. 2018). This approach has been adopted in other previous studies of epistasis in obesity (S. Lee et al. 2012) and T2D studies (Gül et al. 2014) where P-value < $1x10^{-1}$ and P-value < $1x10^{-3}$ were considered respectively.

3.5 Association Rule Mining (ARM)

In bioinformatics, association rules can be utilized to reveal biologically relevant associations between different SNPs. The guiding principles are that if

attributes frequently appear together, there must be an underlying relationship between them. Studies based on SNP data have described large datasets with limited focus on plausible interactions between genetic variants. Therefore, exploring the intrinsic relationships in the data and extracting rules to better understand SNP behaviour and their subsequent interactions between each other are important tasks that can be performed using frequent pattern mining (FPM). This technique extracts all the frequent itemsets from a dataset, which are then used to generate association rules. In the proposed methodology, the idea is to extract important rules identified in cases and controls separately.

The biological extrapolation in this study is to identify frequently occurring SNPs as items, in different individuals in the form of transactions. Applied to GWAS, the individuals are transactions, SNPs are items, and SNP combinations are itemsets. Single SNPs tend to have small effect sizes in polygenic diseases. Therefore, by looking at the joint effect of multiple SNPs, explanatory power can be increased.

SNP genotypes recoded in terms of additive components, following logistic regression analysis, are translated into transaction data. To do so, SNPs are first coded into numeric values: 0, 1 and 2, using the PLINK command --*recodeA*. This produces a single column for genotype data in terms of minor allele numbers, in the format [Variant ID]_[counted allele]. The number of alleles is 0 if the genotype is *AA*, 1 if genotype is *Aa/aA* and, 2 if genotype is *aa*. Once completed, the recoded dataset is partitioned into cases and controls to extract rules for each group separately. Hence, in the rule mining experiments (described later in this thesis), each item (SNP) in each transaction (individual)

is labelled as dominant (D), heterozygous (H) or recessive (R) according to 0, 1 and 2 respectively. An example SNP identified in the study has the following format rs322132_A_D, where A refers to the counted allele and D indicates that is dominant.

ARM is one of the main techniques used to detect and extract useful information from large scale transaction data. This step is conducted after itemset mining to allow appropriate rules to be derived from itemsets. Many ARM algorithms, including the *Apriori* algorithm (Agrawal & Srikant 1994), assume a common strategy for decomposing mining problems into two principal subtasks: 1) *Frequent itemset generation* and, 2) *rule generation*. In the case of frequent itemset generation the aim is to identify all the itemsets that satisfy a minimum support threshold. While rule generation extracts all the high-confidence rules from the frequent itemsets that satisfy a minimum confidence constraint.

ARM is therefore formally defined as:

Definition 1 (Items): Let $I = \{i_1, i_2, \ldots, i_m\}$ be a set of m attributes called *items*.

Definition 2 (Transaction): Let a *transaction database* $T = \{t_1, t_2, \ldots, t_n\}$ be a set of n subset of items called *transactions*. Each *transaction* in T identifies a subset of *items* in I.

In this work, items represent SNPs while transactions represent individual samples.

Definition 3 (Itemset): A subset $X = \{i_1, i_2, \ldots, i_k\} \subseteq I$ is referred to as an *itemset*, or k-itemset with k *items*.

Itemsets are sets of *k*-items where *k* starts with 1 to infinity. Unnecessary itemset candidates are produced if at least one of its subsets is infrequent. Hence, the frequent itemset generation is equipped with pruning steps in which it eliminates some of the *k*-itemset candidates based on a minimum support threshold. Support is the number of transactions that contain a particular itemset as defined in Definition 4.

Definition 4 (Support of an Itemset): *Support(X)*, or support of an *itemset X*, refers to the number of *transactions* in *T* that contains the *itemset X*. *Support(X)* is defined as follows:

$$support(X) = |\{t \in T, X \subseteq I\}| \quad 3\text{-}4$$

Definition 5 (Frequent Itemset): Given a set of $items\ I = \{i1, i2, \ldots, im\}$ and a set of $transactions\ T = \{t1, t2, \ldots, tn\}$, a subset of $I, X \subseteq I$, is considered a frequent $itemset$ if X occurs in a percentage of all transactions in T that exceeds a minimum *support* threshold σ, with $0 \leq \sigma \leq |T|$.

In this work, frequent itemsets are independent sets of SNPs in the Geisinger MyCode dataset whose support is greater than or equal to a given minimum support threshold σ, as defined in Definition 5. Itemsets whose support count is lower than the minimum σ, are pruned (eliminated). This strategy for reducing the exponential search of frequent itemsets based on the support measure is termed support-based pruning, which diminishes the number of candidate patterns.

Once frequent itemsets have been obtained, the generation of association rules is performed. Frequent itemset $I > 1$ is divided into two itemsets, *X* and *Y*, representing the elements of a rule. The rules are created if its *support* and

confidence values exceed a given threshold. Rules are generated from each of the frequent *k*-itemsets. Hence, the total candidate association rules generated is $2^k–2$, excluding those that are null in the antecedent (X) or consequent (Y) (Zaki 2000).

Definition 6 (Association Rule): An *association rule* is defined as an implication of the form $X \rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = \emptyset$. X refers to the *left-hand side (LHS)* or *antecedent of the rule* while Y is the *right-hand side (RHS)* or *consequent*.

Association rule mining allows us to discover associations among a subset of SNPs extracted from the MyCode dataset. The approach discovers SNPs that frequently occur together in the MyCode dataset (cases and controls separately) and creates a relationship between those SNPs in the form $X \rightarrow Y$. As stated in Definition 6, this relationship implies that when X occurs it is likely that Y also occurs. Such a relationship is called an association rule.

The significance of the association rules is measured in terms of their *support* and *confidence* although other interest measures such as *lift* or *Chi-Square* can be used to validate rules. The support of a rule is the probability that the samples in a dataset contain both X and Y. Rules with very low support may occur by chance, therefore, support is an important measure that can be used to eliminate unimportant rules. The confidence of a rule, on the other hand, is the probability that a case contains Y given that it contains X. It provides an estimate of the conditional probability of Y given X. The probability representations for the support and confidence of a rule are defined as:

$$support = P(X \cap Y), \qquad 3\text{-}5$$

and

$$confidence = P(Y|X) = \frac{\mathrm{P(X \cap Y)}}{\mathrm{P(X)}} \quad 3\text{-}6$$

Definition 7 (Support of a Rule): The *support* of an *association rule* $X \rightarrow Y$ with regard to a *transaction* set T is given by the *support* of $X \cup Y$:

$$support(X \rightarrow Y) = \frac{support(X \cup Y)}{|T|} \quad 3\text{-}7$$

Definition 8 (Confidence of a Rule): The *confidence* of a rule $X \rightarrow Y$ with regard to a transaction set T is given by:

$$confidence(X \rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \quad 3\text{-}8$$

In the rule generation phase, only confidence is considered since support is contemplated when mining frequent itemsets. Hence, the usual filtration is based on the user-defined minimum support and confidence values. Once specified, *Apriori* looks for rules satisfying the condition as indicated in Definition 9.

Definition 9 (Association Rule Discovery). Given a set of *transactions* T, search for all the *rules* with *support* $\geq \sigma$ and *confidence* $\geq \delta$ where σ and δ are the corresponding minimum *support* and *confidence* thresholds.

3.5.1 Apriori algorithm

The generation of association rules is conducted using the Apriori algorithm (Agrawal & Srikant 1994). The algorithm was proposed by Agrawal and Srikant in 1994 and is one of the oldest, simplest and most popular frequent itemset mining algorithms. The name Apriori is based on the fact that the algorithm

benefits from prior knowledge of frequent itemset properties. The Apriori algorithm performs a breadth-first search (BFS), enumerating every single frequent itemset by iteratively generating candidate itemsets (Hipp et al. 2000). Candidate itemsets of length *k* are generated from *k*-1 itemsets. The support of every candidate itemset is calculated iteratively where itemsets with support values under a defined threshold are disregarded. For a transaction *T* and support threshold σ, the pseudo code for the Apriori algorithm is given in Algorithm 3-1, which has been extracted from (Agrawal & Srikant 1994). Christian Borgelt's *Apriori* implementation is applied in this study using the R package *arules* (Hahsler et al. 2005; Hahsler et al. 2018; Hahsler et al. 2011), to search for partial dependencies in the filtered set of SNPs.

Apriori Algorithm

```
1:  Apriori (T, σ)
2:  L1 = {large 1 − itemsets}
3:  for(k = 2; L_{k−1} ≠ 0; k + +) do begin
4:      C_k = apriori − gen(L_{k−1}); // New candidates
5:      forall transactions t ∈ T do begin
6:          C_t = subset(C_k, t); //Candidates contained in t
7:          forall candidates c ∈ C_t do
8:              c.count++;
9:      end
10:     L_k ={ c ∈ C_k | c. count ≥ σ}
11: end
12: Answer = ∪_k L_k
```

Algorithm 3-1: Apriori algorithm

To manage the very large number of association rules, the patterns are filtered, grouped and organized. This is a crucial step to focus on the most interesting association rules discovered. Nearly all search algorithms rely on support-based pruning. If an itemset *X* is not frequent (given a minimum support), then none of its supersets $Y \supset X$ can be frequent. This property is

known as anti-monocity of the frequency. Furthermore, if the support value is set too low (close to 0), a large number of spurious rules are generated (type I errors). This makes the problem computationally intractable. Conversely, if the value for support is too high (close to 1), a very small number of rules, if any, are extracted, which means that several significant rules can be missed (type II errors). To reduce the aforementioned errors, the traditional support-confidence framework used in rule mining is replaced by a support-dependence framework.

Standard minimum support and confidence measures set by the user are employed by the algorithm to prune uninterested association rules. However, the minimum frequency and confidence requirements do not guarantee statistical dependence or significance. Hence, it is also possible to add additional objective interest measures to each rule, e.g. P-value thresholds computed using the Chi-square test or Fisher's exact test to evaluate the significance of the rules.

3.5.2 Additional Interest Measures

Limitations of the support-confidence based rule mining framework (Tan et al. 2005; Ahn 2012) have given rise to the use of other interestingness measures to evaluate the quality of the patterns identified. Examples of these measures are lift, P-value thresholds computed using the Chi-square test or Fisher's Exact test, and a collection of other objective symmetric and asymmetric interestingness methods (Tan et al. 2005; Hahsler et al. 2005). In this study, those described previously are used in addition to lift and Chi-squared to determine significant rules, as they allow us to measure which rules are more correlated.

Lift or interest (Brin et al. 1997), is a symmetric measure which divides the rule's confidence by the support of the itemset in the rule consequent as shown in Equation 3-9. It can be used to analyse the relativity of association rules mined and for measuring how many times more frequently *X* and *Y* occur together than expected under statistically independent conditions. Lift indicates a positive correlation between *X* and *Y* when its value is greater than one, negative correlation when its value is lower than one, and independence when lift is equal to one. As an example, a lift($X \rightarrow Y$) > 1 indicates that the appearance of *X* promotes the appearance of *Y*, resulting in a positive correlated rule. Thus, the higher the lift, the stronger the positive correlation and the more dependent the SNPs are. In this study, only positive correlated rules are of interest.

Definition 10 (Lift of a rule): The *lift* of an *association rule* $X \rightarrow Y$ is defined as the ratio of the observed *support* for this *association rule*, to the expected *support* if *X* and *Y* were independent.

$$lift(X \rightarrow Y) = \frac{confidence(X \rightarrow Y)}{support(Y)} = \frac{support(X \cup Y)}{support(X)support(Y)} \qquad 3\text{-}9$$

Finding measures that can be used with lift to make the best selection of rules is crucial. Despite the numerous alternatives for expressing the dependence between the antecedent and the consequent of an association rule, the classic Chi-square test statistic (χ^2), is extensively used to assess the significance of dependencies and determine the statistical significance level of association rules (Liu et al. 1999). Rules can then be pruned in cases of independency, meaning that the itemsets (SNPs in this study) in the rule are not correlated. χ^2 helps to decide whether items in the rules are independent of each other, however it is not useful for ranking purposes by itself.

Definition 11 (Chi-square test): Let f_0 be an observed frequency, and f be an expected frequency. The *Chi-squared test statistic* (χ^2) value is defined as:

$$\chi^2 = \sum \frac{(f_0 - f)^2}{f} \qquad 3\text{-}10$$

χ^2 is a summed normalized square deviation of the observed values from the expected values. A χ^2 value equal to 0 implies that the elements of the rule are statistically independent. An important fact regarding the Chi-square test is that it can be used to calculate the P-value to determine the significance level of the rule (Tan & Kumar 2000). For instance, if the P-value of the rule is lower than a significance threshold 0.05, that is a χ^2 value higher than 3.84, we can conclude that X and Y are significantly dependent (the independence assumption is rejected) and, therefore, the rule $X \rightarrow Y$ can be considered in subsequent analysis (Liu et al. 1999). This is one way to identify the direction of a rule when summarising unpruned rules, i.e. by the type of correlation the rules have, as similarly performed by lift (positive correlation, negative correlation or independence). To some extent, χ^2 improves the traditional framework of the interestingness measure provided by lift.

Association rules are generally considered statistically significant if their occurrence is not due to random chance. A combination of different interest measures is necessary to assess the strength and the dependency of the antecedent and consequent of the rules. Discovered associations are pruned to remove non-significant rules, and then a special subset of unpruned associations forms a summary of the discovered associations which represent candidates for epistatic interactions.

3.5.3 Redundancy

Redundancy elimination tasks can be beneficial to reduce complexity by identifying smaller sets of more general rules which are easier to interpret than larger complex, and frequently overlapping rules. Rules are considered redundant, if a more general rule or rules with the same or higher confidence values are present (Zaki 2000). Formally, for *X'* subset of *X*, a rule $X \rightarrow Y$ is redundant if,

$$confidence(X' \rightarrow Y) \geq confidence(X \rightarrow Y). \quad \text{3-11}$$

The idea is to find statistically significant rules after support and confidence pruning, in addition to redundant rule elimination. For this reason, several assumptions have been considered to rank the rules. First, the rules must be common in, at least, 60% of the individuals. Second, the higher the confidence the more likely it is for Y to be present in transactions that contain X. According to this, a support value of 0.6 and a confidence value of 0.8 were used to generate rules in this thesis. This states that 60% of individuals carry the SNPs in the LHS and RHS of a particular rule together, and those who have the SNPs in the LHS also have the SNPs in the RHS 80% of the time.

3.5.4 Rule visualisation

After significant rules are selected, they can be visualized using tools or packages such as the *arulesViz* package in R (Hahsler & Chelluboina 2011), which uses graph-based visualization (Klemettinen et al. 1994) to construct a genetic interaction network. This network is used to characterise epistatic

interactions within the network. Utilising networks provides an intuitive and interpretable framework for studying and visualising complex relationships between large numbers of biological features (Strogatz 2001). In other words, network plots provide an alternative way of inspecting the rules in a graphical way. It helps, for example, to see which rules form clusters when sharing common SNPs.

The rules are linked together forming a large graph consisting of vertices or nodes with various sizes and colours, and edges with the items label. Unlike the normal graph where the node is usually the item and edges are the relationship between items, this graph has a different notation. Nodes represent the rule $X \rightarrow Y$ while items or SNPs are labelled and represented by an arrow (edge), indicating the direction of the association. Items (SNPs) are connected to nodes via: 1) an outward arrow if they are the antecedents of the rule $X \rightarrow Y$; 2) an inward arrow if they are the consequent. In addition to this outward and inward notation, edges are represented in distinct colours, where a soft purple arrow indicates antecedent and a red arrow indicates the consequent of the rule. The size of the node indicates the support value of rule $X \rightarrow Y$ while the colour indicates the lift value of rule $X \rightarrow Y$. Bigger nodes indicate a higher support value while red darker nodes indicate higher lift values. In other words, more frequent rules present bigger nodes or vertices, whereas rules with higher lift values are represented with more intense red colours.

In Figure 3-6, examples of graph-based visualization for a rule based on the above notation is depicted. Figure 3-6 (a) shows the simplest representation of a rule while Figure 3-6 (b) depicts a graph-based visualisation with two rules.

In this study, the SNP-SNP interactions are represented using this graph notation. Significant association rules are visualized in an epistatic network. The graph allows us to conclude on potential pathways which are not just limited to SNP name, but also information about risk allele. However, this scheme can generate very large and complex graphs when the number of interactions is large. Thus, it is recommended to select a reduced number of possible interactions in the network graph (i.e. top 10 rules) to reduce the complexity of the network graph.

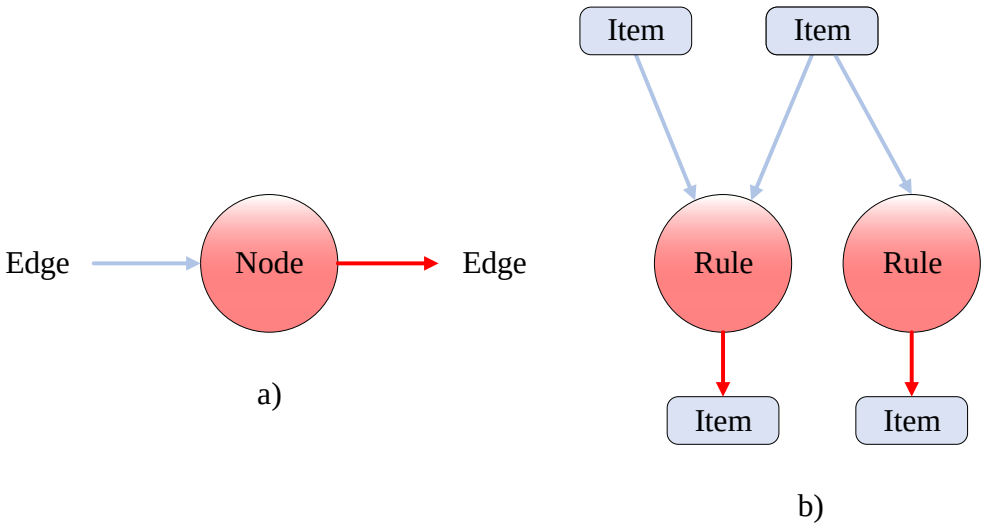


Figure 3-6: (a) Rule notation. (b) Basic example of graph-based visualisations

3.6 Multilayer Perceptron Neural Network (MLP)

In this study, a multi-layer feedforward neural network is implemented based on the formal definitions in (Ng 2011), for classification analysis. This section begins with a discussion on feed forward ANN and backpropagation before discussing how autoencoders are used in this study.

MLPs in this study use labelled training samples $(x^{(i)}, y^{(i)})$ from case-control genetic data to train the network for supervised learning tasks. A complex non-

linear hypothesis $h_{W,b}(x)$ is defined using a feed forward ANN, with parameters W,b fitted to the data. Taking a set of labelled samples $\{x_1, x_2, \ldots x_n\}$ and a bias unit b (+1 intercept term) as input, single computational units or neurons output

$$h_{W,b}(x) = f(W^T x) = f\left(\sum_{i=1}^{n} W_i x_i + b\right) \qquad 3\text{-}12$$

where f: $\mathbb{R} \mapsto \mathbb{R}$ represents the activation function. Figure 3-7 provides a simple example for a single neuron with the elements described above. Each input is connected to an output node by a weighted link. Activation functions, such as the sigmoid function, hyperbolic tangent (tanh) and rectifier linear unit (ReLU) are common activation functions used in many neural network configurations. However, rectifier functions have shown faster learning compared to sigmoid or tanh (Glorot et al. 2011). In the experiments conducted in this thesis, the selection of activation functions is determined using random search optimization methods to simplify model configuration (Bergstra & Bengio 2012).

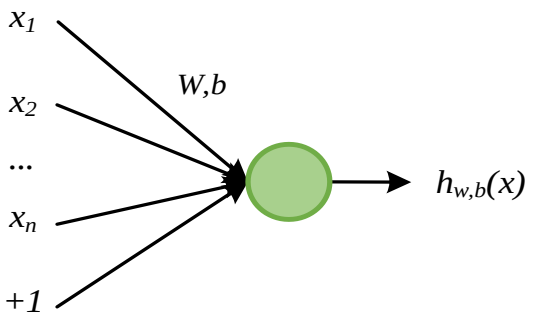


Figure 3-7: Single computational unit or neuron

By connecting multiple single neurons, an ANN architecture can be constructed so that the output of a neuron becomes the input to another one. An

example of such an architecture used in classification tasks is depicted in Figure 3-8.

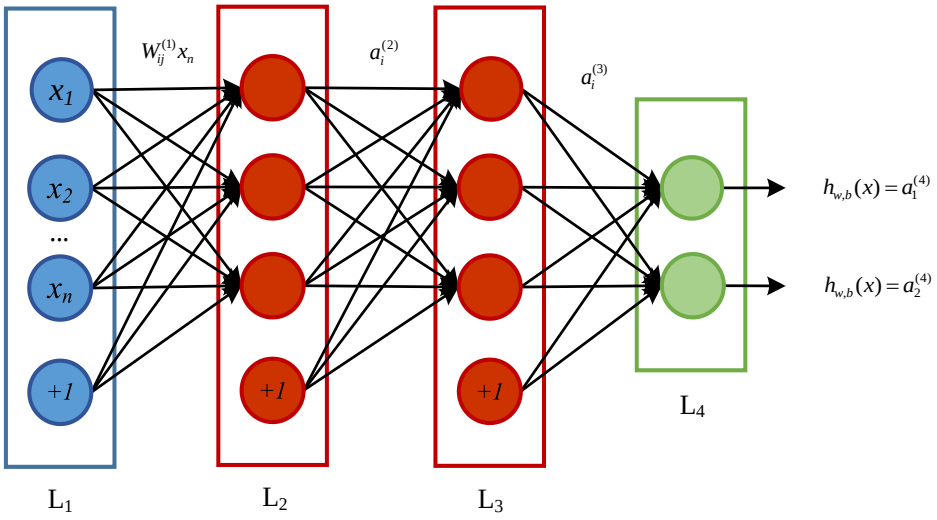


Figure 3-8: MLP network with an input layer L1, two hidden layers L2 and L3 and an output layer L4 with two output units

Therefore, input, hidden and output layers make up the network structure where n_l represents the total number of layers and L_l a particular layer (i.e. L_1 the input layer and Ln_l the output layer). Several parameters are described in the network and summarised in Table 3-2. The parameter $W_{ij}^{(l)}$ denotes the weight of the connection between the j^{th} neuron in layer l, and the i^{th} neuron in layer l+1. The bias unit $b_i^{(l)}$, associated with neuron i in layer l+1, is introduced to counteract the problem associated with input patterns that are zero. The number of nodes in a particular layer (L_l) is denoted by s_l without taking $b_i^{(l)}$ into consideration.

Parameter	Description
n_l	Total number of layers in the network.
l	Denotes a layer.
L_l	A particular layer, i.e. L_1 is the input layer and Ln_l the output layer.
$W_{ij}^{(l)}$	Denotes the weight of the connection between neuron j in layer l, and neuron i in layer l+1.
$b_i^{(l)}$	Bias or intercept term associated with neuron i in layer l+1.
s_l	Number of neurons in layer l (not including the bias unit)
$a_i^{(l)}$	Activation of unit i in layer l.

Table 3-2: Network parameters description

Additionally, the activation or output value of node i in layer l, denoted as $a_i^{(l)}$, is equal to the total weighted sum of inputs (including the bias term), represented as $f(z_i^{(l)})$ and is defined as:

$$a_i^{(l)} = f\left(\sum_{j=1}^{n} W_{ij}^{(l)} a_j^{(l)} + b_i^{(l)}\right) \quad 3\text{-}13$$

$$a_i^{(l)} = f(z_i^{(l)}) \quad 3\text{-}14$$

It is possible then to rewrite Equation 3-13 in a matrix form using a weight matrix W^l for each layer, l. The activation vector a^l is also defined using activation components $a_i^{(l)}$. Given that the values from inputs are denoted by $a^{(1)} = x$ and the activation for layer l is $a^{(l)}$, the activation in the output layer (l+1) can be computed. Thus, a more compact vectorised form of Equation 3-13 can be defined as:

$$z^{(l+1)} = W^{(l)}a^{(l)} + b^{(l)} \quad \text{3-15}$$

$$a^{(l+1)} = f(z^{(l+1)}) \quad \text{3-16}$$

Equations 3-15 and 3-16 can be used to compute the output of the network, successively calculating all the activations in layer L_2, then L_3 and so on up to the output layer Ln_l. Learning using the proposed FNN-based model (MLP) is performed by adjusting the connection weight values to minimise the prediction error on training data. An example of how to compute activations in a neural network is presented in Appendix A for explanatory purposes.

The neural network hypothesis is defined as $h_{W,b}(x)$ based on a given set of fixed parameters W,b. The neural network is trained using training samples ($x^{(i)}$, $y^{(i)}$) where $y^{(i)} \in \mathbb{R}^2$. The parameter x is a vector of input features representing individuals while outputs for the two class labels (obese or non-obese in this study) are represented using y. The weight and bias parameters can be learned by minimising the cost function using gradient descent. For a single training sample (x,y), the cost function is defined as one-half of the squared differences between $h_{W,b}(x)$ and y:

$$J(W,b;x^{(i)},y^{(i)}) = \frac{1}{2}\left\|h_{W,b}(x) - y\right\|^2 \quad \text{3-17}$$

Therefore, given a training set $\{(x^{(1)}, y^{(1)}),\ldots, (x^{(m)}, y^{(m)})\}$ of m samples, the neural network is trained using gradient descent and the overall cost function is defined as:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J(W,b;x^{(i)},y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2$$
$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}||h_{W,b}(x^{(i)}) - y^{(i)}||^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}(W_{ji}^{(l)})^2 \quad \text{3-18}$$

where the first term is an average sum of squared errors and the second, a weight decay or regularization term that helps prevent overfitting by reducing the magnitude of the weights. Hence, relative importance of the first and second terms is controlled by the weight decay parameter λ.

The idea is to minimise *J(W,b)* as a function of *W* and *b*. Parameters $W_{ij}^{(l)}$ and $b_i^{(l)}$ are randomly initialised to values close to zero to train the ANN, as this helps prevent hidden layer neurons learning the same function of the input. Following random initialization, gradient descent updates of *W,b* are achieved as follows:

$$W_{ij}^{(l)} := W_{ij}^{(l)} - \alpha\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b),$$
$$b_i^{(l)} := b_i^{(l)} - \alpha\frac{\partial}{\partial b_i^{(l)}}J(W,b) \quad \text{3-19}$$

where α is the learning rate.

The adjustment of connection weights is conducted using a backpropagation algorithm, based on the amount of error associated with the outputs of the network in comparison with the expected result (cost function). The goal of the backpropagation algorithm (see Algorithm 3-2) is to efficiently compute the partial derivatives $\frac{\partial}{\partial W_{ij}^{(l)}}J(W,b;x,y)$ and $\frac{\partial}{\partial b_i^{(l)}}J(W,b;x,y)$ of the cost function *J(W,b;x,y)* for a single sample with respect to any weight or bias in the network (Rumelhart et al. 1986).

Backpropagation

1: Perform a forward pass and compute activations for L_2, …, Ln_l

2: **for** output unit *i* in layer n_l, **do**

3: $\delta_i^{(n_l)} = \frac{\partial}{\partial_{z_i}^{(n_l)}} \frac{1}{2} ||y - h_{W,b}(x)||^2 = -(y_i - a_i^{(n_l)}) \cdot f'(z_i^{(n_l)})$

4: **end for**

5: **for** $l = n_l - 1$, …, 2, **do**

6: **for** node *i* in layer *l*, **do**

7: $$\delta_i^{(l)} = (\sum_{j=1}^{S_l+1} W_{ji}^{(l)} \delta_j^{(l+1)}) f'(z_i^{(l)})$$

8: **end for**

9: **end for**

10: Compute the desired partial derivatives:

11: $\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x,y) = a_j^{(l)} \delta_i^{(l+1)}$

12: $\frac{\partial}{\partial b_i^{(l)}} J(W,b;x,y) = \delta_i^{(l+1)}$

Algorithm 3-2: Backpropagation algorithm

The backpropagation algorithm starts by performing a feed-forward pass to compute all the activations $a_i^{(l)}$, including the output $h_{W,b}(x)$, across the network. An error term $\delta_i^{(l)}$ is calculated for each node *i* in layer *l* to quantify the node's contribution to errors in the output. The error term $\delta_i^{(n_l)}$ for an output node (n_l is the output layer), measures the difference between the activation and the true target value for an output node of the network. Conversely, hidden units compute $\delta_i^{(l)}$ by means of a weighted average of the error terms of the nodes that use $a_i^{(l)}$ as input.

The derivatives for the overall cost function can be calculated once the partial derivatives have been computed. Hence:

$$\frac{\partial}{\partial W_{ij}^{(l)}} J(W,b) = \left[\frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial W_{ij}^{(l)}} J(W,b;x^{(i)},y^{(i)}) \right] + \lambda W_{ij}^{(l)},$$
$$\frac{\partial}{\partial b_i^{(l)}} J(W,b) = \frac{1}{m} \sum_{i=1}^{m} \frac{\partial}{\partial b_i^{(l)}} J(W,b;x^{(i)},y^{(i)}) \qquad 3\text{-}20$$

where weight decay affects W but not b.

Next, the gradient descent algorithm is described (see Algorithm 3-3), where $\Delta W^{(l)}$ is a matrix with dimension equal to $W^{(l)}$, and $\Delta b^{(l)}$ is a vector with dimension equal to $b^{(l)}$. Gradient descent uses the derivatives to compute the adjustments to be made to the weights of the network. Gradient descent is used repeatedly to reduce the cost function $J(W,b)$ when training the neural network used in this study for classification purposes.

Gradient Descent

1: Set $\Delta W^{(l)} := 0$, $\Delta b^{(l)} := 0$ (matrix/vector of zeros) for all l.

2: **for** $i = 1, \ldots, m$, **do**

3: Use backpropagation to compute $\nabla_{W^{(l)}} J(W,b;x,y)$ and $\nabla_{b^{(l)}} J(W,b;x,y)$

4: Set $\Delta W^{(l)} := \Delta W^{(l)} + \nabla_{W^{(l)}} J(W,b;x,y)$

5: Set $\Delta b^{(l)} := \Delta b^{(l)} + \nabla_{b^{(l)}} J(W,b;x,y)$

6: **end for**

7: Update the parameters:

8: $W^{(l)} := W^{(l)} - \alpha \left[\left(\frac{1}{m} \Delta W^{(l)} \right) + \lambda W^{(l)} \right]$

9: $b^{(l)} := b^{(l)} - \alpha \left[\frac{1}{m} \Delta b^{(l)} \right]$

Algorithm 3-3: Gradient Descent algorithm

The learning process is made using the back-propagation algorithm which uses gradient descent to adjust the connection weights between neurons. This is performed to reduce the value of the error function. Predicted and actual values are compared to compute the value of the predefined error function. This information is supplied to the network and the connection weights are adjusted. Finally, this process is repeated according to the number of epochs until the error rate is sufficiently small.

Since binary classification is conducted, the output layer in the network will have two neurons, where each output value represents a probability value between 0 and 1. Therefore, two neurons will compose the output layer since classification of obese and non-obese individuals is performed in this thesis.

Classification tasks carried out in this thesis using an MLP model required different network configurations depending on the experiment conducted. In this sense, the performance produced in different classification tasks is obtained using a variable number of input neurons. The number of hidden layers was fixed to two with a variable number of neurons. The output layer comprises two output nodes. These configurations will be discussed in more detail in the results chapter. Hyper-parameters in this thesis were selected using random search optimization methods to simplify model configuration.

3.7 Stacked Autoencoders (SAE)

Based on the previous definition of a multilayer feedforward ANN, AE and SAE are used in this study to learn the epistatic relationships between filtered SNPs in the rules and produce a significant smaller input feature space to initialise the

weights of a multi-layer feedforward ANN (MLP) classifier. This latent information is representative of the epistatic interactions that occur between SNPs.

A basic AE is a three-layered neural network that applies backpropagation to learn an output $\hat{x}$ that is similar to the input x. Hence, an AE tries to learn a function $h_{W,b}(x) \approx x$, given a set of unlabelled training samples $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\}$, where $x^{(i)} \in \mathbb{R}^n$. An example of a single layer AE is illustrated in Figure 3-9, where the first and the third layers are the input and the reconstruction or output layer with 5 units, respectively. The second layer or hidden layer aims to generate the deep features by minimizing the error between the input vector and the reconstruction vector. Thus, an AE is a neural network with a single hidden layer composed by two parts, an encoder and a decoder.

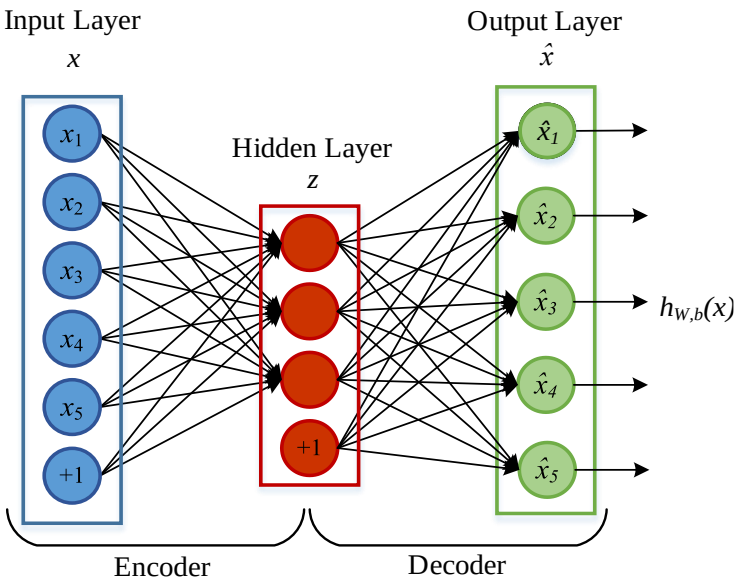


Figure 3-9: Single layer Autoencoder. The model learns a hidden feature z from input x by reconstructing it on x̂

The output of the encoder z is a reduced representation of x used by the decoder to reconstruct the original input x. An autoencoder with a code

dimension lower than the input dimension is termed undercomplete (Goodfellow et al. 2016). This forces the autoencoder to capture the most prominent features of the training data. First, the encode phase maps input data into a feature vector z so that, for each sample $x^{(i)}$ from the input set $\{x^{(1)}, x^{(2)}, x^{(3)}, \ldots\}$, we have

$$z^{(i)} = f(W^{(1)}x^{(i)} + b^{(1)}) \quad \text{3-21}$$

while in the decode phase, the decoder reconstructs the input x, producing a reconstructed space $\hat{x}$ defined by

$$\hat{x}^{(i)} = f(W^{(2)}z^{(i)} + b^{(2)}) \quad \text{3-22}$$

where $W^{(1)}$ and $W^{(2)}$ represent the input-to-hidden and the hidden-to-output weights respectively, $b^{(1)}$ and $b^{(2)}$ represent the bias of hidden and output neurons, whereas $f(\cdot)$ denotes the activation function. As indicated in the previous section, several alternatives for $f(\cdot)$ exist, including the sigmoid, tanh and rectifier linear functions. In this thesis, the best activation function was selected by a random search according to each experiment conducted.

Parameters $W^{(1)}$, $W^{(2)}$, $b^{(1)}$ and $b^{(2)}$ in the AE can be learned by minimising the reconstruction error

$$J(W,b;x,x) = \frac{1}{2}\left\| h_{W,b}(x) - x \right\|^2 \quad \text{3-23}$$

This is a measurement of discrepancy between input x and reconstructed $\hat{x}$ with respect to a single sample. For a training set of m samples, the cost function of an autoencoder is as shown below:

$$J(W,b) = \left[\frac{1}{m}\sum_{i=1}^{m} J\left(W,b;x^{(i)},x^{(i)}\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(l)}\right)^2$$
$$= \left[\frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}\left\|h_{W,b}\left(x^{(i)}\right)-x^{(i)}\right\|^2\right)\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_l+1}\left(W_{ji}^{(l)}\right)^2 \quad \text{3-24}$$

where *m* denotes the overall training set size and the square error is used as the reconstruction error for each training sample. The second term remains as explained in the previous section and represents a weight decay term introduced to decrease the magnitude of the weights and aid to prevent overfitting. Equation 3-24 can be minimised using stochastic gradient descent as described in the previous section (see Section 3.6).

The AE will learn any structure present in the data. Basic AEs typically learn a low-dimensional representation as similarly performed by principal component analysis (PCA). The hidden layer is forced to summarise the data, to compress it. After training an AE, the output layer (reconstruction) and its parameters are discarded, and the learned reduced features remain in the hidden layer which can then be used for classification or as the input of an extended network to extract deeper features. The strength of AEs lies in this type of reconstruction-oriented training that only uses information in the hidden layer which represents learned features from the input. Therefore, the learned non-linear transformation, defined by weights and biases, describes a feature extraction step.

By stacking a sequence of AEs layer by layer, an SAE can be constructed (Bengio et al. 2007). Once a single layer AE has been trained, a second AE can be trained using the hidden layer from the first AE as shown in Figure 3-10. By repeating this procedure, it is possible to create SAEs of arbitrary depth. The

first single layer AE maps inputs into the first hidden vector. Once the first layer AE is trained, its reconstruction layer is removed, and the hidden layer becomes the input layer of the next AE.

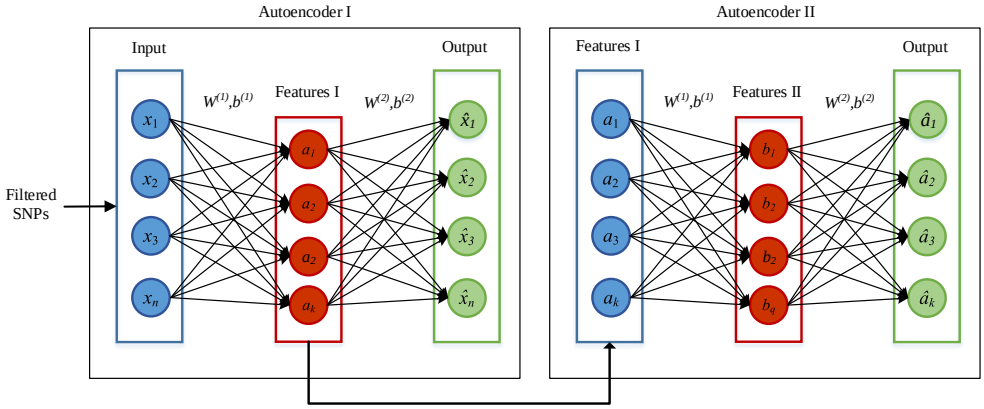


Figure 3-10: Example of SAE formed by two single AEs

In this study, AEs are stacked to enable greedy layer-wise learning where the l_{th} hidden layer is used as input to the l+1 hidden layer in the stack. The results produced by the SAE are utilised to pre-train the weights for a proposed MLP, rather than randomly initialising the weights to small values to classify extreme cases of obesity and normal individuals. Greedy layer-wise pre-training helps the model initialise the parameters near to a good local minimum and transform the problem space to a better form of optimisation (Bengio et al. 2007). By adopting this approach, it is expected to achieve smoother convergence and higher overall performance in the classification task (Danaee et al. 2017).

An SAE with 2,000, 1,000, 500 and 50 hidden neurons in each hidden layer was considered in the experiments conducted in this thesis (See Figure 3-11). Consequently, selected layers are used as input features for classification using an MLP. In Figure 3-11 an instance of the SAE proposed in this thesis connected with an MLP is depicted. The classification scheme represented is composed of 8 layers: one input layer, four hidden layers from AEs, two hidden layers from

the MLP, and an output layer. This represents the network configuration where SNPs are compressed progressively from 2,465 to 50 neurons.

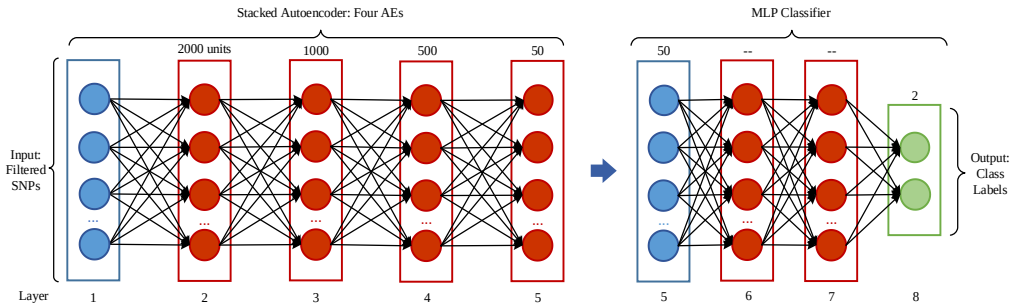


Figure 3-11: Instance of proposed SAE connected with an MLP

The goal of the proposed SAE architecture is to extract a mapping that decodes the input (set of SNPs) as closely as possible without losing significant SNP-SNP patterns. The encoder decreases the dimensionality of the original data (SNP set) stack by stack, leading to a reduction in noise while preserving important information patterns (Bengio et al. 2007). Consequently, AEs are used in this study to gradually extract deep features representative of obesity epistasis.

3.8 Performance Assessment

3.8.1 Model validation

For classification model assessment, a three-way data split procedure is utilised (training, validation and test) - 60% for training, 20% for validation and 20% for model testing. The dataset is thus partitioned based on 3:1:1 ratio, using a 60/20/20 split as recommended in (Lever et al. 2016). This resulted in 1,198 samples to be used for training, 399 for validation and 399 for testing. The training set is used to calculate the optimal weights and bias in the network using

the backpropagation algorithm and the validation set is used to identify optimal hyper-parameters to minimise overfitting. If performance using the training dataset increases but performance using the validation dataset decreases, it is likely that overfitting is occurring, and the training process should stop. Early stopping strategies can be adopted to stop training when training and validation losses begin to diverge. Introducing weight penalties such as L1 and L2 regularisation can also help to reduce overfitting. Finally, the test set is employed to assess the performance of the final model. The testing process is conducted only after model tuning and regularization parameters have been optimised using the model with the best performance on the validation set. The steps necessary to train, validate and test a classifier can be summarised according to (Dougherty 2013):

a) Split available data into training, validation and test sets.
b) Architecture and training parameter selection.
c) Model training using training set.
d) Model evaluation using validation set.
e) Repeat steps (b)-(d) testing different architectures and training parameters.
f) Best model selection.
g) Best model assessment with test set.

These steps are valid if the selection of hyper-parameters is manually managed. Since a random search has been considered in this thesis, step (e) from the procedure proposed by (Dougherty 2013) is replaced by optimal hyper-parameter selection using random search.

3.8.2 Binary class performance evaluation

Model performance in this thesis is measured using a range of numerical and graphical approaches (Salari et al. 2014; Fergus, Curbelo et al. 2018).

The performance metrics for binary classification are derived from a 2x2 contingency table (matrix) to calculate sensitivity (SE) and specificity (SP) - among other metrics - where rows in the table indicate the actual class and columns the predicted class, as shown in Table 3-3.

		Predicted class		
		Positive	*Negative*	
Actual class	*Positive*	**TP**	**FN**	TPR
	Negative	**FP**	**TN**	TNR
		PPV	NPV	

Table 3-3: Conventional data layout for the 2x2 confusion matrix

The confusion matrix has four terms: true positive (TP), false positive (FP), true negative (TN) and false negative (FN); which are utilised to compare the class labels assigned by a classifier, against the desired correct class labels. From this notation, TP and TN represent the number of positive and negative cases that are classified correctly, whereas the positive and negative cases falsely classified are denoted by FP and FN. In obesity prediction, an individual can be classified as obese or normal. This results in four possible combinations. Predicting obese when the individual is obese (TP) and predicting non-obese when he/she is not obese (TN). If the prediction says obese when the individual is healthy, it is considered an FP (Type I error in statistics), while predicting

someone is healthy when they are obese is considered an FN (Type II error). Below, several performance metrics (Hoens & Chawla 2013) are presented using the confusion matrix:

$$\text{Accuracy (Acc)} = \frac{TP + TN}{N}; \quad N = TP + TN + FP + FN \tag{3-25}$$

$$\text{Sensitivity (SE)} = \text{Recall} = TPR = \frac{TP}{TP + FN} \tag{3-26}$$

$$\text{Specificity (SP)} = TNR = \frac{TN}{TN + FP} \tag{3-27}$$

$$FPR = (1 - Specificity) = \frac{FP}{FP + TN} \tag{3-28}$$

$$FNR = \frac{FN}{FN + TP} \tag{3-29}$$

$$\text{Precision} = PPV = \frac{TP}{TP + FP} \tag{3-30}$$

$$NPV = \frac{TN}{FN + TN} \tag{3-31}$$

$$\mathrm{F_1} = 2 \times \left(\frac{\mathrm{PPV} \times \mathrm{TPR}}{\mathrm{PPV} + \mathrm{TPR}} \right) \tag{3-32}$$

In this study, SE or true positive rate (TPR) is used to quantify how effectively the classifiers correctly recognise actual positive cases (i.e. obese individuals), whilst SP or true negative rate (TNR) represents the classifier's ability to correctly recognise actual negative cases (i.e. non-obese individuals). Moreover, the proportion of actual negatives predicted as positives is termed

false positive rate (FPR) while a false negative rate (FNR) is the proportion of items wrongly identified as negative out of total true positives. The proportion of predicted positives that are actual positives is called precision or positive predictive value (PPV). Conversely, a negative predictive value (NPV) is the proportion of predicted negatives that are actual negatives. Classification accuracy is commonly utilised to evaluate the quality of predictive models (it represents the percentage of total items correctly classified). However, this performance measure could be misleading particularly in large class imbalance datasets, since overall accuracy varies with class frequency (Hoens & Chawla 2013; Salari et al. 2014).

The receiver operating characteristic (ROC) curve is a standard technique used as a graphical performance measure to summarise the predictive performance of binary classification models (Hoens & Chawla 2013; Fawcett 2004). The ROC curve plots the TPR against false positive rate (FPR) measurements produced by a classification model, where each point on the ROC curve corresponds to a classifier, as depicted in Figure 3-12.

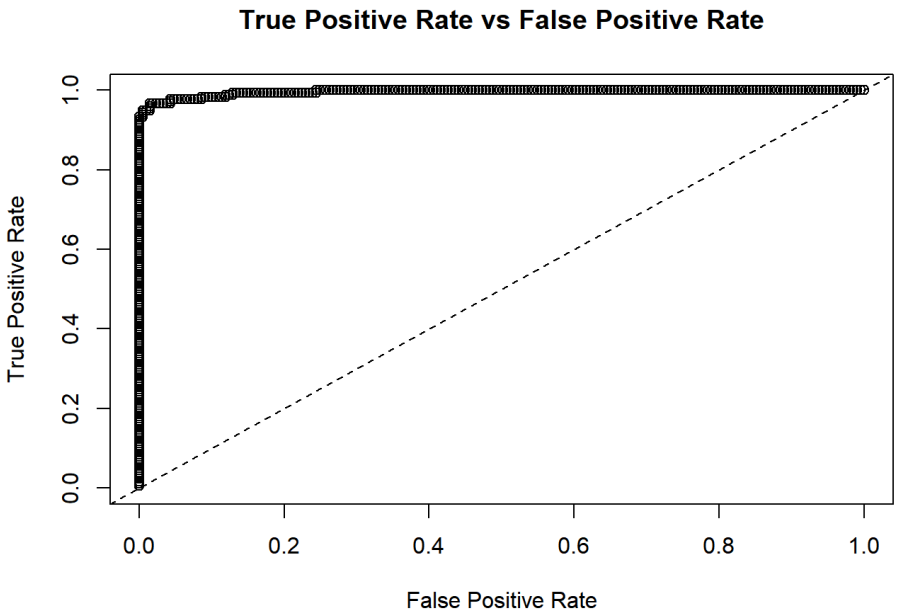


Figure 3-12: ROC curve example

Therefore, an ideal predictive model should have an ROC curve closer to the top left corner of the ROC space (see Figure 3-12), which indicates that the model is able to accurately classify both positive and negative classes. In contrast, models with predictive performance close to the ROC curves diagonal line represent classifiers that classify randomly.

Additionally, the ROC is commonly summarised by a single measure known as the area under the ROC curve (AUC). AUC measures the probability that test values from a randomly selected pair of case-control samples are correctly ranked and is thus a convenient global measure for the quantification of classification accuracy. For a classifier that perfectly classifies, the AUC will be 1, whereas a classifier that randomly assigns labels, will be 0.5 (de Ridder et al. 2013).

The F1 metric, also known as the F-score or F-measure, takes precision and recall into account and represents the harmonic mean between the two (See Equation 3-32).

Other performance measures such as Logarithmic Loss (logloss), mean square error (MSE) and Gini coefficient (Gini) are also utilised to provide an objective classification measure of performance (Fergus, Curbelo et al. 2018; Curbelo, Fergus, Curbelo, et al. 2018), in addition to those introduced above (SE, SP, AUC and ROC). Logarithmic loss (logloss) is a classification loss function which provides a quantification of accuracy for a classifier by penalising false classifications. Minimising the logloss is correlated with accuracy; as one increases the other decreases. The logloss for a binary class classifier is defined by:

$$\text{Logloss} = -\frac{1}{N}\sum_{i=1}^{N}\left[y_i \log(p_i) + (1-y_i)\log(1-p_i)\right] \qquad 3\text{-}33$$

where N is the number of samples, *pi* is the probability of the i_{th} sample belonging to class C1 and y_i is the actual label of the i_{th} sample, which could be either 0 or 1. In the case of misclassifications, logloss values are progressively larger, whereas logloss for models that classify all instances correctly will be 0. Thus, the robustness of the model increases by minimising the logloss value.

A common performance metric utilised to measure the average sum of the square difference between actual values and predicted values is the mean squared error (MSE). The standard definition of MSE is defined as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - \hat{y}^{(i)})^2 \quad \text{3-34}$$

where $(y^{(i)} - \hat{y}^{(i)})$ is termed the residual, and the aim of the MSE is to minimise the residual sum of squares. MSE values close to 0 indicate that the model correctly classifies all class instances. Progressively larger values of MSE indicate the proportion of misclassifications that occur. Hence, the lower the MSE value the better.

The Gini coefficient is occasionally used in binary classification studies and represents the area between the diagonal line and the ROC curve:

$$Gini = 2 * AUC - 1 \quad \text{3-35}$$

The Gini coefficient quantifies dispersion among values of a frequency distribution. Gini values close to 1 indicate a good model, whereas a coefficient of 0 indicates that the features (SNPs) have no predictive capacity.

3.9 SNP to Gene Context

To report the context of the SNPs identified, the SNPnexus tool is utilised (Dayem Ullah et al. 2012; Dayem Ullah et al. 2013; Dayem Ullah et al. 2018) although others are available (Hinrichs et al. 2016). It was developed to facilitate the selection of functionally relevant SNPs in large-scale genotyping studies of multifactorial diseases, via single or batch queries using dbSNP identifiers or genomic coordinates. This tool uses the ensemble gene annotation system as a reference to conduct the annotation for gene-overlapped and intergenic variants, through a query and output interface freely accessible on the web (Dayem Ullah et al. 2018). Variant-centric functional annotations are

reported from different sources (primary annotations datasets) for GRCh38/hg38, GRCh37/hg19 and, NCBI36/hg18 assemblies. However, in the consultations made in this thesis, the GRCh37/hg19 assembly was used as it matches with the characteristics of the MyCode dataset. In addition to genomic mapping or gene annotation, SNPnexus also reports phenotype and disease association, retrieved from the queried SNPs and several databases (Dayem Ullah et al. 2018), including The Genetic Association Database (Becker et al. 2004). This can be used to mine any disease or phenotype related information about the queried SNPs that have been reported in the literature.

SNPnexus is thus used to map genes to SNPs identified in the experiments conducted in this thesis to help understand the functional role of SNPs and their impact on health and disease.

3.10 Chapter Summary

In this chapter, a novel methodology has been described. First, a subset of SNPs after QC and association analysis is selected. Epistatic analysis of the filtered set of genetic variants is explored using ARM and deep learning SAE, where ARM provides model interpretation for the extracted features using an SAE. Finally, classification analysis is performed using an MLP neural network to evaluate the discriminatory capacity of identified features. Thus, standard GWAS analysis has been combined with the novel approach presented to extract the epistatic interactions between SNPs in obesity case-control observations (identified by ARM) using a deep learning SAE. These two techniques combined form a tight correlation such that the manipulation of support and confidence parameters affects SAE layers and the level of epistatic information

extracted, which in turn affects the classification accuracy of case-control classification tasks. The different techniques utilised have been combined to produce SAERMA - an algorithm for the analysis and interpretation of interactions between SNPs in case-control GWAS.

In the proposed methodology, the use of association rule mining provides an intuitive framework for studying and visualizing complex relationships between a large number of biological entities. In contrast, AE is applied for layer-wise training while SAE is adopted as the corresponding deep neural network architecture. AEs are a powerful approach capable of extracting both linear and non-linear relationships inherent in the input data. It has been proved that by reducing dimensionality gradually, a multi-layered architecture using SAEs may extract valuable patterns from data without losing important information (Bengio et al. 2007). Features extracted using the SAE are then used to pre-train an MLP in different scenarios which will be reported in the next chapter.

Finally, detailed annotations, from SNPs identified by association analysis and rule mining experiments, are obtained utilising the SNPnexus tool while functional validation is provided via gene set enrichment analysis.

Chapter 4. RESULTS

4.1 Introduction

This chapter implements the methodology previously discussed and presents the results obtained. The results are reported in six experimental sections:

1. Quality control.
2. Association analysis (Statistical filtering).
3. Logistic regression classification (Gold Standard) using SNPs with P-value < 10^{-2}.
4. MLP classification using SNPs with P-value < 10^{-2}.
5. SAE-based classification using non-linear SNP-SNP interactions. Again, using P-values < 10^{-2}.
6. The proposed SAERMA approach posited in this thesis.

The results for each of the steps conducted in the proposed methodology are presented in this chapter and discussed later in Chapter 5.

4.2 Quality Control

This section includes the results after applying QC to the MyCode data. Quality control results and processes presented in this thesis have been published in several high quality conference and journal papers (Curbelo, Fergus, Hussain, Al-Jumeily, Dorak, et al. 2017; Curbelo, Fergus, Hussain, Al-Jumeily, Abdulaimma, et al. 2017; Curbelo, Fergus, Curbelo, et al. 2018; Fergus, Curbelo et al. 2018; Curbelo, Fergus, Chalmers, et al. 2018).

Before conducting QC analysis, the number of individuals and variants per individual are shown in Table 4-1. The total number of subjects reported is slightly higher than the number provided in the Data Description section in Chapter 3 (2,193). This was due to discrepancies in the binary files (particularly in .fam file) when merging the two datasets. In particular, mismatches in sample numbers due to sample duplication. This was not reported by dbGaP in the clinical information.

Individuals	Males	Females	Cases	Controls	SNPs per individual
2,270	942	1,328	1,006	1,264	594,034

Table 4-1: Number of individuals and genetic variants before QC. Information extracted from binary files

The QC steps conducted for individuals, resulted in 295 individuals being removed. However, some of these individuals' IDs were duplicated. Thus, 273 unique individuals were removed from the main data frame for subsequent analysis. After applying individual level QC, SNP level QC was conducted. In this process, a total of 353,084 SNPs failed QC due to pruning steps explained in Chapter 2 and Chapter 3, and were thus removed from the data set.

A summary table with samples and SNPs rejected for each QC criterion is given in Table 4-2:

Criteria	n
Individual criteria	
Sex check	3
Missing Genotype and heterozygosity	43
Relatedness or duplicates	156
Population outliers	93
SNP criteria	
Missing genotype	324,244
HWE	0
MAF	28,840

Table 4-2: Summary of QC steps applied for individual and genetic variants to the MyCode dataset

A total of 240,950 variants and 1,997 individuals passed subsequent filter analysis and QC. Among the remaining phenotypes, 879 are cases and 1,118 are controls. The number of individuals and associated genetic variants that remain are summarised in Table 4-3.

Individuals	Males	Females	Cases	Controls	SNPs per individual
1,997	840	1,157	879	1,118	240,950

Table 4-3: Number of individuals and genetic variants passing filters and QC

4.3 Association Analysis

After QC, 1,997 individuals and 240,950 SNPs are used in subsequent analysis as indicated in Table 4-3. These are then used to conduct association analysis with obesity trait, as a statistical filtering strategy.

Association tests between SNP genotypes and extreme obesity were conducted under an additive model using logistic regression implemented in

PLINK version 1.09 (Purcell & Chang 2018). Logistic regression was adjusted using Genomic Control (GC) to control population structure. Correction for multiple testing was considered using Bonferroni correction but only to demonstrate standard GWAS results. Hence, analysis to assess associations of all Illumina HumanOmniExpress-12 v1.0 arrays SNPs with the obesity phenotype is performed, in which the total number of SNPs tested is adjusted using significance cut-off P-values < $2.1x10^{-7}$ following Bonferroni's correction (P-value = 0.05/240,950). The adjusted significance threshold is used here only to present standard results from association analysis. For epistasis, a modified P-value threshold that allows for a larger subset of SNPs is selected as discussed later.

In Figure 4-1 a quantile-quantile (QQ) plot depicts the relationship between the expected distribution of P-values (null hypothesis) and the observed distribution of P-values for the association test. The genomic control inflation factor λ, which measures the degree of deviation from y = x, is λ=1.0384. A value close to one suggests appropriate adjustment for potential substructure. Thus, population stratification was assessed, and standard errors were adjusted using the genomic inflation statistic (λ).

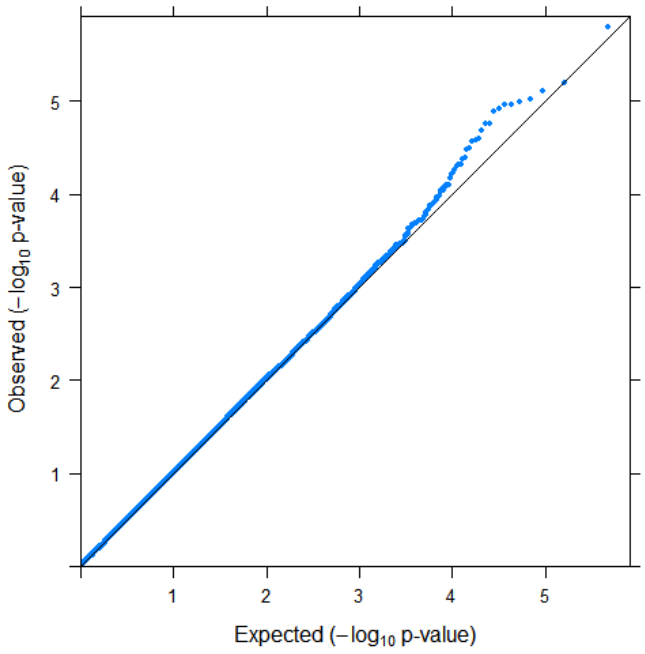


Figure 4-1: Quantile-quantile plot for association analysis using logistic regression

In Figure 4-2 the P-values for associations with extreme obesity are illustrated using a Manhattan plot. Each point represents an SNP distributed through the human chromosomes from left to right, whereas the heights correspond to the strength of the association to disease. P-values in $-\log_{10}$ scale are distributed in the *y*-axis while the physical position of the SNPs along chromosomes is represented in the *x*-axis. The smallest P-values suggest potential disease-related SNPs. The Bonferroni corrected significance threshold and the suggestive threshold of association are represented in Figure 4-2 using red and blue lines, respectively. Therefore, the red upper horizontal line indicates the threshold for genome-wide significance (P-value < 2×10^{-7}) and the lower line indicates the threshold for suggestive association (P-value < 1×10^{-5}).

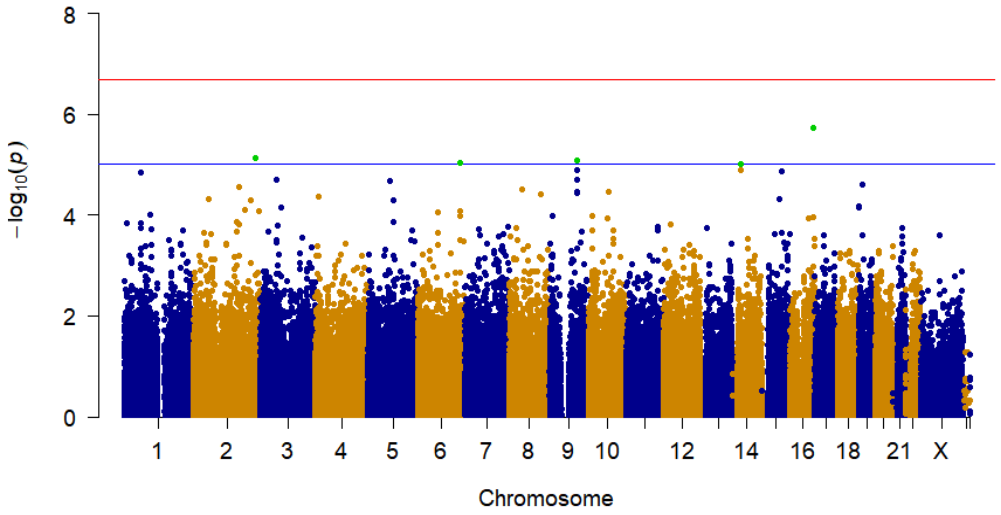


Figure 4-2: Manhattan plot of association results using logistic test adjusted GC in MyCode dataset

In Figure 4-3 the Manhattan plot for the Logistic test is represented with SNP ID annotations.

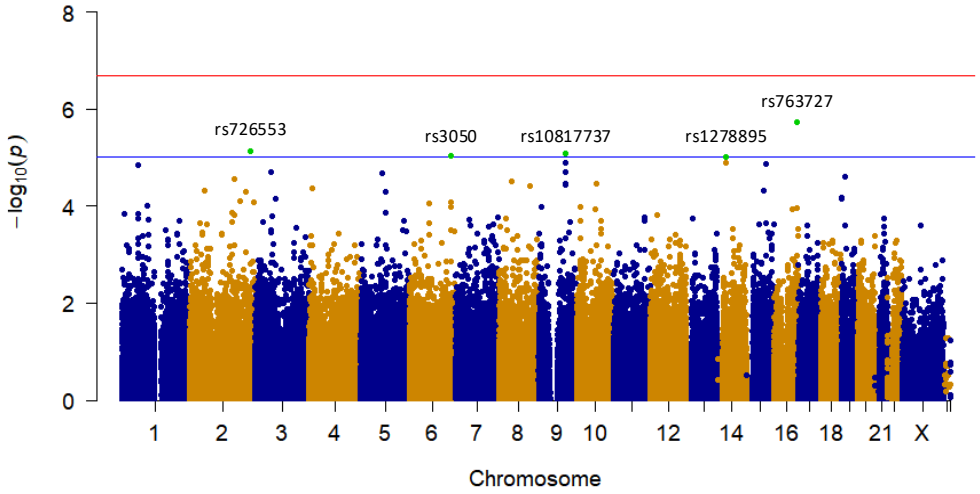


Figure 4-3: Manhattan plot for logistic test adjusted GC with SNP labels

None of the SNPs identified in logistic analysis reached the Bonferroni level of significance (P-value < $2x10^{-7}$ – red horizontal line in Figure 4-2 and Figure 4-3); however, five SNPs were suggestive of association (P-value < $1x10^{-5}$ – blue horizontal line in Figure 4-2 and Figure 4-3). All suggestive association

signals with P-value < $1x10^{-5}$ are shown in Table 4-4 where SNP information is based on the Genome Reference Consortium Human Build 37 (GRCh37) (Schneider & Church 2013). The information for each SNP includes risk allele, chromosome number, position of the SNP in the genome, SNP name, distance in bases (b) of the SNP to the closest gene (in case of no overlapping) and, the P-value of association.

SNP	Allele	Chr#	Position	Closest Gene	Distance	P-value
rs763727	A	16	83342301	CDH13	0	$1.821x10^{-06}$
rs726553	G	2	226016494	DOCK10	109,332	$7.330x10^{-06}$
rs10817737	A	9	100306267	TMOD1	0	$8.319x10^{-06}$
rs3050	A	6	150923115	PLEKHG1	0	$9.061x10^{-06}$
rs1278895	T	14	32400170	RP11-159D23.2	818	$9.979x10^{-06}$

Table 4-4: Top suggestive results (P-value < $1x10^{-5}$) obtained from association analysis in the MyCode dataset

QC analysis and successive association tests were conducted using PLINK v1.9 (Purcell et al. 2007) and the language and environment for statistical computing and graphics, R (R Development Core Team 2008). A Linux Ubuntu version 16.04 LTS based machine, with 64GiB of Memory and an Intel® Core™ i7-7700K CPU @ 4.20GHz x 8, was utilized to conduct the analysis.

4.4 Generalised Linear Model classification

The first experiment conducted after QC and association analysis involves classification tasks using the filtered SNPs obtained from association analysis (statistical filtering). Logistic regression methods are the most commonly used parametric models for the analysis of binary outcome variables and are currently the industry standard. Thus, before conducting experiments with more complex

approaches such as ANNs or SAEs, classification analysis is conducted using an extension of traditional linear models, the generalised linear model (GLM) (McCullagh 1984). This is performed using H2O's implementation of GLM (Nykodym et al. 2018).

Multiple testing is a common statistical practice used in genetic association studies. However, it has been suggested that, when conducting multiple association test in case-control studies, using Bonferroni adjustments can be too strict and may lead to missing associations that are potentially significant. Thus, in the following experiments, multiple testing based solely on Bonferroni was not taken into consideration. Instead, the results from association tests with P-value < $1x10^{-2}$ were considered (Perneger 1998; Harbron et al. 2014). The resulting outcomes may be therefore, considered as hypothesis generating.

After QC and association analysis using logistic regression (statistical filtering), four different sets of SNPs were derived using different P-value thresholds as indicated in Table 4-5. The suggestive threshold of association ($1x10^{-5}$) was considered, in addition to $1x10^{-4}$, $1x10^{-3}$, and $1x10^{-2}$. Therefore, four sets of SNPs (5, 32, 248 and 2,465 SNPs) are used to train a GLM to classify extremely obese and non-obese observations. MSE, Logloss, AUC, Gini, Sensitivity and Specificity values are used to measure the performance of each model. The data set is split randomly into training (60%), validation (20%) and testing (20%).

Set	P-value	Number of SNPs
1	$1x10^{-5}$	5
2	$1x10^{-4}$	32
3	$1x10^{-3}$	248
4	$1x10^{-2}$	2,465

Table 4-5: Four sets of SNPs selected based on different P-value thresholds

4.4.1 Regularisation parameter selection

H2O supports elastic net regularisation, which combines Lasso (L1) and Ridge (L2) penalties parametrised by the *alpha* and *lambda* parameters. These penalties are introduced to the model to avoid overfitting, reduce variance of the predictor error, and handle correlated predictors (Cook 2016). The alpha parameter controls the elastic net distribution between L1 and L2 penalties while lambda controls the amount of regularisation applied (penalty strength).

To get the best possible model, regularisation parameters alpha and lambda were tuned, and the optimal values were obtained using a random search. Based on empirical analysis, *alpha* and *lambda* values reported in Table 4-6 produced the best classification results.

P-value	Parameter	Value
$1x10^{-5}$	Alpha	0.5
	Lambda	0.00598
$1x10^{-4}$	Alpha	0.5
	Lambda	0.00204
$1x10^{-3}$	Alpha	0.5
	Lambda	0.00970
$1x10^{-2}$	Alpha	0.5
	Lambda	0.00151

Table 4-6: Regularisation parameters for classification task with GLM

Based on empirical analysis, these configurations produced the best results.

4.4.2 Classifier performance

The metric values provided in Table 4-7 and Table 4-8 describe the results for the four SNP configurations in Table 4-5, for the validation and test sets respectively. In H2O, the default confusion matrix is computed at thresholds that optimise the F1 score (F1-optimal threshold).

Hence, using optimised F1 threshold values 0.3527, 0.4532, 0.3969 and 0.6684 the results in the validation set were obtained as shown in Table 4-7 for 5 SNPs ($1x10^{-5}$), 32 SNPs ($1x10^{-4}$), 248 SNPs ($1x10^{-5}$) and 2,465 SNPs ($1x10^{-2}$) respectively.

P-value	SE	SP	Gini	LogLoss	AUC	MSE
$1x10^{-5}$	0.8723	0.2819	0.2563	0.6619	0.6281	0.2348
$1x10^{-4}$	0.6862	0.7225	0.5010	0.5865	0.7505	0.2004
$1x10^{-3}$	0.8298	0.8194	0.8081	0.3938	0.9041	0.1261
$1x10^{-2}$	0.7606	0.9383	0.8317	0.3841	0.9158	0.1150

Table 4-7: Performance metrics for validation set

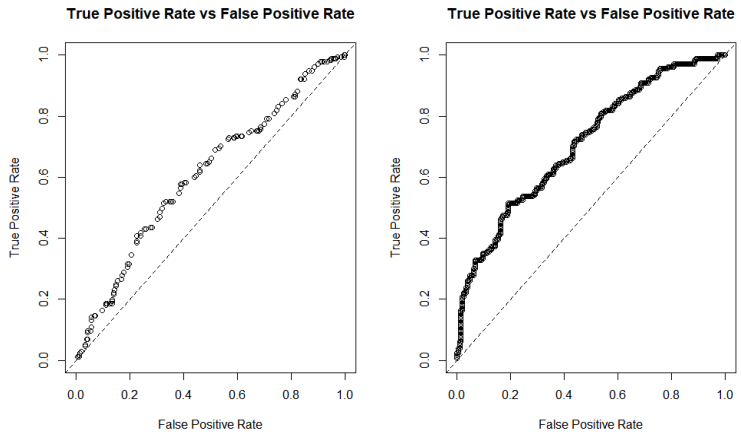
The performance metrics for the test set are shown in Table 4-8. These metric values were obtained using optimised F1 thresholds 0.2893, 0.4533, 0.2368 and 0.4665 for $1x10^{-5}$, $1x10^{-4}$, $1x10^{-3}$ and $1x10^{-2}$, respectively.

P-value	SE	SP	Gini	LogLoss	AUC	MSE
$1x10^{-5}$	0.9774	0.0909	0.2145	0.6736	0.6073	0.2404
$1x10^{-4}$	0.9548	0.2440	0.4186	0.6185	0.7093	0.2153
$1x10^{-3}$	0.9548	0.6316	0.7798	0.4119	0.8899	0.1350
$1x10^{-2}$	0.8531	0.9043	0.8725	0.3288	0.9362	0.0976

Table 4-8: Performance metrics for test set

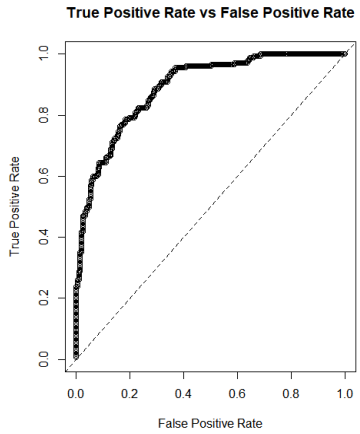
4.4.3 Model selection

The ROC curve comparison depicted in Figure 4-4 is used as a graphical performance measure to summarise the predictive performance of the GLM models. The cut-off values for the false and true positive rates using the test set are shown in each of the ROC curves for the different implemented classifiers. In this first evaluation, there is a clear deterioration in performance as the number of SNPs decreases (P-value threshold increases). Note that SNPs with conservative P-value thresholds are an indication of how significant associations are. This demonstrates the limitations of the most significant SNPs in classifying case-control samples. The highest performance was obtained with 2,465 NSPs whereas the lowest was achieved with 5 SNPs.

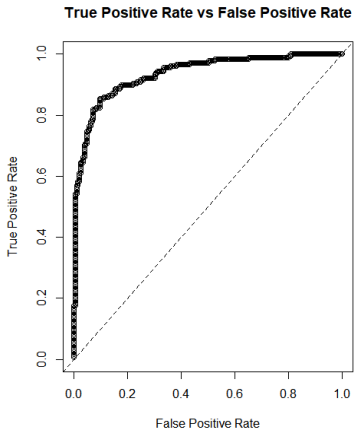


(a) ROC for P-val < 10^{-5}

(b) ROC for P-val < 10^{-4}



(c) ROC for P-val < 10^{-3}



(d) ROC for P-val < 10^{-2}

Figure 4-4: From (a) to (d) ROC curves for the test set using GLM models trained with different P-value thresholds

4.5 MLP Classification using Suggestive SNPs

Methods and results presented in this section were published in 2018 (Curbelo, Fergus, Curbelo, et al. 2018). Association results from GWAS were used to train a multi-layer perceptron neural network (MLP) framework to test the predictive capacity of statistically significant SNPs associated with the extremely obese phenotype. Various modified suggestive thresholds were considered to increase the number of SNPs for investigation based on our previous work (Curbelo, Fergus, Hussain, Al-Jumeily, Dorak, et al. 2017; Fergus, Curbelo et al. 2018).

These associations capture the linear interactions between SNPs and phenotype but not the cumulative interactions between them. H2O (Candel & LeDell 2018; Kraljevic 2018) is again used for classification analysis using the network architecture presented in Figure 3-8.

After QC and association analysis using logistic regression, four different sets of SNPs (5, 32, 248 and 2,465 SNPs) were derived using different P-value thresholds as indicated in Table 4-5. These were used to train an MLP to classify extremely obese and non-obese observations. The performance of each model is measured using MSE, Logloss, AUC, Gini, Sensitivity and Specificity values. The data set is split randomly into training (60%), validation (20%) and testing (20%).

4.5.1 Hyper-parameters selection

For each implemented classifier, the network architecture and the regularization parameters were tuned. To achieve this, random search was utilised and a maximum of 200 models were generated to obtain the best parameters. Early stopping was adopted to avoid overfitting. The model stops when the logloss value fails to improve by at least 1% (stopping tolerance) for two scoring events (stopping rounds). The adaptive learning rate ADADELTA (Zeiler 2012) was used for stochastic gradient descent optimisation, with parameters *rho* and *epsilon* set to 0.99 and $1x10^{-8}$ respectively, to balance the global and local search efficiencies. In Table 4-9 these parameters are summarised.

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	Yes
Stopping tolerance	0.01
Stopping rounds	2
Max model generated	200

Table 4-9: Tuning parameters for classification tasks with MLP

Model-specific tuning parameters summarised in Table 4-10 were considered for each model in the training phase to obtain optimal results. To prevent overfitting and to add stability and improve generalisation, Lasso (L1) and Ridge (L2) regularisation, and input dropout ratio were tuned. L1 only allows strong weights to survive, L2 prevents them from getting too big and input dropout ratio regulates the number of neurons randomly dropped in the input layer, whereas hidden dropout ratios do the same in hidden layers. Based on empirical analysis, these configurations produced the best results.

Input	Parameter	Value
$1x10^{-5}$	Activation	TanhWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$6.8x10^{-5}$
	L2	$5.1x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
$1x10^{-4}$	Activation	RectifierWithDropout
	Hidden	2
	Neurons	50
	Epochs	100
	L1	$2.2x10^{-5}$
	L2	$5.7x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5
$1x10^{-3}$	Activation	TanhWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$6.8x10^{-5}$
	L2	$5.1x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
$1x10^{-2}$	Activation	RectifierWithDropout
	Hidden	2
	Neurons	50
	Epochs	50
	L1	$8.6x10^{-5}$
	L2	$4.3x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5

Table 4-10: Model-specific tuning parameters

4.5.2 Classifier performance

The performance metrics for the validation set are provided in Table 4-11. These metric values describe the results for the four SNP configurations, 5 SNPs ($1x10^{-5}$), 32 SNPs ($1x10^{-4}$), 248 SNPs ($1x10^{-5}$) and 2,465 SNPs ($1x10^{-2}$) using optimized F1 threshold values 0.2674, 0.4463, 0.3551 and 0.8084, respectively.

P-value	SE	SP	Gini	LogLoss	AUC	MSE
$1x10^{-5}$	0.9415	0.1806	0.2556	0.6606	0.6278	0.2342
$1x10^{-4}$	0.6915	0.7490	0.5117	0.5828	0.7558	0.1987
$1x10^{-3}$	0.8564	0.819383	0.8474	0.3510	0.9237	0.1120
$1x10^{-2}$	0.9628	0.9780	0.9923	0.0883	0.9961	0.0259

Table 4-11: Validation set performance

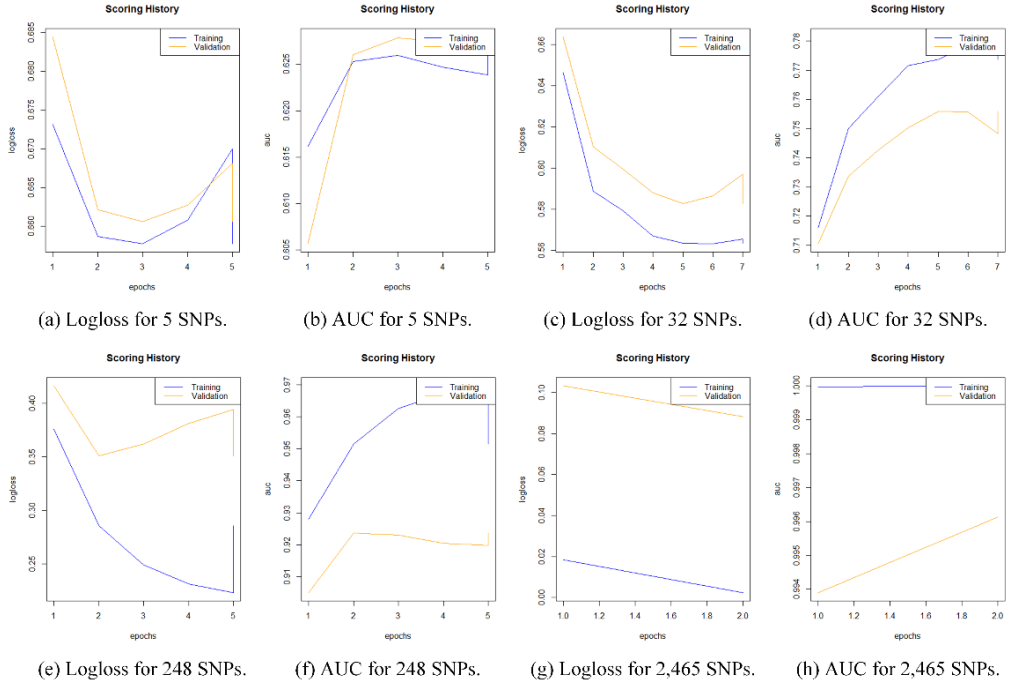
Table 4-12 shows the performance metrics for the test data using the trained models. This time, metric values were obtained using optimised F1 thresholds 0.2675, 0.2157, 0.4312 and 0.6303 for $1x10^{-5}$, $1x10^{-4}$, $1x10^{-3}$ and $1x10^{-2}$, respectively. The results are generally lower than those achieved with the validation set but, in some cases, not by much.

P-value	SE	SP	Gini	LogLoss	AUC	MSE
$1x10^{-5}$	0.9943	0.0622	0.2074	0.6750	0.6037	0.2410
$1x10^{-4}$	0.9491	0.2871	0.4331	0.6151	0.7165	0.2140
$1x10^{-3}$	0.9039	0.7942	0.8512	0.3476	0.9256	0.1094
$1x10^{-2}$	0.9548	0.9761	0.9878	0.1061	0.9938	0.0291

Table 4-12: Test set performance

Figure 4-5 helps to check if overfitting is appropriately managed. Epochs represent the inflection points where performance on the validation set starts to decrease while performance on the training set continues to improve as the model starts to overfit. The AUC plots provide useful information about early

divergence between the training and validation curves and highlight if and when overfitting occurs. Bearing in mind the scale of the plots in Figure 4-5, there are small signs of overfitting, particularly in the model trained with 248 SNPs (Plots (e) and (f) in Figure 4-5).



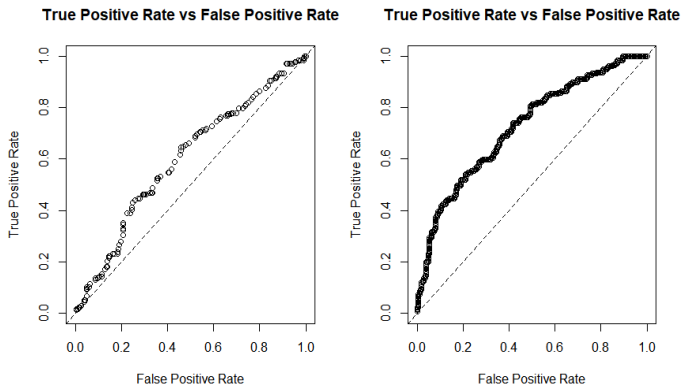
(a) Logloss for 5 SNPs. (b) AUC for 5 SNPs. (c) Logloss for 32 SNPs. (d) AUC for 32 SNPs.

(e) Logloss for 248 SNPs. (f) AUC for 248 SNPs. (g) Logloss for 2,465 SNPs. (h) AUC for 2,465 SNPs.

Figure 4-5: From (a) to (h), Logloss and AUC plots against epochs for SNPs derived from P-values $1x10^{-5}$, $1x10^{-4}$, $1x10^{-3}$ and $1x10^{-2}$ respectively

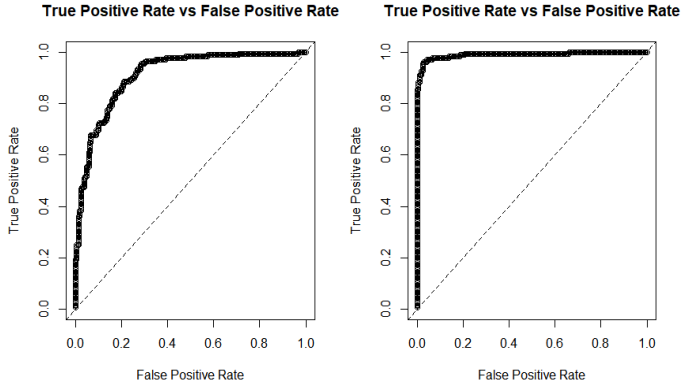
4.5.3 Model Selection

The ROC curves in Figure 4-6 show the cut-off values for the false and true positive rates using the test set. In this second evaluation, there is a clear deterioration in performance as the number of SNPs decreases (P-value threshold increases). In this instance, machine learning demonstrates the limited predictive capacity of highly ranked SNPs when discriminating between case and control samples (extremely obese and non-obese individuals).



(a) ROC for P-val < 10^{-5}

(b) ROC for P-val < 10^{-4}



(c) ROC for P-val < 10^{-3}

(d) ROC for P-val < 10^{-2}

Figure 4-6: From (a) to (d) ROC curves for test set using the MLP trained with different P-value thresholds

4.6 Epistatic interactions using Stacked Autoencoders

The methodology and results shown in this section were published in 2018 (Curbelo, Fergus, Chalmers, et al. 2018; Fergus, Curbelo et al. 2018).

In this evaluation, only SNPs with P-values lower than $1x10^{-2}$ were considered for machine learning analysis. The R H2O package is used in this research to implement the SAE and MLP network. First, an SAE configuration is utilised to learn the deep features that exist in a subset of 2,465 SNPs (P-value < $1x10^{-2}$), to capture information about important SNPs and the cumulative epistatic interactions between them. A layer wise configuration is utilised by

stacking four single layer AEs with 2,000-1,000-500-50 hidden units, where the original 2,465 SNPs are compressed into progressively smaller hidden layers as discussed in Chapter 3. The final SAE hidden layer is then used to initialise the weights of an MLP. The data set is split randomly into training (60%), validation (20%) and testing (20%). The classifier is trained with stochastic gradient descent using backpropagation and fine-tuned to classify case-control instances in the validation and test sets. Again, four classification experiments are presented and evaluated using MSE, Logloss, AUC, Gini, Sensitivity and Specificity values to measure the performance of each model.

4.6.1 Hyper-parameter selection

The structure and selection of the parameters are conducted as in the previous section (Section 4.5) were parameters obtained from the best model were selected via random search. These are presented in Table 4-13 and Table 4-14.

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	Yes
Stopping tolerance	0.01
Stopping rounds	2
Max model generated	200

Table 4-13: Tuning parameters for classification tasks in the third experiment

The specific parameters required for each model are presented in Table 4-14.

Input	Parameter	Value
2,000	Activation	RectifierWithDropout
	Hidden	2
	Neurons	10
	Epochs	50
	L1	$4.7x10^{-5}$
	L2	$2.0x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5
2,000-1,000	Activation	RectifierWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$8.5x10^{-5}$
	L2	$6.0x10^{-6}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5
2,000-1,000-500	Activation	TanhWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$6.8x10^{-5}$
	L2	$5.1x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
2,000-1,000-500-50	Activation	RectifierWithDropout
	Hidden	2
	Neurons	50
	Epochs	100
	L1	$4.2\ x10^{-5}$
	L2	$6.1\ x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5

Table 4-14: Model-specific tuning parameters

Based on empirical analysis, these configurations produced the best results.

4.6.2 Classifier performance

To measure the performance, each MLP classifier was initialized using the different compressed units obtained using the SAE defined in the study. Performance metrics for the validation set are provided in Table 4-15 while Table 4-16 shows the performance metrics on the test data when the trained models are used.

The first layer compresses the input space to 2,000 hidden units to initialize the weights of an MLP before the fully connected layers are fine-tuned for classification tasks. An optimised F1 threshold value of 0.4977 was used to extract the validation set metrics as indicated in Table 4-15. Successive layers of the SAE were used to initialise and fine-tune the remaining MLP models with 1,000, 500 and 50 hidden units respectively. On this occasion, metrics were obtained using optimised F1 threshold values 0.6188, 0.4978 and 0.2701 for each of the remaining models respectively.

Layers	**SE**	**SP**	**Gini**	**LogLoss**	**AUC**	**MSE**
2,000	0.9202	0.9383	0.9608	0.1817	0.9804	0.0547
2,000-1,000	0.8404	0.9383	0.9034	0.2889	0.9517	0.0848
2,000-1,000-500	0.8670	0.8899	0.8828	0.3146	0.9414	0.0963
2,000-1,000-500-50	0.9202	0.5771	0.6976	0.4776	0.8488	0.1593

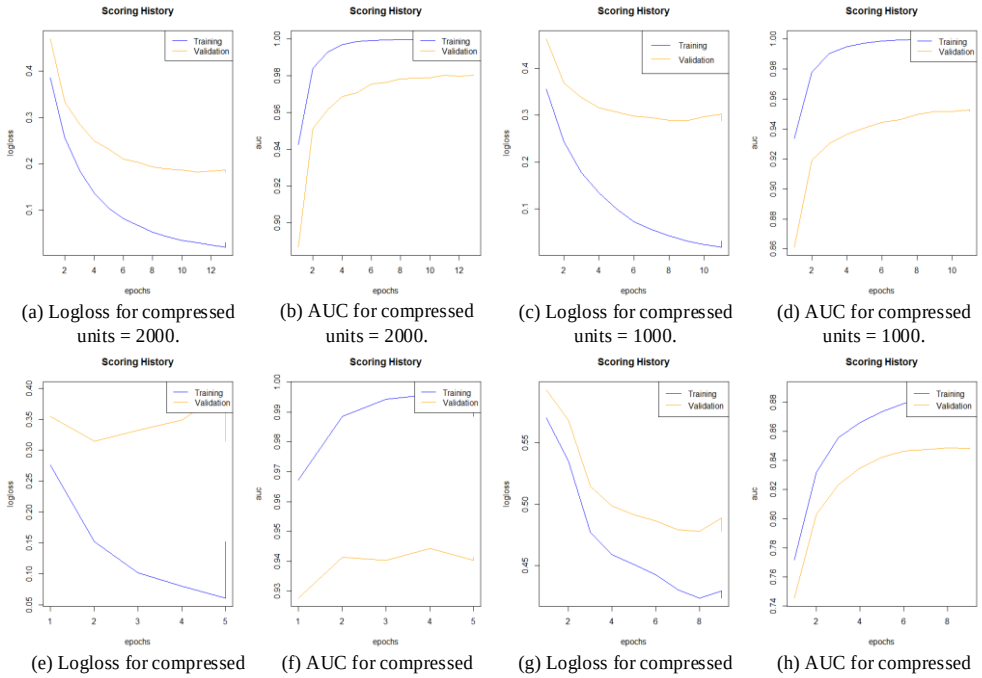
Table 4-15: Performance metrics for validation set

Table 4-16 shows the performance metrics obtained using the test set. Optimised F1 threshold values 0.5363, 0.3356, 0.3899 and 0.4615 were used to obtain these metrics by training the models with 2,000, 1,000, 500 and 50 compressed input units respectively.

Layers	SE	SP	Gini	LogLoss	AUC	MSE
2,000	0.9491	0.9330	0.9499	0.1956	0.9750	0.0540
2,000-1,000	0.9152	0.8756	0.9102	0.2948	0.9551	0.0875
2,000-1,000-500	0.9096	0.8756	0.9005	0.2851	0.9502	0.0872
2,000-1,000-500-50	0.7853	0.7990	0.7036	0.4769	0.8518	0.1563

Table 4-16: Performance metrics for test set

Early stopping was adopted to avoid overfitting. Model building stops when the logloss on the validation set does not improve by at least 1 percent for 2 consecutive scoring epochs (stopping rounds). As shown in Figure 4-7 the AUC plots provide useful information about early divergence between the training and validation curves. In this instance, there is limited evidence to suggest overfitting has occurred except when 500 hidden units are used.



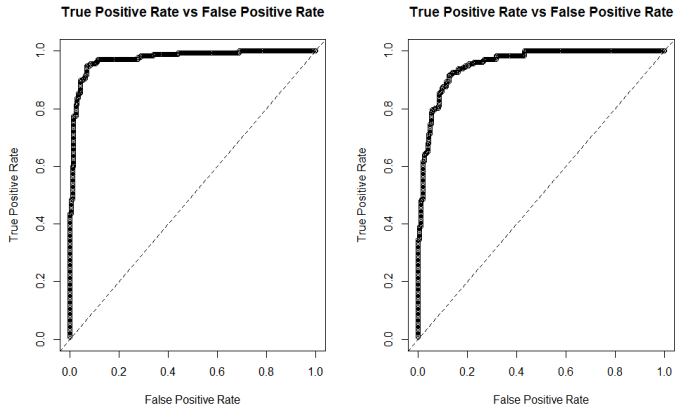
(a) Logloss for compressed units = 2000. (b) AUC for compressed units = 2000. (c) Logloss for compressed units = 1000. (d) AUC for compressed units = 1000.

(e) Logloss for compressed units = 500. (f) AUC for compressed units = 500. (g) Logloss for compressed units = 50. (h) AUC for compressed units = 50.

Figure 4-7: From (a) to (h), Logloss and AUC plots against epochs for 2,000-1,000-500-50 compressed units

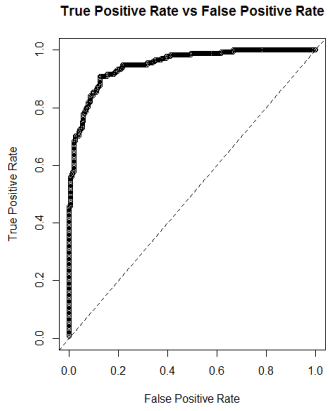
4.6.3 Model selection

The cut-off values for the false and true positive rates in the test set are depicted in Figure 4-8. The ROC curves show a gradual deterioration in classifier performance as the initial 2,465 features (SNPs) are progressively compressed to 50 hidden units in the SAE. Despite the observable deterioration, the results remain high with 50 compressed hidden units. This is in stark contrast to the P-value approach adopted in the previous experiments with GLM and MLP without SAE weight initialisation.

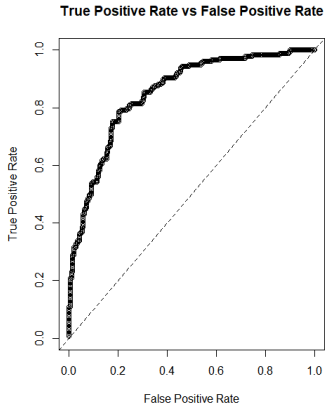


(a) ROC for compressed units = 2000.

(b) ROC for compressed units = 1000.



(c) ROC for compressed units = 500.



(d) ROC for compressed units = 50.

Figure 4-8: From (a) to (d) performance ROC curves for the test set using trained models with the different compressed units considered for the SAE

4.7 SAERMA: Stacked Autoencoder Rule Mining Algorithm for the Interpretation of Epistatic Interactions in GWAS of Extreme Obesity

In this final experiment QC, association analysis, rule mining, SAE and MLP classification are combined to form the SAERMA algorithm.

4.7.1 ARM

In this work, SNPs are referred to as items whilst individuals are referred to as transactions in the MyCode dataset. The R package *arules* is used for the rule generation process (Hahsler et al. 2005; Hahsler et al. 2011; Hahsler et al. 2018). Table 4-17 shows a summary containing the number of rules generated using the *Apriori* algorithm in cases and controls, using support $\sigma = 0.6$ and confidence $\delta = 0.8$. The time in minutes required to generate the rules and the maximum subset size with the maximum number of SNPs per rule is also reported.

Group	Total rules	Time (min)	Max subset size	Algorithm
Cases	208,553,621	33.5	15	Apriori
Controls	218,816,734	34.9	15	Apriori

Table 4-17: ARM summary

It can be observed that the rule generation process time for controls was slightly higher than for cases, possibly due to the number of transactions (individuals) which was also higher than in cases (879 cases and 1,118 controls). Redundant rules were removed; hence, this resulted in a substantial reduction, leaving 18,250,501 rules for cases and 17,949,083 rules for controls.

To select and arrange the most significant rules, those with lift values higher than 1 and χ^2 values higher that 3.84 were filtered and retained. Finally, the rules were ranked based on the highest lift value confirmed by χ^2, which represents the lowest P-values, i.e. the most statistically significant association rules. Results obtained after applying this criterion are shown in Table 4-18 and Table 4-19 for cases and controls respectively. In each case, only the top 10 rules identified by the *Apriori* algorithm and related information about the interestingness measures *support* (Supp.), *confidence* (Conf.), *lift* and χ^2 are reported. Although the number of rules identified was much higher, only the top 10 rules were included for simplicity when presenting the results. This number facilitates an effective visualisation of the rules and concentrates on the most relevant ones. Graph-based visualisation for rule representations tend to become saturated and are only viable for small rule sets (Hahsler 2017).

Rank	**Rule**	**Supp.**	**Conf.**	**Lift**	χ^2
1	{rs12053340_C_D} => {rs1527944_T_D}	0.6075	1	1.6399	870.64
2	{rs1046724_T_D} => {rs7448421_C_D}	0.6109	1	1.6369	879
3	{rs13171869_T_D} => {rs1046724_T_D}	0.6064	0.9944	1.6277	849.82
4	{rs13171869_T_D} => {rs7448421_C_D}	0.6064	0.9944	1.6277	849.82
5	{rs2073950_A_D, rs2301621_A_D} => {rs10849949_C_D}	0.6098	0.9907	1.6248	858.18
6	{rs2832503_G_D} => {rs977779_C_D}	0.6166	1	1.6218	879
7	{rs12315146_A_D, rs2301621_A_D} => {rs2073950_A_D}	0.6018	1	1.6218	826.05
8	{rs2301621_A_D} => {rs10849949_C_D}	0.6098	0.9889	1.6218	854.06
9	{rs2073950_A_D} => {rs10849949_C_D}	0.6098	0.9889	1.6218	854.06
10	{rs11682173_T_D} => {rs11692215_T_D}	0.6086	0.9853	1.6188	845.91

Table 4-18: Top 10 rules identified in cases

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs10501544_C_D} => {rs12280583_T_D}	0.6082	1	1.6369	1105.46
2	{rs10828296_G_D, rs11593316_T_D} => {rs6482203_A_D}	0.6055	1	1.6345	1088.99
3	{rs10828296_G_D, rs1926690_G_D} => {rs6482203_A_D}	0.6046	1	1.6345	1084.92
4	{rs7171993_G_D} => {rs3743121_A_D}	0.6127	1	1.6297	1113.79
5	{rs10828296_G_D} => {rs6482203_A_D}	0.6091	0.9971	1.6297	1097.06
6	{rs11593316_T_D, rs6482203_A_D} => {rs1926690_G_D}	0.6064	0.9985	1.6273	1080.66
7	{rs10828296_G_D, rs11593316_T_D} => {rs1926690_G_D}	0.6046	0.9985	1.6273	1072.57
8	{rs6482203_A_D} => {rs1926690_G_D}	0.6064	0.9912	1.6154	1059.80
9	{rs10828296_G_D} => {rs1926690_G_D}	0.6046	0.9897	1.6130	1047.57
10	{rs2042867_T_D, rs4979935_T_D} => {rs735638_G_D}	0.6002	1	1.6040	1013.69

Table 4-19: Top 10 rules identified in controls

Among the top 10 rules identified in cases (see Table 4-18), no rules were common with those identified in controls (see Table 4-19). The most significant rule identified in cases, involved an interaction between two SNPs: {rs12053340_C_D} => {rs1527944_T_D}. This rule indicates that 61% of the cases had the variants rs12053340_C_D and rs1527944_T_D together, and those who had rs12053340_C_D also had rs1527944_T_D 100% of the time (confidence = 1). Furthermore, the lift value for this rule was higher than 1 (1.6399) and the highest among all rules; and χ^2 (870.64) was significantly higher than 3.84, indicating strong positive correlation. Although the top 10 rules were unique, some of them have SNPs in common as can be seen in Table 4-18. Rules 2-4 as well as rules 5, 7, 8 and 9 had some SNPs in common. Conversely, rules 1, 6 and 10 did not share any SNPs between the top 10 rules in cases. Results also revealed that all the SNPs within the top 10 rules in cases, were labelled as D, indicating that the minor allele counting for the SNPs was 0 (the SNPs are homozygous major alleles).

In controls, it can be observed that the most significant rule is {rs10501544_C_D} => {rs12280583_T_D} as shown in Table 4-19. This rule

was also common in 61% of the control samples and its consequent (rs12280583_T_D) was always present when the antecedent (rs10501544_C_D) was present (confidence = 1). The lift value for this rule was higher than 1 (1.636896); and χ^2 (1105.465) much higher than 3.84, indicating strong positive correlation. As in cases, rules 2, 3, 5, 6, 7, 8 and 9 shared common SNPs. Rules 1, 4 and 10 did not share any SNPs. SNPs within the top 10 rules in controls were also labelled as D, indicating that SNPs are homozygous major alleles.

Finally, the top 10 rules were represented as a network using graph-based visualization techniques. Both in cases and in controls, individual rules and clusters can be observed. Clusters represent association rules that contain one or more SNPs in common. Even though lift and χ^2 measure dependency and independency in rules, high χ^2 values do not necessarily imply the highest lift, as observed in Table 4-18 and Table 4-19. Graph-based visualizations for the top 10 rules obtained for cases and controls are shown in Figure 4-9 and Figure 4-10 respectively. In Figure 4-9, rule 1, 6 and 10 are the only individual rules depicted. Rules 2, 3 and 4 form a cluster with common SNPs rs1046724_T_D, rs7448421_C_D and rs13171869_T_D. Additionally, a second cluster formed by rules 5, 7, 8 and 9 can be observed. This represents a combination of the only two rules with three SNPs in Table 4-18 (rules 5 and 7) and two rules with two SNPs (rules 8 and 9), where the shared SNPs are rs2073950_A_D, rs2301621_A_D and rs10849949_C_D. Conversely, in Figure 4-10, the main cluster is composed by rules 2, 3, 5, 6, 7, 8, 9 and 10. In this case, the common SNPs in the cluster are rs10828296_G_D, rs11593316_T_D, rs6482203_A_D and rs1926690_G_D. This time, the cluster is a combination of rules composed

by three items (2, 3, 6 and 7) and two items (5, 8 and 9). Rules 1, 4 and 10 are individual rules with no cluster formation. Extended plots with the top 100 rules for cases and controls can be found in Appendix B.

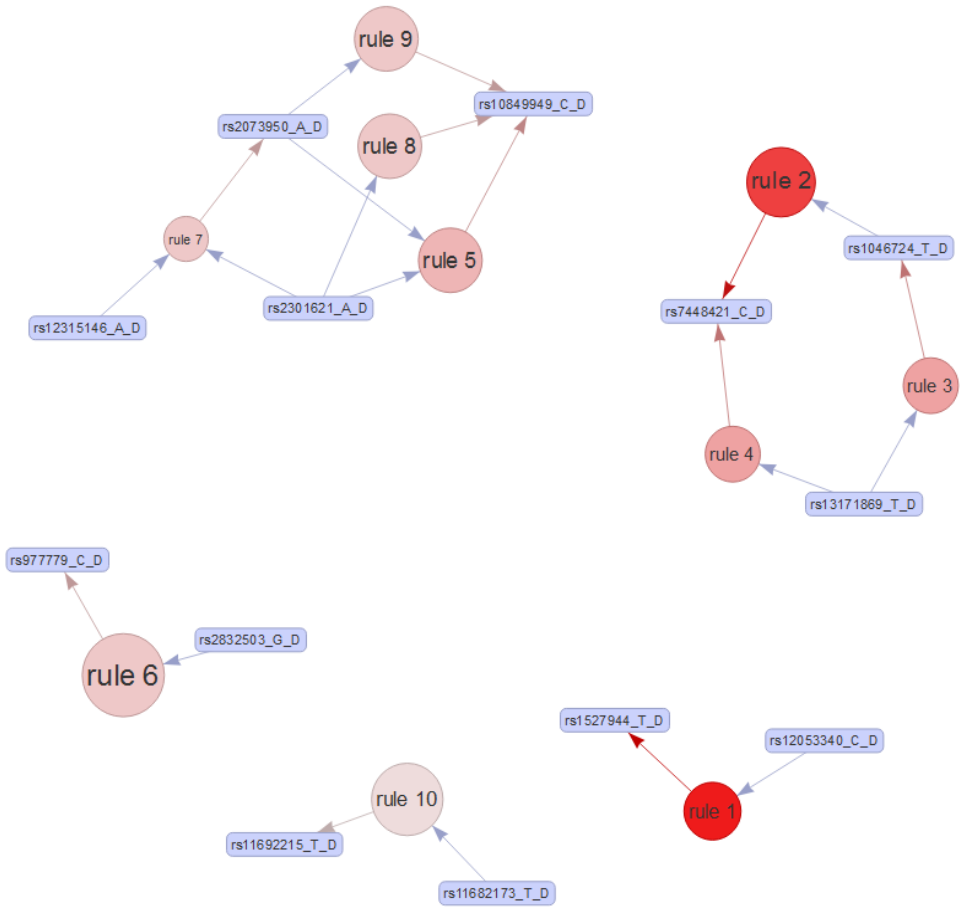


Figure 4-9: Rule visualisation network for the top 10 rules identified in cases

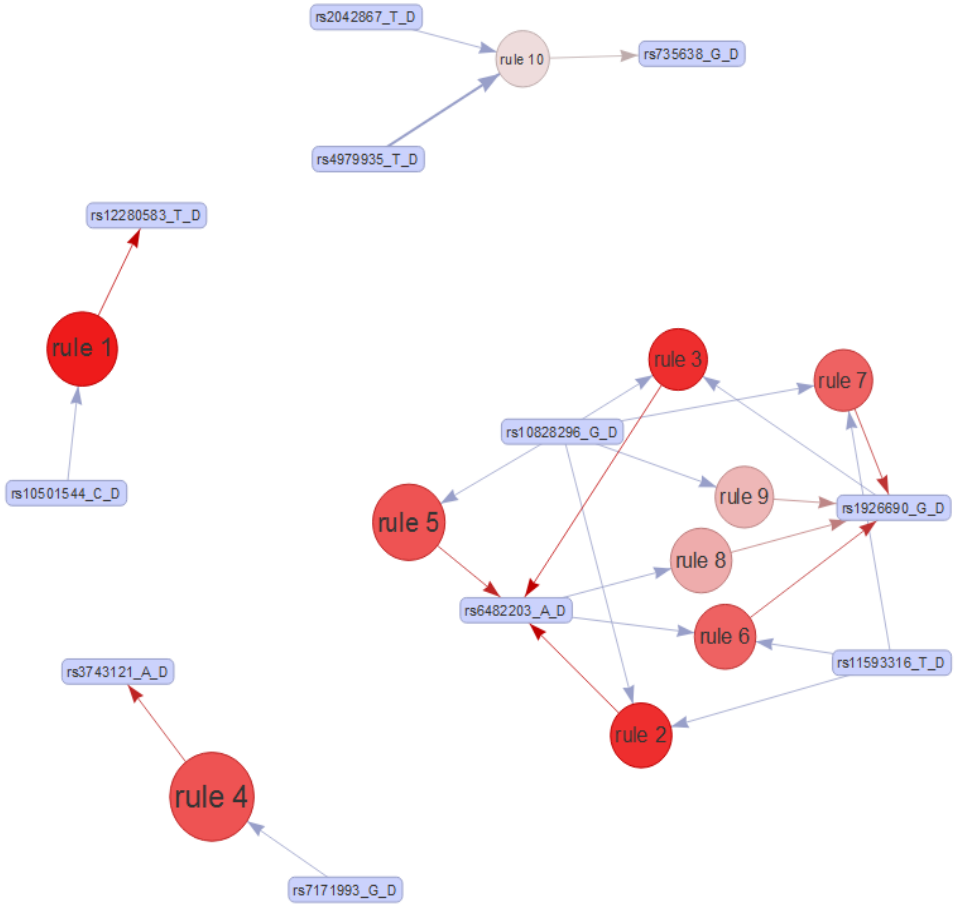


Figure 4-10: Rule visualisation network for the top 10 rules identified in controls

Association rule mining tasks were conducted using Christian Borgelt's *Apriori* implementation via the *arules* and *arulesViz* R packages. A Windows Server 2016 Standard 64-bit version-based machine, with 768 GiB of Memory and an Intel® Xeon® CPU E5-2620 v4 @ 2.10 GHz (16 CPUs), was utilized to conduct the analysis.

4.7.2 SAERMA model performance

Classification analysis was conducted in this experiment using a second feature selection step based on ARM. Rule mining allows us to find the most frequent SNPs (from the 2,465 SNPs considered) among individuals in cases and controls and then extract rules from them. These rules can be plotted as

discussed earlier, to provide insights through rule inspection. Additionally, items from the rules (SNPs) can be utilised as input features in our SAE for deep feature extraction (epistatic relationships between SNPs) and used to initialise the weights of an MLP before fine-tuning and classification purposes. By adjusting support and confidence parameters in the rule generation process, the number of rules can be increased or decreased. This, in turn, impacts the performance of the models generated for feature extraction and classification tasks. Hence, results presented next are derived from the SNPs contained within the most significant rules extracted with support $\sigma = 0.6$ and confidence $\delta = 0.8$ as discussed earlier. These are the lowest interest measure values which allow rule generation without overloading the system used in this study.

Several classification tasks were conducted using the top 300, 200, 100 and 50 rules from the previous ARM analysis as can be observed in Table 4-20 to Table 4-23. To accomplish this, the SNPs from each set of rules were compressed using SAEs as conducted in Section 4.6. However, this time by utilising three AEs instead of four (since the number of input features was considerably lower) with a variable number of hidden units. The number of AEs and hidden neurons were arbitrarily selected to gradually reduce the number of initial features. The final layers of the SAEs were then utilised to initialise the weights of the MLPs before being fine-tuned for classification tasks.

Table 4-20 contains the classifier performance values for both the validation and test set using the SAE and SNPs from the top 300 rules. In this instance 204 SNPs are compressed using the following layer configuration: 150-100-50.

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
150	Validation	0.7340	0.7621	0.5839	0.5508	0.7920	0.1841
	Test	0.7966	0.6268	0.5594	0.5770	0.7797	0.1952
150-100	Validation	0.7979	0.6476	0.5389	0.5783	0.7695	0.1964
	Test	0.7684	0.6794	0.5349	0.5769	0.7675	0.1968
150-100-50	Validation	0.8617	0.3921	0.3588	0.6411	0.6794	0.2252
	Test	0.7797	0.5789	0.4591	0.6125	0.7295	0.2117

Table 4-20: Classifier results for top 300 rules

Table 4-21 contains the classifier performance values for the validation and test sets using an SAE and SNPs from the top 200 rules. This resulted in 161 SNPs that were then compressed using the following layer configuration: 125-75-50.

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
125	Validation	0.7766	0.6255	0.4456	0.6046	0.7228	0.2076
	Test	0.7401	0.6651	0.4715	0.6099	0.7357	0.2104
125-75	Validation	0.8777	0.5198	0.4509	0.6083	0.7254	0.2099
	Test	0.7006	0.6746	0.3788	0.6553	0.6894	0.2259
125-75-50	Validation	0.7606	0.5859	0.3394	0.6462	0.6697	0.2275
	Test	0.7853	0.4976	0.3959	0.6280	0.6980	0.2191

Table 4-21: Classifier results for top 200 rules

Table 4-22 contains the classifier performance values for the validation and test sets using an SAE and SNPs from the top 100 rules. This resulted in 124 SNPs that were compressed using the following layer configuration: 90-50-25.

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
90	Validation	0.7872	0.5022	0.3738	0.6351	0.6869	0.2220
	Test	0.6949	0.6603	0.4170	0.6231	0.7085	0.2167
90-50	Validation	0.9361	0.2247	0.3521	0.6396	0.6760	0.2245
	Test	0.7853	0.5024	0.4065	0.6285	0.7032	0.2185
90-50-25	Validation	0.9202	0.2687	0.3137	0.6499	0.6568	0.2293
	Test	0.8136	0.4832	0.3674	0.6425	0.6837	0.2257

Table 4-22: Classifier results for top 100 rules

Finally, Table 4-23 contains classifier performance values for the validation and test sets using an SAE and SNPs from the top 50 rules. This resulted in 92 SNPs that were compressed using the following layer configurations: 75-50-25.

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
75	Validation	0.7606	0.4714	0.2898	0.6615	0.6449	0.2340
	Test	0.8305	0.4545	0.3949	0.6338	0.6974	0.2210
75-50	Validation	0.9149	0.2731	0.3171	0.6471	0.6585	0.2282
	Test	0.7740	0.6268	0.4529	0.6178	0.7265	0.2142
75-50-25	Validation	0.9681	0.1542	0.3482	0.6464	0.6741	0.2277
	Test	0.8362	0.4402	0.3735	0.6372	0.6867	0.2242

Table 4-23: Classifier results for top 50 rules

Logloss, AUC and ROC plots for Table 4-20 to Table 4-23 as well as tuning parameters used by each model are presented and summarised in Appendix C. As observed in previous experiments, the results utilising the four different rule configurations (204, 161, 124 and 92 SNPs) were all affected when the input features of the classifier using AEs was reduced.

4.8 Chapter summary

This chapter presents the results obtained following the detection of epistatic interactions in association studies of binary traits utilising several traditional

and more advanced statistical and computational methods. The results were reported in four different experiments following a stepwise approach. Based on the limitations derived from each experiment, the different components proposed in the methodology chapter (Chapter 3) were incorporated to overcome these limitations.

Chapter 5. POST ANALYTICS INTERROGATION

Direct biological inference from the results using statistical tests is a complex task since statistical interactions do not automatically entail interaction at the biological or mechanistic level (Cordell 2002). Hits identified by GWAS represent genomic regions rather than specific genes. Therefore, it is important to identify the gene underlying associations after conducting genome-wide association experiments. This task is conducted using the SNPnexus tool.

While no SNPs reached genome-wide significance levels of association in association analysis, five SNPs were suggestive of association as shown in Table 4-4. Among these suggestive SNPs, rs763727, rs10817737 and rs3050 were identified in the genes CDH13, TMOD1 and PLEKHG1 respectively, while rs1278895 and rs726553 were located in intergenic regions, 818 b to the nearest upstream gene RP11-159D23.2 and 109,332 b to the nearest upstream gene DOCK10 respectively.

In the rules

In the rule generation process, redundant rules were removed, and a number of assumptions were applied to reduce and rank the most significant rules as shown in Table 4-18 and Table 4-19. The most significant rule identified in cases was {rs12053340} => {rs1527944}, composed of SNPs located within the genes SGOL2 and AOX1 respectively, both protein-coding genes in chromosome 2 with risk allele C and T respectively. The remaining rules were composed by the following genes: rule 2, {rs1046724} => {ZNF354B}, where rs1046724 is located 563 b upstream of the protein-coding gene ZNF354B in chromosome 5.

Rule 3, {ZFP2} => {rs1046724}, with rs1046724 variant located 563 b upstream of the protein-coding gene ZNF354B in chromosome 5. Rule 4, {ZFP2} => {ZNF354B}, both protein-coding genes located in chromosome 5. Rule 5, {ATXN2, ATXN2} => {ATXN2}, with three variants interacting within the gene ATXN2, located in chromosome 12. Rule 6, {rs2832503} => {rs977779}. This rule is composed of two intergenic variants located in chromosome 21, where the closest upstream gene is GRIK1 with distances 11,830 b and 13,187 b respectively. Rule 7, {MAPKAPK5, ATXN2} => {ATXN2}, involved two different protein-coding genes located in chromosome 12. Rule 8, {ATXN2} => {ATXN2}, is composed by SNPs located in the protein-coding ATXN2 (chromosome 12). Rule 9, {ATXN2} => {ATXN2}, also composed by interactions between SNPs located in the gene ATXN2. Finally, rule 10, {rs11682173} => {AFF3}, where the intergenic variant rs11682173 is located 609 b upstream of the protein-coding gene AFF3, in chromosome 2. In Table 5-1 the top 10 rules identified in cases have been presented with the mapped genes as items.

Rank	Rule: SNPs as items	Rule: Genes as items
1	{rs12053340_C_D} => {rs1527944_T_D}	{SGOL2} => {AOX1}
2	{rs1046724_T_D} => {rs7448421_C_D}	{ZNF354B} => {ZNF354B}
3	{rs13171869_T_D} => {rs1046724_T_D}	{ZFP2} => {ZNF354B}
4	{rs13171869_T_D} => {rs7448421_C_D}	{ZFP2} => {ZNF354B}
5	{rs2073950_A_D, rs2301621_A_D} => {rs10849949_C_D}	{ATXN2, ATXN2} => {ATXN2}
6	{rs2832503_G_D} => {rs977779_C_D}	{GRIK1} => {GRIK1}
7	{rs12315146_A_D, rs2301621_A_D} => {rs2073950_A_D}	{MAPKAPK5, ATXN2} => {ATXN2}
8	{rs2301621_A_D} => {rs10849949_C_D}	{ATXN2} => {ATXN2}
9	{rs2073950_A_D} => {rs10849949_C_D}	{ATXN2} => {ATXN2}
10	{rs11682173_T_D} => {rs11692215_T_D}	{AFF3} => {AFF3}

Table 5-1: Equivalent rules with closest genes as items in the case set

It can be observed that rules 1, 6 and 10 were the only rules in cases that did not share SNPs with other rules, while rules 2, 3, 4, 5, 7, 8 and 9 were composed by SNPs shared between these rules (see Table 4-18). Furthermore, rules 2, 3 and 4 showed interactions between the genes ZNF354B and ZFP2, while rules 5, 7, 8 and 9 indicated interactions between variants in the gene ATXN2, although rule 7 also interacted with the MAPKAPK5 gene.

The most significant rule in controls was {rs10501544_C_D} => {rs12280583_T_D}. The SNPs in this rule are located within the protein-coding gene DLG2 in chromosome 11, with risk alleles C and T respectively. Rule 2, {rs10828296_G_D, rs11593316_T_D} => {rs6482203_A_D}, is composed of three intergenic variants situated 6,107 b upstream of the closest pseudogene ADIPOR1P1, 20,245 b downstream of the closest protein-coding gene EBLN1 and, 17,807 b downstream of the gene EBLN1 respectively, all situated in chromosome 10. Rule 3, {rs10828296_G_D, rs1926690_G_D} => {rs6482203_A_D}, combines three intergenic variants close to the pseudogene ADIPOR1P1 and the protein-coding gene EBLN1 ({ADIPOR1P1, ADIPOR1P1} => {EBLN1}) with distances 6,107 b upstream, 17,018 b upstream and 17,807 b downstream respectively. Rule 4, {rs7171993_G_D} => {rs3743121_A_D}, combines two intergenic variants situated 2,469 b upstream of the pseudogene RP11-83J16.3 and 387 b downstream of the protein-coding gene AQR respectively, both in chromosome 15. Rule 5, {rs10828296_G_D} => {rs6482203_A_D}, is composed by two intergenic variants situated 6,107 b upstream the closest pseudogene ADIPOR1P1 and 17,807 b downstream the closest protein-coding gene EBLN1 respectively, both in chromosome 10. Rule 6, {rs11593316_T_D, rs6482203_A_D} => {rs1926690_G_D}, combines

three intergenic variants located 20,245 b downstream the closest protein-coding gene EBLN1, 17,807 b downstream EBLN1, and 17,018 upstream the pseudogene ADIPOR1P1 respectively, in chromosome 10. Rule 7, {rs10828296_G_D, rs11593316_T_D} => {rs1926690_G_D}, involved also three intergenic variants in chromosome 10, located 6,107 b upstream of the ADIPOR1P1 pseudogene, 20,245 b downstream of the protein-coding gene EBLN1, and 17,018 upstream of the pseudogene ADIPOR1P1 respectively. Rule 8, {rs6482203_A_D} => {rs1926690_G_D}, combines two intergenic variants located 17,807 b downstream of the gene EBLN1 and 17,018 upstream of the pseudogene ADIPOR1P1 respectively, in chromosome 10. Rule 9, {rs10828296_G_D} => {rs1926690_G_D}, combined two intergenic variants close to the pseudogene ADIPOR1P1 and the protein-coding gene EBLN1 with distances 6,107 b upstream and 17,018 b upstream respectively. Rule number 10, {rs2042867_T_D, POLR3A} => {rs735638_G_D}, is formed by two intergenic variants and a protein-coding gene POLR3A. The variant rs2042867 is located 8,608 b upstream of the H2AFZP5 pseudogene while the closest gene of rs735638 is located 2,186 b downstream of the protein-coding gene POLR3A. In Table 5-2 the top 10 rules identified in control samples have been presented with the mapped genes as items.

Rank	Rule: SNPs as items	Rule: Genes as items
1	{rs10501544_C_D} => {rs12280583_T_D}	{DLG2} => {DLG2}
2	{rs10828296_G_D, rs11593316_T_D} => {rs6482203_A_D}	{ADIPOR1P1, EBLN1} => {EBLN1}
3	{rs10828296_G_D, rs1926690_G_D} => {rs6482203_A_D}	{ADIPOR1P1, ADIPOR1P1} => {EBLN1}
4	{rs7171993_G_D} => {rs3743121_A_D}	{RP11-83J16.3} => {AQR}
5	{rs10828296_G_D} => {rs6482203_A_D}	{ADIPOR1P1} => {EBLN1}
6	{rs11593316_T_D, rs6482203_A_D} => {rs1926690_G_D}	{EBLN1, EBLN1} => {ADIPOR1P1}
7	{rs10828296_G_D, rs11593316_T_D} => {rs1926690_G_D}	{ADIPOR1P1, EBLN1} => {ADIPOR1P1}
8	{rs6482203_A_D} => {rs1926690_G_D}	{EBLN1} => {ADIPOR1P1}
9	{rs10828296_G_D} => {rs1926690_G_D}	{ADIPOR1P1} => {EBLN1}
10	{rs2042867_T_D, rs4979935_T_D} => {rs735638_G_D}	{H2AFZP5, POLR3A} => {POLR3A}

Table 5-2: Equivalent rules with closest genes as items in the control set

It can be noted that rules 1, 3 and 10 in controls are the only rules which do not share any common SNPs as can be seen in Table 4-19. A full discussion and interpretation of the findings in this chapter will be presented in Chapter 6.

Although mapping variants to genes can be useful to extend the information about the rules, generating a list of genes does not provide any evidences about epistasis from a biological point of view. Retrieving a functional profile of the gene set to better understand the underlying biological mechanism in obesity represents a more insightful approach. Therefore, in this chapter, biological interpretation of the rules identified is provided. This will test whether the genes identified within the most significant rules (top 300) in cases and controls belong to any relevant biological pathway or not. If positive matches are identified, potential true biological epistasis can be discovered by ARM without preliminary knowledge, relying solely on statistical approaches. Finally, as a proof of concept approach, instead of relying on statistical filtering to preselect SNPs based on a P-value threshold $< 10^{-2}$, SNPs will be selected based on gene set enrichment analysis (GSEA) (Subramanian et al. 2005) from GWAS results (all SNPs and P-values after Logistic Regression), to identify the correlation

between pathways and the phenotype under investigation. Subsequently, ARM analysis will be repeated with the identified variants to explore epistasis. Additionally, classification analysis will be performed using MLP to measure the prediction capacity of genetic variants identified in the rules within the pathways. These experiments are intended to serve as a proof of concept in order to validate the results derived from this thesis from a biological perspective.

5.1 Biological Interpretation of the Results

A common approach to provide biological interpretation of genetic variants is via gene set enrichment analysis which is based on the functional annotation of gene sets. This approach represents an advantage especially if we want to identify whether a set of genes are associated with specific biological process /molecular function or not. Popular tools for gene set enrichment and pathway analysis such as DAVID, GSEA or Reactome are available and can be used for this purpose.

Utilising functional enrichment strategies have the potential to increase the probabilities for researchers to identify biological processes more relevant to the disease under investigation (Huang et al. 2009).

5.1.1 Biological Implication of Association Rules from SAERMA

To validate the rules identified by SAERMA, we tested whether the genes forming the rules were involved in biological pathways or not. Any identified rules including more than one gene involved in a particular pathway can be considered potential true obesity epistasis. The DAVID Functional Annotation Tool (Dennis et al. 2003) was used to perform this task. DAVID stand for

database for annotation, visualisation and integrated discovery (DAVID) and is a widely used tool for functional interpretation and biological meaning of genes and proteins, cited in more than 6,000 scientific publications (Jiao et al. 2012).

5.1.1.1 Results from DAVID

A total of 132 genes were extracted from the top 300 rules in cases and controls and used as input in DAVID. Gene-specific functional data was obtained using the KEGG (Kanehisa & Goto 2000) and the Reactome pathway knowledgebase (Fabregat, Jupe, et al. 2018; Joshi-Tope 2004) options from the annotation categories in the *Annotation Summary Results* page, although other categories are available for selection (Dennis et al. 2003). Instead of selecting highly enriched pathways in the annotation categories, all pathways were considered as the main goal in this experiment is to find items in the rules matching genes in pathways relevant to obesity. A list with all the most relevant pathways identified from KEGG and Reactome for the Homo sapiens are reported in Table 5-3. The table includes the pathway name and information about reported source (KEGG/Reactome), the number of genes mapped to the pathway (count), the gene official names, association rules mapping the genes in the pathway and, finally, a list of the genes in the rules that were not found in the pathway. For each rule identified, CA indicates that the rule was identified in cases whereas CO indicates control. Additionally, the gene/s in the rules mapping genes in the pathways were highlighted.

Pathway Name	Source	Count	Genes Found	Rule	Genes in Rule not in Pathway
FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes	Reactome	2	HDAC2, NPY	CA: {**HDCA2**, ACAD10}=>{BRAP} CO: {ATXN2, **NPY**}=>{ATXN2}	ACAD10, BRAP
Metabolism of lipids	Reactome	3	SLC10A2, ACAD10, MTMR7	CA: {**ACAD10**, RP3-462E2.3}=>{**ACAD10**} CO: {MAPKAP5, ATXN2, **ACAD10**, **SLC10A2**}=>{ATXN2} CO: {**MTMR7**}=>{**MTMR7**}	RP3-462E2.3, MAPKAP5, ATXN2
GPCR ligand binding	Reactome	2	NLN, NPY	CO: {SYT14, **NLN**}=>{SYT14} CO: {ATXN2, **NPY**}=>{ATXN2}	SYT14, ATXN2
Axon guidance.	Reactome	6	ROBO2, COL6A3, NEO1, GFRA2, ITGA9, RGMA	CO: {SYT14, **ROBO2**}=>{SYT14} CO: {**COL6A3**,GRIK1}=>{GRIK1} CO: {ATXN2, **NEO1**}=>{ATXN2} CO: {**GFRA2**, CSNK1E}=>{CSNK1E} CO: {**ITGA9**}=>{**ITGA9**} CO: {AC097713.3, **RGMA**}=>{AC097713.3}	SYT14, GRIK1, ATXN2, CSNK1E, AC097713.3
MAPK family signaling cascades	Reactome	4	BRAP, DLG2, MAPKAPK5, GFRA2	CA: {**MAPKAPK5**, **BRAP**}=>{ACAD10} CA: {**BRAP,** RP3-462E2.3}=>{ACAD10} CA: {**BRAP**}=>{**BRAP**} CA: {**MAPKAPK5**, ATXN2}=>{ATXN2} CO: {**DLG2**}=>{**DLG2**} CO: {**BRAP**, ATXN2}=>{ATXN2} CO: {**MAPKAPK5**, ATXN2, snoU13}=>{ATXN2} CO: {**GFRA2**, CSNK1E}=>{CSNK1E}	ACAD10, ATXN2, snoU13, CSNK1E
M Phase	Reactome	3	SGOL2, CSNK1E, PHF8	CA: {**SGOL2**}=>{AOX1} CO: {AOX1, DOCK4}=>{**SGOL2**} CO: {DDX18, **CSNK1E**}=>{**CSNK1E**} CO: {**PHF8**, LINC00460}=>{RNA5SP38}	AOX1, DOCK4, DDX18, LINC00460, RNA5SP38
Factors involved in megakaryocyte development and platelet production	Reactome	4	DOCK6, DOCK10, DOCK4, HDAC2	CA: {**DOCK6**}=>{**DOCK6**} CA: {**HDAC2**, ACAD10}=>{BRAP} CO: {**DOCK10**}=>{ **DOCK10**} CO: {AOX1, **DOCK4**}=>{SGOL2}	ACAD10, BRAP, AOX1, SGOL2
RNA Polymerase III Transcription Termination	Reactome	2	POLR3A, NFIA	CO: {H2AFZP5, **POLR3A**}=>{**POLR3A** } CO: {**NFIA**, AC097713.3}=>{AC097713.3}	H2AFZP5, AC097713.3

NCAM1 interactions	Reactome	2	COL6A3, GFRA2	CO: {**COL6A3**,GRIK1}=>{GRIK1} CO: {**GFRA2**, CSNK1E}=>{CSNK1E}	GRIK1, CSNK1E
Netrin-1 signaling	Reactome	2	NEO1, RGMA	CO: {ATXN2, **NEO1**}=>{ATXN2} CO: {AC097713.3, **RGMA**}=>{AC097713.3}	ATXN2, AC097713.3
ECM proteoglycans	Reactome	2	COL6A3, ITGA9	CO: {**COL6A3**,GRIK1}=>{GRIK1} CO: {**ITGA9**}=>{**ITGA9**}	GRIK1
Positive epigenetic regulation of rRNA expression	Reactome	2	HDAC2, TTF1	CA: {**HDAC2**, ACAD10}=>{BRAP} CA: {TTC27, **TTF1**}=>{TTC27} CO: {AC097713.3, **TTF1**}=>{AC097713.3}	ACAD10, BRAP, TTC27, AC097713.3
Peptide ligand-binding receptors	Reactome	2	NLN, NPY	CO: {SYT14, **NLN**}=>{SYT14} CO: {ATXN2, **NPY**}=>{ATXN2}	SYT14, ATXN2
Transmission across Chemical Synapses	Reactome	3	DLG2, ALDH2, GRIK1	CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2**, RP3-462E2.3}=>{**ALDH2**} CA: {NAA25, **ALDH2**}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{ACAD10} CA: {**GRIK1**, **GRIK1**}=>{ **GRIK1**} CO: {**DLG2**}=>{**DLG2**} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2} CO: { **GRIK1**, **GRIK1**}=>{ **GRIK1**}	ACAD10, RP3-462E2.3, NAA25, ATXN2
Neuronal System	Reactome	4	PTPRD, DLG2, ALDH2, GRIK1	CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2**, RP3-462E2.3}=>{**ALDH2**} CA: {NAA25, **ALDH2**}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{ACAD10} CA: {**GRIK1**, **GRIK1**}=>{ **GRIK1**} CO: {**PTPRD**, SYT14}=>{SYT14} CO: {ZNF366, **PTPRD**}=>{ZNF366} CO: {**PTPRD**, ALPK3}=>{ALPK3} CO: {**DLG2**}=>{**DLG2**} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2} CO: { **GRIK1**, **GRIK1**}=>{ **GRIK1**}	ACAD10, RP3-462E2.3, NAA25, SYT14, ZNF366, ALPK3, ATXN2
Diseases of signal transduction	Reactome	3	BRAP, HDAC2, KREMEN1	CA: {**BRAP**}=>{**BRAP**} CA: {**BRAP**, RP3-462E2.3}=>{ACAD10} CA: {**HDAC2**, ACAD10}=>{**BRAP**} CO: {**BRAP**, ATXN2}=>{ATXN2} CO: {AC097713.3, **KREMEN1**}=>{AC097713.3}	RP3-462E2.3, ACAD10, ATXN2, AC097713.3

Metabolism of RNA	Reactome	4	AQR, RPPH1, CDKAL1, CSNK1E	CA: {MAPKAPK5, **RPPH1**, HECTD4}=>{RP3-462E2.3} CO: {**AQR**}=>{**AQR**} CO: {**CDKAL1**}=>{**CDKAL1**} CO: {DDX18, **CSNK1E**}=>{**CSNK1E**}	MAPKAPK5, HECTD4, RP3-462E2.3, DDX18
Metabolism	Reactome	7	ALDH2, ENTPD4, SLC10A2, AOX1, ACAD10, MTMR7, HK1	CA: {**ACAD10**}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ACAD10**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CA: {MAPKAPK5**, ACAD10**, HECTD4}=>{**ALDH2**} CA: {NAA25**, ACAD10**}=>{**ACAD10**} CA: {SGOL2}=>{**AOX1**} CO: {ATXN2, **ALDH2**, **ACAD10**}=>{ATXN2} CO: {**ENTPD4**, ATXN2}=>{ATXN2} CO: {MAPKAPK5, ATXN2, **ACAD10**, **SLC10A2**}=>{ATXN2} CO: {**AOX1**, DOCK4}=>{SGOL2} CO: {ATXN2**, ACAD10**, HECTD4}=>{ATXN2} CO: {**MTMR7**}=>{**MTMR7**} CO: {**HK1,** GRIK1}=>{GRIK1}	RP3-462E2.3, MAPKAPK5, HECTD4, NAA25, SGOL2, ATXN2, DOCK4, GRIK1
ECM-receptor interaction	KEGG	3	ITGA9, LAMB4, COL6A3	CO: {**ITGA9**}=>{**ITGA9**} CO: {**LAMB4**, ALPK3}=>{ALPK3} CO: {**COL6A3**, GRIK1}=>{GRIK1}	ALPK3, GRIK1
Cell adhesion molecules (CAMs)	KEGG	3	ITGA9, NEO1, ICOSLG	CO: {**ITGA9**}=>{**ITGA9**} CO: {ATXN2, **NEO1**}=>{ATXN2} CO: {ALPK3, **ICOSLG** }=>{ALPK3}	ALPK3, ATXN2
Tryptophan metabolism	KEGG	2	AOX1, ALDH2	CA: {SGOL2}=>{**AOX1**} CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CO: {**AOX1**, DOCK4}=>{SGOL2} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2}	SGOL2, ACAD10, RP3-462E2.3, DOCK4, ATXN2
Valine, leucine and isoleucine degradation	KEGG	2	AOX1, ALDH2	CA: {SGOL2}=>{**AOX1**} CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CO: {**AOX1**, DOCK4}=>{SGOL2} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2}	SGOL2, ACAD10, RP3-462E2.3, DOCK4, ATXN2
Purine metabolism	KEGG	3	POLR3A, PDE8A, ENTPD4	CO: {H2AFZP5, **POLR3A**}=>{**POLR3A**} CO: { **PDE8A**, **ENTPD4**}=>{SLC28A1}	H2AFZP5, SLC28A1

Alcoholism	KEGG	3	HDAC2, NPY, CREB5	CA: {**HDCA2**, ACAD10}=>{BRAP} CO: {ATXN2, **NPY**}=>{ATXN2} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ACAD10, BRAP, ATXN2, AC097713.3
Lysine degradation	KEGG	2	PLOD1, ALDH2	CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CO: {SYT14, **PLOD1**}=>{SYT14} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2}	ACAD10, RP3-462E2.3, SYT14, ATXN2
PI3K-Akt signaling pathway	KEGG	4	ITGA9, LAMB4, COL6A3, CREB5	CO: {**ITGA9**}=>{**ITGA9**} CO: {**LAMB4**, ALPK3}=>{ALPK3} CO: {**COL6A3**, GRIK1}=>{GRIK1} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ALPK3, GRIK1, AC097713.3
cAMP signaling pathway	KEGG	3	ATP2B4, NPY, CREB5	CO: {ZNF366, **ATP2B4**}=>{ZNF366} CO: {ATXN2, **NPY**}=>{ATXN2} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ZNF366, ATXN2, AC097713.3
Focal adhesion	KEGG	3	ITGA9, LAMB4, COL6A3	CO: {**ITGA9**}=>{**ITGA9**} CO: {**LAMB4**, ALPK3}=>{ALPK3} CO: {**COL6A3**, GRIK1}=>{GRIK1}	ALPK3, GRIK1
Glycolysis / Gluconeogenesis	KEGG	2	ALDH2, HK1	CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2} CO: {**HK1**, GRIK1}=>{GRIK1}	ACAD10, RP3-462E2.3, ATXN2, GRIK1
Thyroid hormone synthesis	KEGG	2	TTF1, CREB5	CA: {TTC27, **TTF1**}=>{TTC27} CO: {AC097713.3, **TTF1**}=>{AC097713.3} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	TTC27, AC097713.3
Pyrimidine metabolism	KEGG	2	POLR3A, ENTPD4	CO: {H2AFZP5, **POLR3A**}=>{**POLR3A**} CO: {**ENTPD4**, ATXN2}=>{ATXN2}	H2AFZP5, ATXN2
Adrenergic signaling in cardiomyocytes	KEGG	2	ATP2B4, CREB5	CO: {ZNF366, **ATP2B4**}=>{ZNF366} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ZNF366, AC097713.3
Hippo signaling pathway	KEGG	2	CSNK1E, DLG2	CO: {DDX18, **CSNK1E** }=>{**CSNK1E** } CO: {**DLG2**}=>{**DLG2**.3}	DDX18
cGMP-PKG signaling pathway	KEGG	2	ATP2B4, CREB5	CO: {ZNF366, **ATP2B4**}=>{ZNF366} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ZNF366, AC097713.3
Transcriptional misregulation in cancer	KEGG	2	HDAC2, SIX1	CA: {**HDAC2**, ACAD10}=>{BRAP} CO: {**SIX1**, PDE8A}=>{SLC28A1}	ACAD10, BRAP, PDE8A, SLC28A1

Huntington's disease	KEGG	2	HDAC2, CREB5	CA: {**HDAC2**, ACAD10}=>{BRAP} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ACAD10, BRAP, AC097713.3
Viral carcinogenesis	KEGG	2	HDAC2, CREB5	CA: {**HDAC2**, ACAD10}=>{BRAP} CO: {AC097713.3, **CREB5**}=>{AC097713.3}	ACAD10, BRAP, AC097713.3
Regulation of actin cytoskeleton	KEGG	2	ARHGEF4, ITGA9	CO: {AC097713.3, **ARHGEF4**}=>{AC097713.3} CO: {**ITGA9**}=>{**ITGA9**}	AC097713.3
Rap1 signaling pathway	KEGG	2	MAGI1, DOCK4	CO: {**MAGI1**, ATXN2}=>{ATXN2} CO: {AOX1, **DOCK4**}=>{**SGOL2**}	ATXN2, AOX1
Biosynthesis of antibiotics	KEGG	2	ALDH2, HK1	CA: {ACAD10}=>{**ALDH2**} CA: {**ALDH2,** RP3-462E2.3}=>{**ALDH2**} CO: {ATXN2, **ALDH2**, ACAD10}=>{ATXN2} CO: {**HK1**, GRIK1}=>{GRIK1}	ACAD10, RP3-462E2.3, ATXN2, GRIK1
Pathways in cancer	KEGG	2	LAMB4, HDAC2	CA: {**HDAC2**, ACAD10}=>{BRAP} CO: {**LAMB4**, ALPK3}=>{ALPK3}	ACAD10, BRAP, ALPK3

Table 5-3: Relevant pathways identified in KEGG and Reactome

As observed in Table 5-3 several genes from the top rules were identified in biological pathways relevant to obesity. To demonstrate the results, three of the relevant obesity-related pathways identified were reported: metabolism pathway (super pathway), metabolism of lipids (contained pathway) and FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes pathway (super pathway). The explanative information reported in the next sections has been extracted from Reactome reports (Fabregat et al. 2017; Sidiropoulos et al. 2017; Fabregat, Jupe, et al. 2018; Fabregat, Korninger, et al. 2018), supported by the European Bioinformatics Institute, New York University Langone Medical Center, Ontario Institute for Cancer Research and Oregon Health and Science University (Joshi-Tope 2004; Croft et al. 2014).

5.1.1.1.1 Metabolism Pathway

A total of seven genes from the most significant rules were mapped into metabolic processes as shown in Table 5-3. These genes are Aldehyde Dehydrogenase 2 Family Member (ALDH2), Ectonucleoside Triphosphate Diphosphohydrolase 4 (ENTPD4), Solute Carrier Family 10 Member 2 (SLC10A2), Aldehyde Oxidase 1 (AOX1), Acyl-CoA Dehydrogenase Family Member 10 (ACAD10), Myotubularin-Related Protein 7 (MTMR7) and Hexokinase 1 (HK1). A diagram for the metabolism pathway is depicted in Figure 5-1. It can be noted that those sub pathways (contained pathways) with mapping genes from the rules are identified by a vertical olive-green line in the figures.

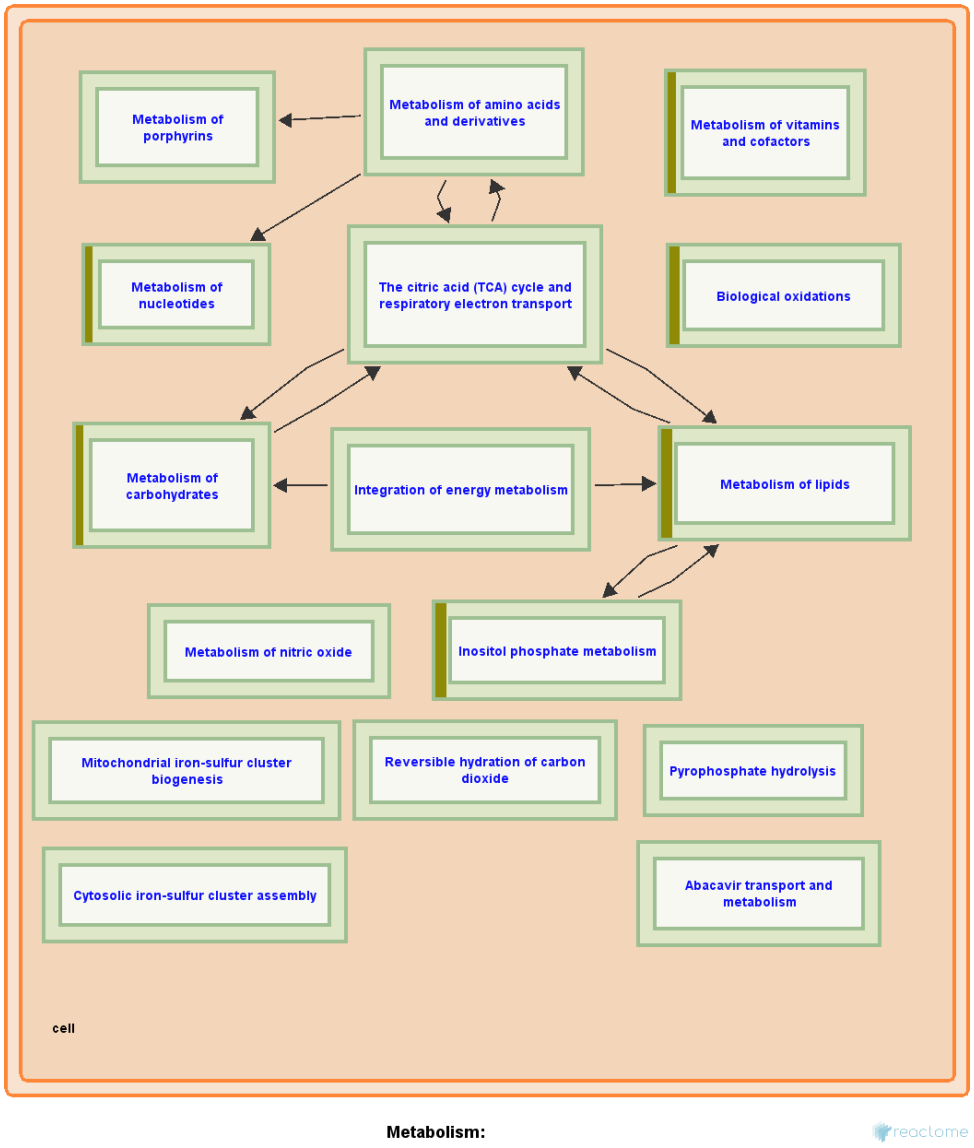


Figure 5-1: Metabolism pathway (Jassal 2011)

Metabolic processes in human cells are important as they generate energy via different processes (Komoda & Matsunaga 2015). They represent step-by-step interconnected biochemical reactions that transform substrate molecule/s via a series of metabolic intermediates, ultimately producing a final product/s.

Processes in metabolism pathway with mapped genes are reported and a diagram for each of them is provided below.

Metabolism of carbohydrates

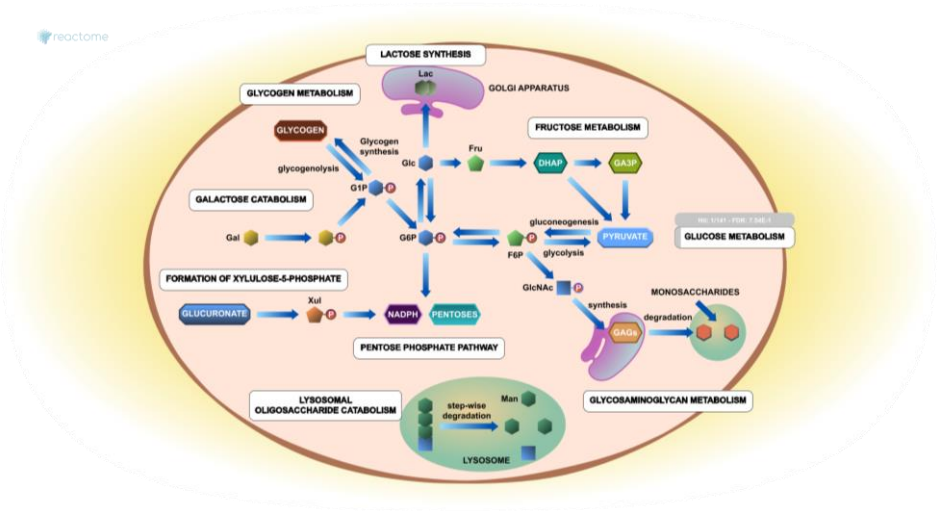


Figure 5-2: Metabolism of carbohydrates (D'Eustachio & Schmidt 2003)

The gene from the rules involved in the metabolism of carbohydrates is HK1. Starches and sugars are major constituents of the human diet and the catabolism of monosaccharides, notably glucose, derived from them is an essential part of human energy metabolism (Dashty 2013).

Inositol phosphate metabolism

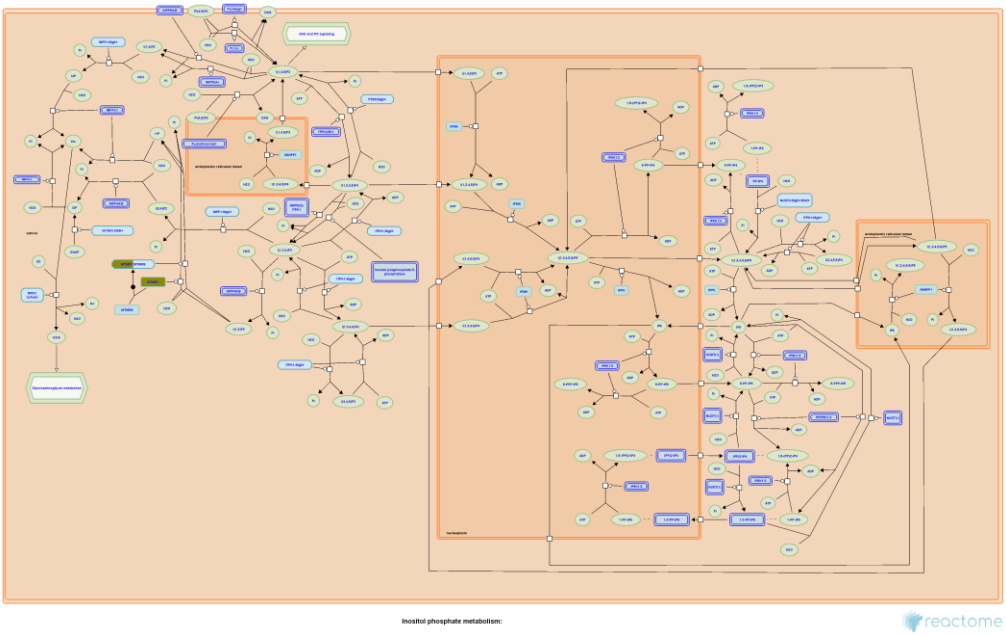


Figure 5-3: Inositol phosphate metabolism (Williams 2011a)

The gene from the rules involved in Inositol phosphate metabolism is MTMR7. Inositol phosphates (IPs) are molecules involves in signalling processes in eukaryotes.

Metabolism of lipids

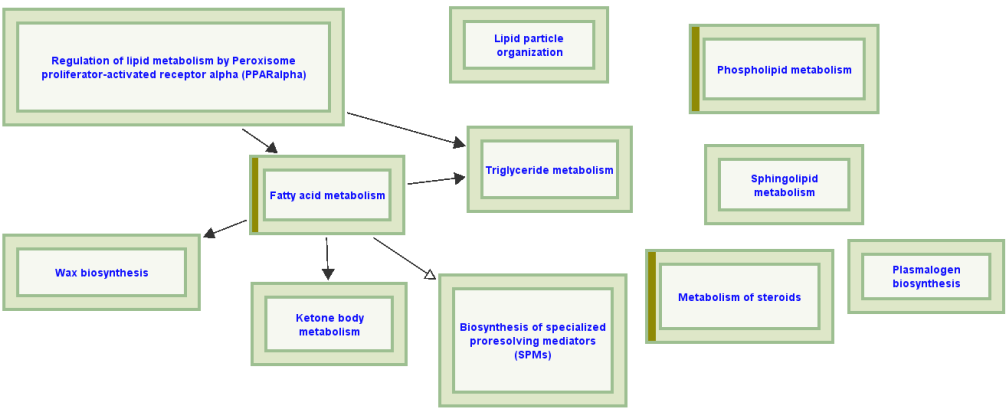


Figure 5-4: Metabolism of lipids (Jassal, Gillespie, Gopinathrao & Peter D'Eustachio 2007)

The genes from the rules involved in the metabolism of lipids are ACAD10, MTMR7 and SLC10A2. Information about this sub pathway is provided in the next section.

Metabolism of nucleotides

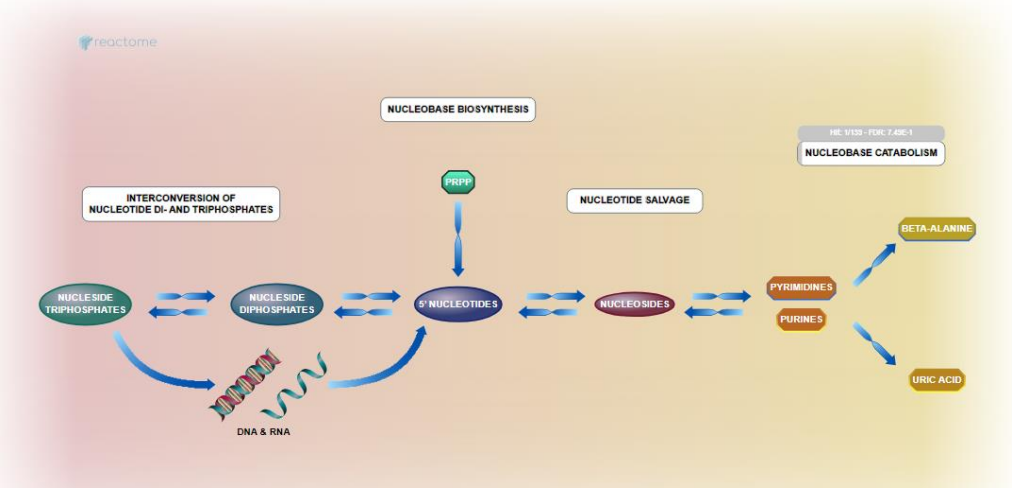


Figure 5-5: Metabolism of nucleotides (Jassal 2003)

The gene from the rules involved in the metabolism of nucleotides is ENTPD4. Nucleotides and their derivatives are used in different processes, including short-term energy storage (ATP, GTP), intra- and extracellular signaling (cAMP; adenosine), as enzyme cofactors (NAD, FAD), and for DNA and RNA synthesis. Additionally, these processes are of major clinical interest as they are the means by which nucleotide analogues used as anti-viral and anti-tumor drugs are taken up by cells, activated, and catabolized (Welin & Nordlund 2010).

Metabolism of vitamins and cofactors

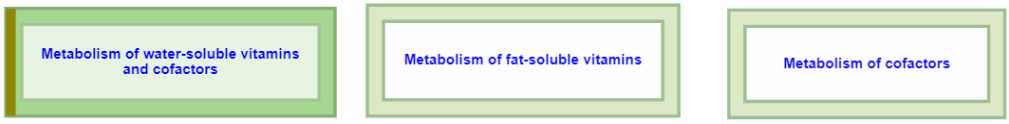


Figure 5-6: Metabolism of vitamins and cofactors (Jassal 2007b)

The gene from the rules involved in the metabolism of vitamins and cofactors is AOX1. Vitamins belong to a diverse group of organic compounds that can be classified based on their solubility in water (water-soluble) or fat (fat-soluble) and are typically not synthesised (or synthesised in limited amounts) by human cells. Furthermore, in small amounts, vitamins are necessary in the diet and have various biochemical roles. Several processes dependent on vitamin-requiring reactions are associated with diverse and severe group of diseases and vitamin deficiencies; these include aspects of intermediary metabolism, vision, bone formation, and blood coagulation.

Biological oxidations

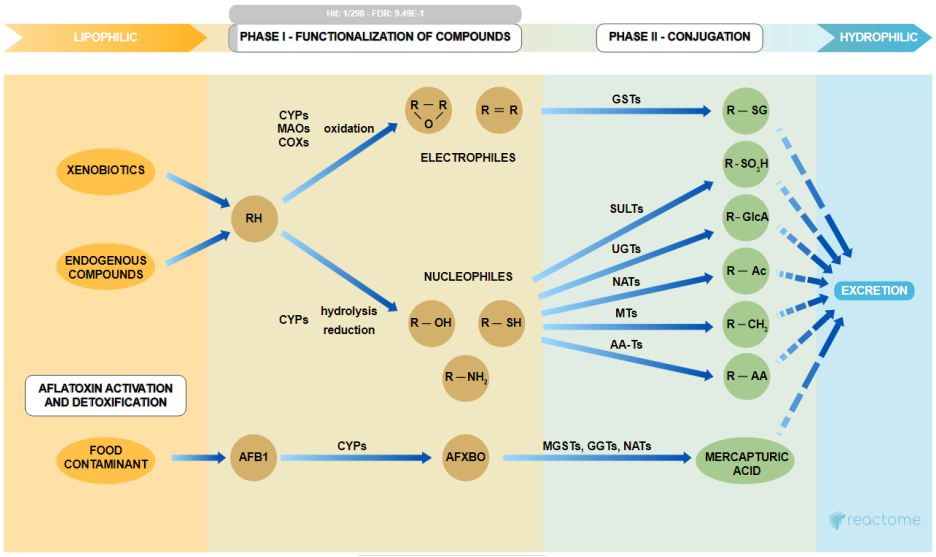


Figure 5-7: Biological oxidations (Jassal 2008)

The gene from the rules involved in biological oxidations is ALDH2.

5.1.1.1.2 Metabolism of Lipids

As mentioned earlier, three entities/genes were found in the Metabolism of lipids pathway: ACAD10, MTMR7 and SLC10A2. The Metabolism of lipids pathway is shown in Figure 5-8.

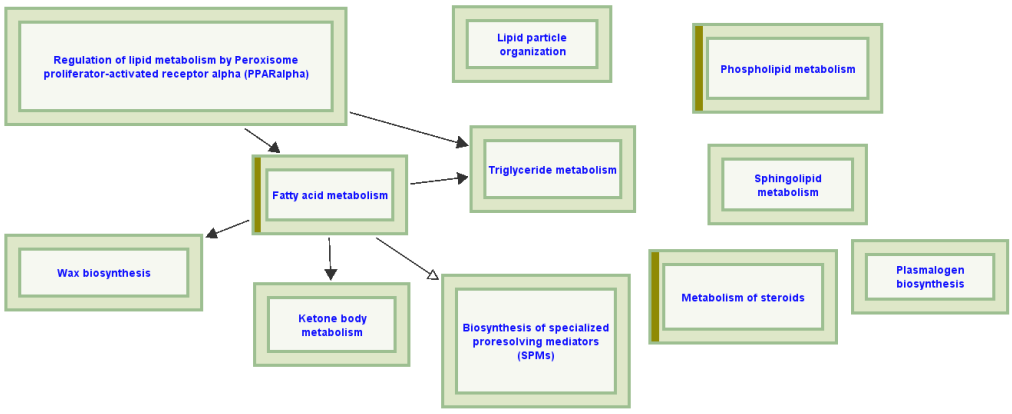


Figure 5-8: Metabolism of lipids pathway

Lipids are hydrophobic chemically diverse molecules with a wide range of roles in human biology. Lipids are responsible of many key roles (Vella 2008), including:

- They represent a major source of energy (fatty acids, triacylglycerols, and ketone bodies).
- Are major constituents of cell membranes (cholesterol and phospholipids).
- Play a major role in their own digestion and uptake (bile salts).
- Take part in numerous signaling and regulatory processes (steroid hormones, eicosanoids, phosphatidylinositols, and sphingolipids).

Aspects of lipid metabolism with mapping genes from the rules are depicted below.

Fatty acid metabolism

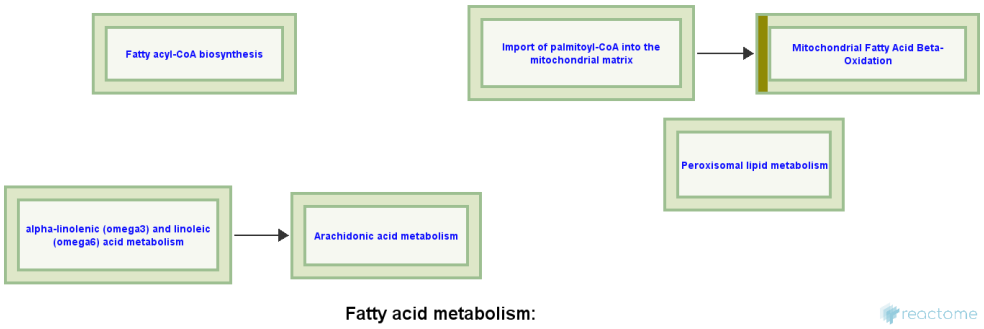


Figure 5-9: Fatty acid metabolism (Jassal, Gillespie, Gopinathrao & P D'Eustachio 2007)

The gene involved in this process located in the metabolism of lipids pathway is ACAD10. The synthesis and breakdown of fatty acids are a central part of human energy metabolism (Vella 2008). Processes annotated in this module include the synthesis of fatty acids from acetyl-CoA, mitochondrial and peroxisomal breakdown of fatty acids, and the metabolism of eicosanoids and related molecules.

Phospholipid metabolism

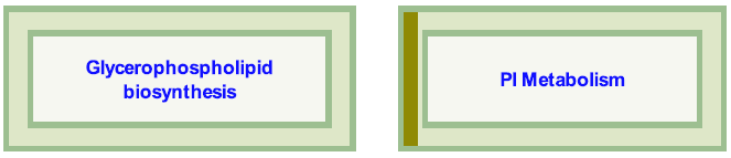


Figure 5-10: Phospholipid metabolism (Williams 2011b)

The gene involved in this process located in the metabolism of lipids pathway is MTMR7. Phospholipids contain a polar head group and two long-chain fatty

acyl moieties, one of which is generally unsaturated. These molecules are a major constituent of cellular membranes, where their diverse structures and asymmetric distributions play major roles in determining membrane properties (Dowhan 1997).

Metabolism of steroids

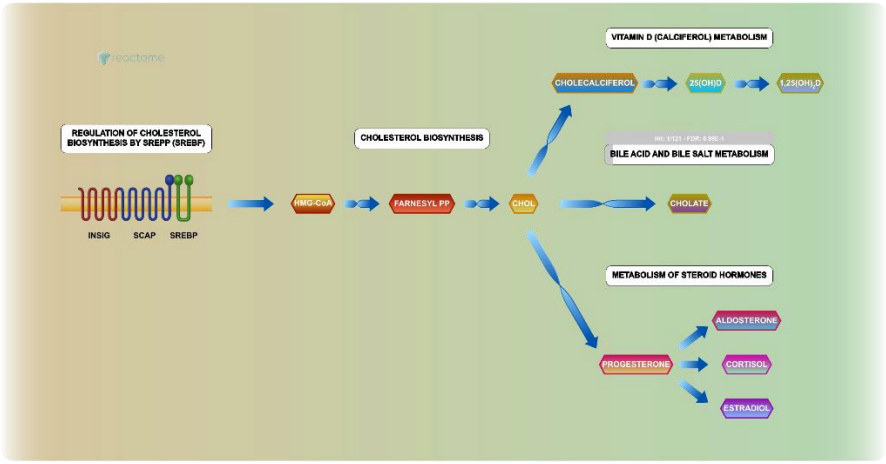


Figure 5-11: Metabolism of steroids (Jassal 2007a)

The gene involved in this process located in the metabolism of lipids pathway is SLC10A2. Three groups of molecules synthesised from steroids, include cholesterol and bile acids and salts, steroid hormones, and vitamin D. In this module, pathways for the synthesis of cholesterol from HMG-CoA (hydroxymethylglutaryl-coenzyme A), and for its conversion to bile acids and salts, steroid hormones, and vitamin D are annotated, together with the SREBP-mediated regulatory process that normally links the rate of cholesterol synthesis to levels of cellular cholesterol (Brown & Goldstein 2009).

5.1.1.1.3 FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes

Another important obesity related pathway where genes from the rules were mapped to, is the FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes pathway (Orlic-Milacic 2018a), depicted in Figure 5-12. Two entities/genes from the 132 genes forming the top 300 rules in cases and controls were mapped to this pathway: Neuropeptide Y (NPY) and Histone Deacetylase 2 (HDAC2). Processes with mapping genes from the rules are identified by a vertical olive-green line in the figures.

FOXO transcription factors regulate transcription of several genes whose protein products are secreted from hypothalamic neurons to control appetite and food intake: NPY gene, AGRP gene and POMC gene. At low insulin levels (characteristic of starvation) FOXO transcription factors bind to insulin responsive elements (IRES) in the regulatory regions of NPY, AGRP and POMC gene. FOXO1 directly stimulates transcription of the NPY gene, encoding neuropeptide-Y, and the AGRP gene, encoding Agouti-related protein, which both stimulate food intake (Kim et al. 2006; Hong et al. 2012).

A gene involved in lipid homeostasis and regulated by FOXOs, is the glucokinase (GCK) gene. FOXO1, acting with the SIN3A: HDAC complex, directly represses the GCK gene transcription, thus repressing lipogenesis in the absence of insulin (Langlet et al. 2017).

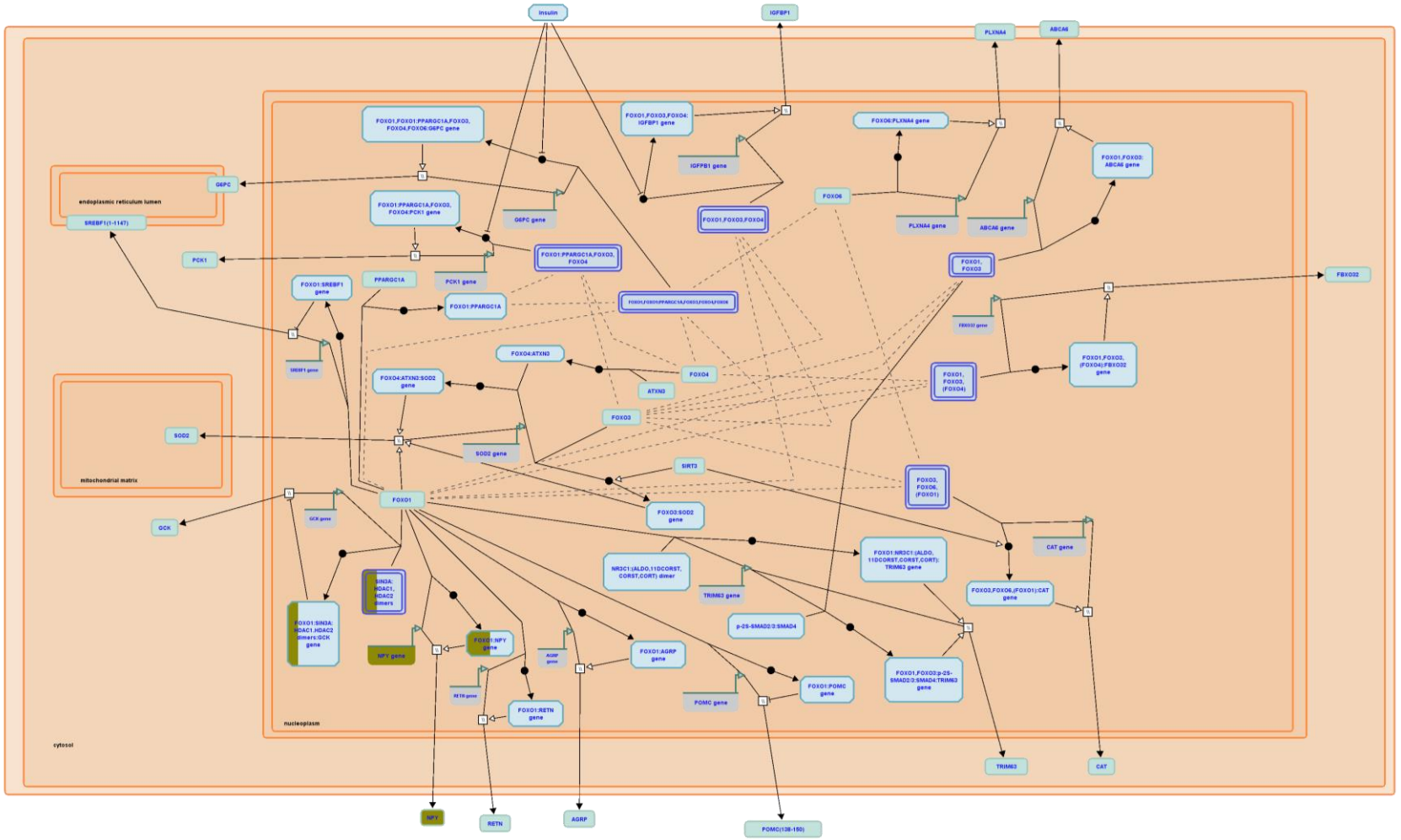


Figure 5-12: FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes pathway (Orlic-Milacic 2018a)

FOXO1 binds NPY gene promoter

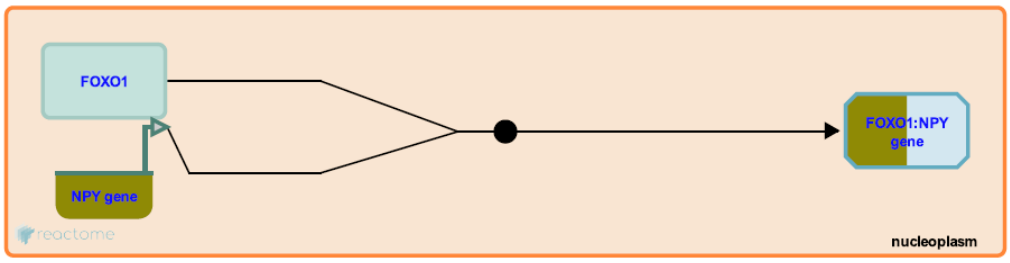


Figure 5-13: FOXO1 binds NPY gene promoter (Orlic-Milacic 2018c)

In the hypothalamic orexigenic neurons, FOXO1 binds to the insulin responsive elements (IREs) in the promoter of the NPY gene (Kim et al. 2006), encoding neuropeptide-Y.

NPY gene expression is stimulated by FOXO1

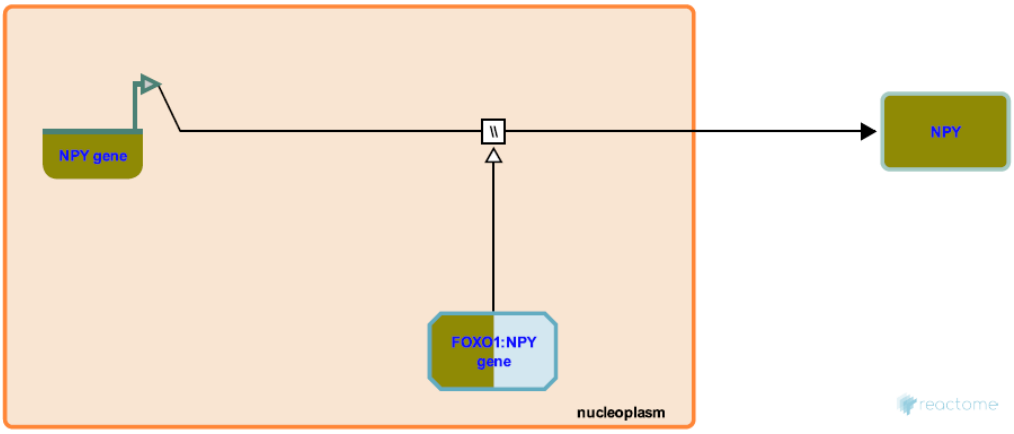


Figure 5-14: NPY gene expression is stimulated by FOXO1 (Orlic-Milacic 2018e)

Transcription of the NPY gene is directly stimulated by FOXO1 (encoding neuropeptide-Y) in hypothalamic orexigenic neurons. NPY stimulates food intake and weight gain. Insulin and leptin, through PI3K/AKT signaling, inhibit FOXO1-mediated upregulation of NPY expression (Kim et al. 2006). NPY may act through a positive feedback loop to increase the transcriptional activity of FOXO1 through the PKA/CREB pathway (Hong et al. 2012).

FOXO1 and SIN3A: HDAC complex bind GCK gene promoter

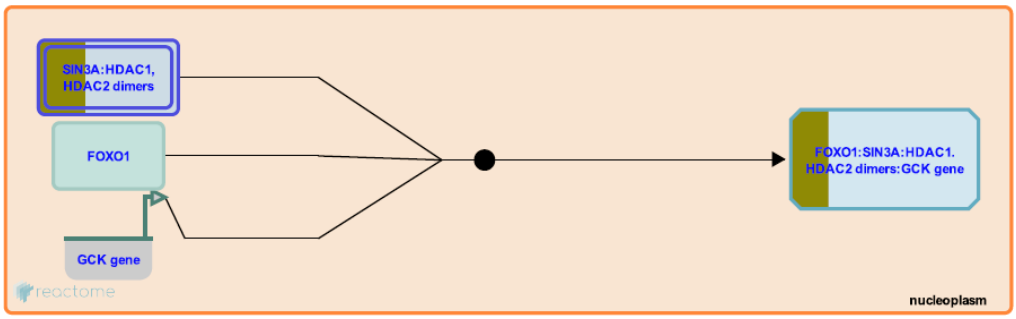


Figure 5-15: FOXO1 and SIN3A: HDAC complex bind GCK gene promoter (Orlic-Milacic 2018b)

Based on studies in mice, FOXO1 recruits transcriptional repressor SIN3A and histone deacetylases (HDACs) of the I class to the promoter of the GCK gene, encoding glucokinase (Langlet et al. 2017).

GCK gene expression is inhibited by FOXO1, SIN3A and HDACs

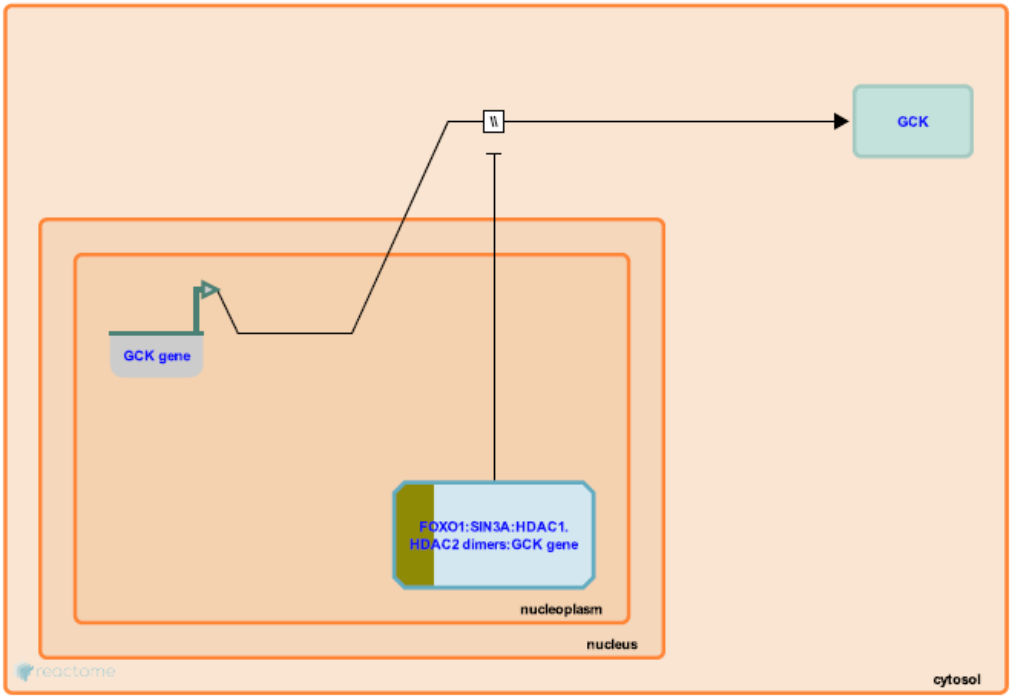


Figure 5-16: GCK gene expression is inhibited by FOXO1, SIN3A and HDACs (Orlic-Milacic 2018d)

The combination of FOXO1, SIN3A and histone deacetylases (HDACs), represses transcription of the GCK gene, encoding glucokinase as reported in studies in mice. Insulin interferes with FOXO1-mediated repression of GCK expression, resulting in upregulation of GCK and stimulation of lipogenesis (Langlet et al. 2017).

Evidences extracted from the highlighted pathways support the potential of using ARM in the identification of epistasis without initial knowledge, based solely on statistical approaches.

5.1.2 Biological Filtering using GSEA: Proof of Concept

As discussed earlier in this thesis, GWAS identify genetic variants that have been independently tested, and it only consider a number of the most significant ones for subsequent experiments (i.e. statistical filtering). From a system point of view, this is seen as a limitation since the combined effect of less significant variants is overlooked. To overcome this limitation, the *improved* gene-set enrichment analysis for GWAS (*i*-GSEA4GWAS) web tool was utilised (Zhang et al. 2010). This tool performs GSEA on GWAS data using SNP label permutation to analyse the P-values reported by association analysis (i.e. using logistic regression). To ensure comprehensiveness and reliability, *i*-GSEA4GWAS relies on a collection of pathways and annotated gene sets curated from Molecular Signatures Database (MSigDB) (Liberzon et al. 2015; Subramanian et al. 2005). MSigDB includes canonical pathways and gene sets integrated and curated from a variety of reference knowledge based sources, including KEGG and curated gene ontology (GO) terms (Ashburner et al. 2000) among other sources. By using this tool, the aim is to identify pathways or gene

sets correlated with obesity as a biological filtering strategy, while providing new insights into epistasis using ARM. This approach is presented, thus, as an alternative to the statistical filtering approach used in the initial method of SAERMA.

Following the identification of significant pathways associated with GWAS results, a series of analysis were conducted using ARM for each of the pathways as well as the union of the genes within all pathways. The analysis concluded by testing a number of classifiers (using MLP) to measure the performance of the variants within the identified pathways when discriminating between cases and controls. A diagram with the steps conducted is shown in Figure 5-17.

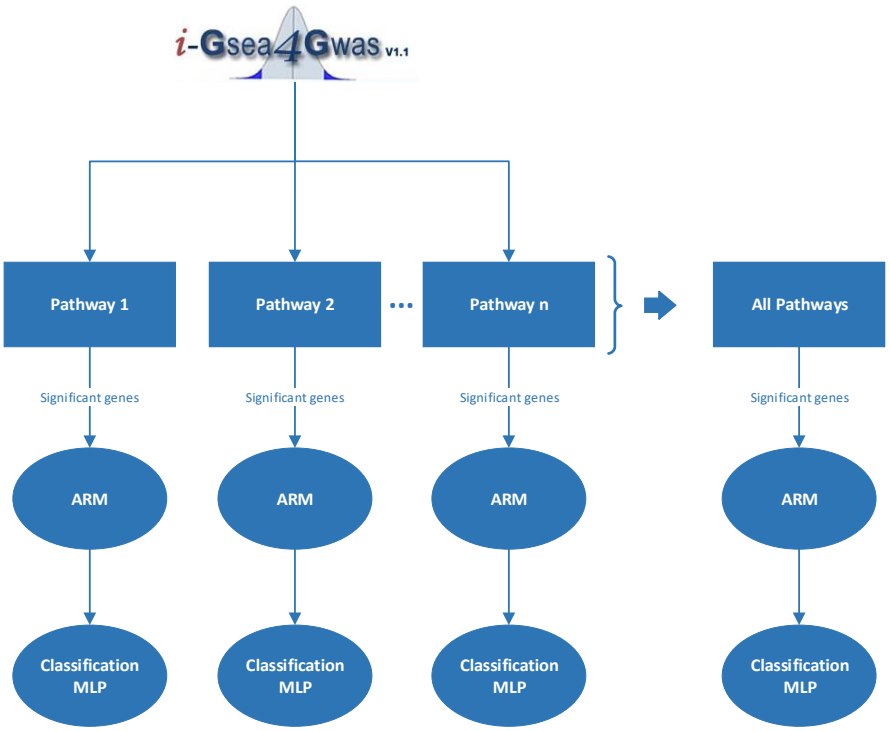


Figure 5-17: Diagram for proposed proof of concept biological filtering approach

In the following section, the results obtained using the steps depicted in Figure 5-17 are presented.

5.1.2.1 *i*-GSEA4GWAS Results

All the variants and corresponding P-values after association analysis with logistic regression (240,950 SNPs) were used as input to the *i*-GSEA4GWAS web tool. Next, the identification of pathway/gene sets associated with obesity was conducted using the default tool parameters.

After running *i*-GSEA4GWAS, four canonical pathways were identified based on enrichment analysis. Additionally, 164 GO terms were also found and reported in Appendix F. In Table 5-4, a list with the significant canonical pathways identified is reported. The table include the pathway name, description, gene set P-value, gene set false discovery rate (FDR) and the number of significant genes involved in each pathway. These results are reported as a proof of concept for future work so a deeper investigation of the findings will be required.

Pathway Name	**Description**	**Gene Set P-Value**	**Gene Set FDR**	**Mapped Genes**
WNT Signaling	Wnt Signaling genes	0.002	0.075	26
ECM Receptor Interaction	Genes involved in ECM-receptor interaction	0.001	0.076	42
Peptide GPCRS	Peptide G protein-coupled receptors (GPCRs)	< 0.001	0.089	28
Prostate Cancer	Genes involved in prostate cancer	0.002	0.090	30

Table 5-4: Canonical pathways identified by i-GSEA4GWAS

For each of the canonical pathways identified, a table with the significant gene set and the pathway diagram generated by KEGG are provided below. For each associated gene, the SNP ID, the -log(P-value) from association analysis (logistic regression), the chromosome where it is located, and gene start-end positions are given. The figures display pathway maps with the genes

highlighted in red to ease biological interpretation in a network context, a feature provided by the DAVID Functional Annotation Tool.

5.1.2.1.1 WNT Signaling Pathway

In Table 5-5 the genes associated with the Wnt Signaling pathway are listed.

Gene Name	SNP ID	-log(P-value)	Chrom #	Gene Start	Gene End
FZD5	rs10188753	4.294992	2	208627310	208634287
WNT11	rs593241	2.74958	11	75897370	75917574
FZD10	rs615099	2.5247648	12	130647032	130650284
WNT7A	rs13077668	2.4329736	3	13857755	13921618
CCND2	rs7307707	2.2549253	12	4382902	4414521
CSNK1E	rs17753394	2.2016256	22	38686697	38794527
MAPK10	rs7439032	2.1237822	4	86936276	87374296
CTNNB1	rs442115	2.100946	3	41236328	41301587
PRKD1	rs225987	2.08302	14	30045687	30396948
JUN	rs2764900	2.0379622	1	59246465	59249785
MYC	rs4733616	1.9792246	8	128747680	128753674
PRKCD	rs2230493	1.8843895	3	53190025	53226733
APC	rs454886	1.870955	5	112043218	112181936
FZD6	rs6990501	1.8335699	8	104311100	104345087
PRKCE	rs6725257	1.7708303	2	45878484	46415129
PPP2R5E	rs1255771	1.7687854	14	63838075	64010092
PRKCH	rs2255146	1.740645	14	61788435	62017698
PRKCA	rs9889698	1.7110804	17	64298926	64806862
FZD8	rs2696309	1.6639407	10	35927177	35930362
CCND1	rs587230	1.6343249	11	69455873	69469241
PRKCQ	rs1409874	1.5793842	10	6469105	6622263
FZD2	rs12450493	1.5554868	17	42634925	42636907
SFRP4	rs2722279	1.501276	7	37945534	38065297
WNT5B	rs4765829	1.4882503	12	1726222	1756409
FZD7	rs10931982	1.4625587	2	202899310	202903160
LDLR	rs2569538	1.418961	19	11200057	11244506

Table 5-5: Significant genes in the Wnt Signaling pathway

Figure 5-18 shows the pathway diagram for the Wnt Signaling pathway with the associated genes highlighted (see red stars).

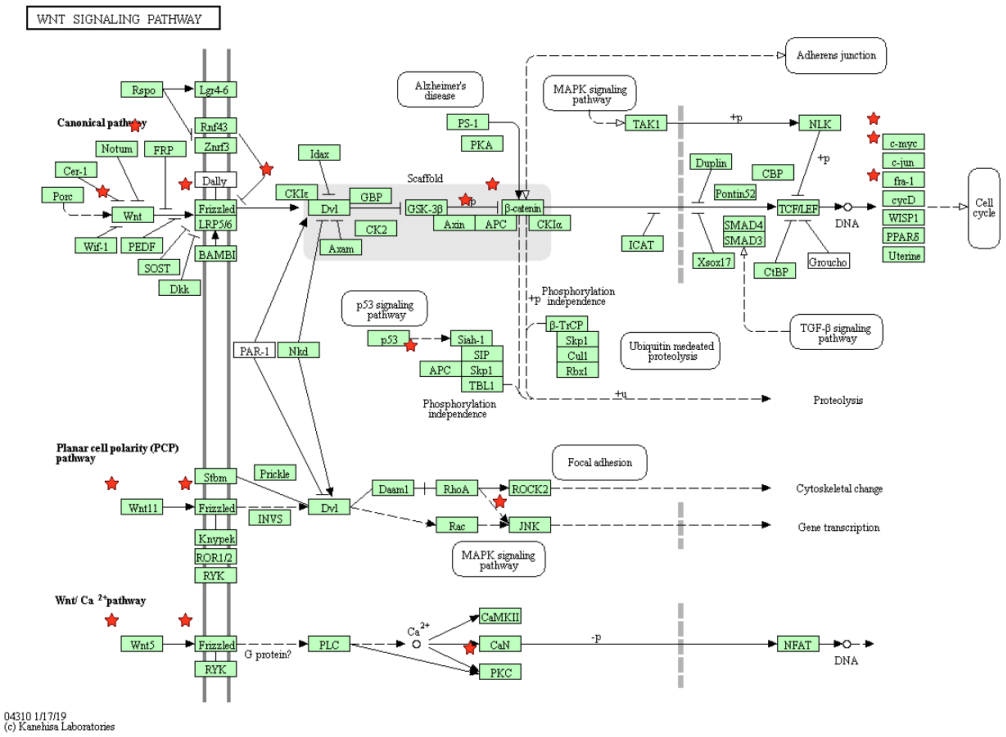


Figure 5-18: WNT Signaling pathway

5.1.2.1.2 ECM Receptor Interaction Pathway

In Table 5-6 the genes associated with the ECM Receptor Interaction pathway are listed.

Gene Name	SNP ID	-log(P-value)	Chrom #	Gene Start	Gene End
VWF	rs1063856	3.019951	12	6058040	6233836
SV2C	rs17564993	2.5041783	5	75379304	75621416
ITGA9	rs4678980	2.4678829	3	37493606	37865005
COL4A2	rs4773184	2.4429743	13	110958159	111165374
LAMB4	rs1627354	2.4280584	7	107663993	107770801
ITGA6	rs1076597	2.3964226	2	173292082	173371181
ITGA2	rs3212578	2.3619103	5	52285183	52390609
THBS4	rs2288394	2.3162327	5	79330991	79379110
LAMC2	rs17479287	2.27311	1	183155174	183214035
RELN	rs362777	2.2685723	7	103112231	103629963
LAMA2	rs17057233	2.252355	6	129204286	129837714
CD36	rs1511682	2.2392006	7	79998891	80308593
COL6A3	rs12328617	2.1219954	2	238232646	238323018
ITGA8	rs1925809	2.0971272	10	15555948	15762124
ITGB1	rs11009132	2.0827045	10	33189247	33294720
ITGA1	rs4865534	2.0795636	5	52083754	52252327
COL3A1	rs17241561	2.0243845	2	189839046	189877472
COL5A1	rs11103417	1.9854795	9	137533620	137736686
LAMB1	rs6943225	1.9593977	7	107564244	107643804
TNR	rs2012430	1.9262817	1	175291935	175712906
COL6A1	rs2839086	1.924453	21	47401651	47424964
ITGA11	rs8041354	1.9037855	15	68594050	68724492
TNC	rs1330368	1.7859514	9	117782806	117880486
ITGA4	rs6740847	1.779892	2	182321619	182400914
SDC2	rs6983702	1.7302536	8	97505882	97624037
COL2A1	rs7299271	1.7272304	12	48366748	48398285
SV2B	rs7182678	1.7203331	15	91643515	91844539
ITGA10	rs1109216	1.7153437	1	145524891	145543868
COL1A2	rs17166182	1.7046529	7	94023873	94060544
CD44	rs10128562	1.7020208	11	35160417	35253946
COL11A1	rs7537288	1.6968039	1	103342023	103574052
THBS2	rs4708599	1.6850799	6	169615875	169654139
IBSP	rs2627704	1.6627405	4	88720710	88733587
HSPG2	rs747546	1.5295895	1	22148738	22263790
LAMA4	rs9374352	1.4971634	6	112429134	112575849
FN1	rs10804242	1.488384	2	216225163	216300895
ITGA3	rs9890077	1.4320331	17	48133340	48167848
SDC1	rs17652287	1.3715088	2	20400558	20425194
COL6A2	rs9975613	1.3639135	21	47518011	47552763

COL4A1	rs2275843	1.3638132	13	110801318	110959496
LAMA1	rs6650624	1.3352641	18	6941743	7117813
SDC3	rs4949184	1.3220283	1	31342313	31381608

Table 5-6: Significant genes in the ECM Receptor Interaction pathway

Figure 5-19 shows the pathway diagram for the ECM Receptor Interaction pathway with the associated genes highlighted (see red stars).

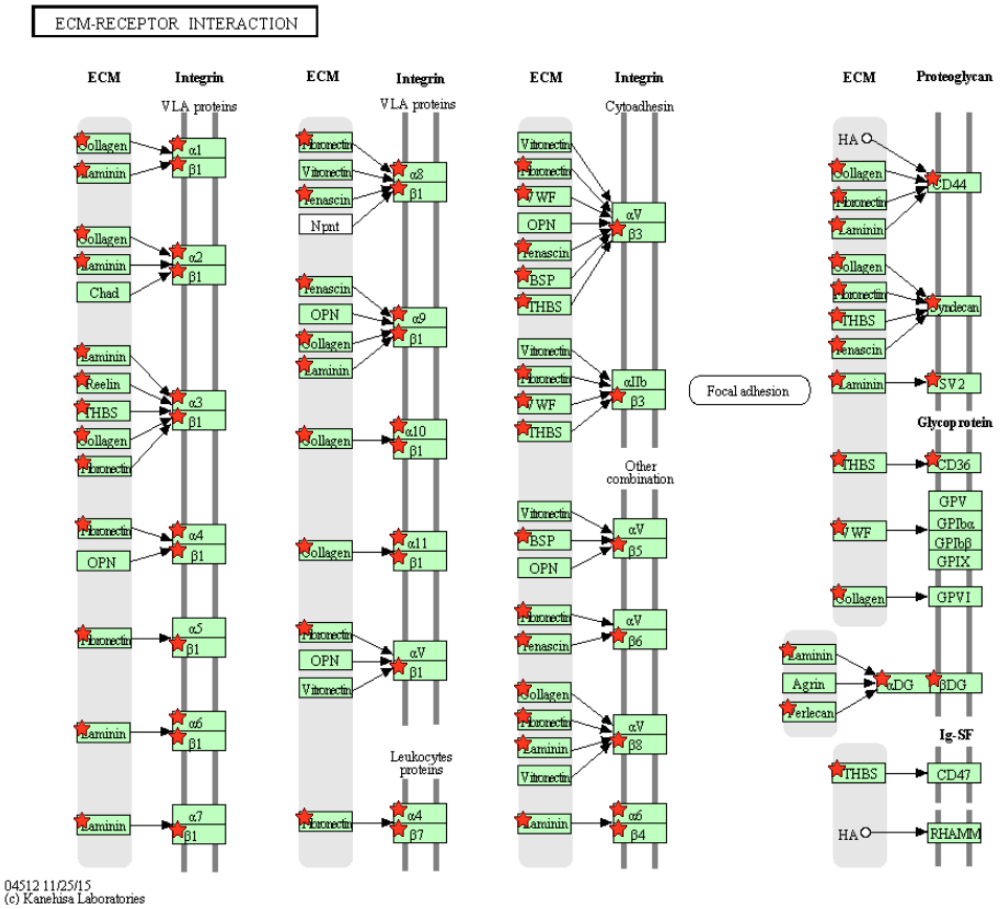


Figure 5-19: ECM Receptor Interaction pathway

5.1.2.1.3 Peptide GPCRS Pathway

In Table 5-7 the genes associated with the Peptide GPCRS pathway are listed.

Gene Name	SNP ID	-log(P-value)	Chrom #	Gene Start	Gene End
TSHR	rs12883673	3.2008662	14	81421775	81612646
SSTR4	rs12480596	2.8787684	20	23016057	23017314
FSHR	rs10469872	2.698319	2	49189296	49381676
OPRM1	rs1937834	2.183825	6	154331636	154568001
EDNRA	rs1400558	2.1297717	4	148401907	148466106
ATP8A1	rs13139219	2.1161683	4	42410390	42659122
TACR3	rs12501131	2.09023	4	104507188	104640973
EDNRB	rs1324791	2.0281677	13	78469616	78493903
TAC4	rs8080832	1.8127614	17	47915671	47925379
TACR1	rs7576919	1.7654827	2	75276231	75426826
NMBR	rs6902780	1.6998395	6	142379467	142409936
MC4R	rs474112	1.6145725	18	58038564	58040001
TRHR	rs6469232	1.583859	8	110099724	110131813
SSTR2	rs1037257	1.5492891	17	71161160	71168060
FPR1	rs4801891	1.5422695	19	52249027	52255150
SSTR1	rs11628551	1.5297365	14	38677204	38682268
NPY2R	rs17032433	1.4866492	4	156129781	156138227
CCR2	rs4513489	1.4795165	3	46395225	46402419
CCKBR	rs11040816	1.4710833	11	6280966	6293357
OPRD1	rs1485471	1.4704413	1	29138654	29190208
CCR3	rs3136667	1.4464811	3	46205096	46308111
CCR1	rs3136667	1.4464811	3	46243200	46249887
AVPR1A	rs12815070	1.4332091	12	63540216	63546590
CX3CR1	rs4676487	1.4060497	3	39304985	39323186
NTSR2	rs7578132	1.4027437	2	11798304	11810290
BDKRB2	rs11624761	1.3831046	14	96671135	96710666
GALR1	rs2717121	1.3497896	18	74962505	74980858
CCR6	rs3798315	1.3431355	6	167525295	167553184

Table 5-7: Significant genes in the Peptide GPCRS pathway

Figure 5-20 shows the pathway diagram for the Peptide GPCRS pathway with the associated genes highlighted (see red stars).

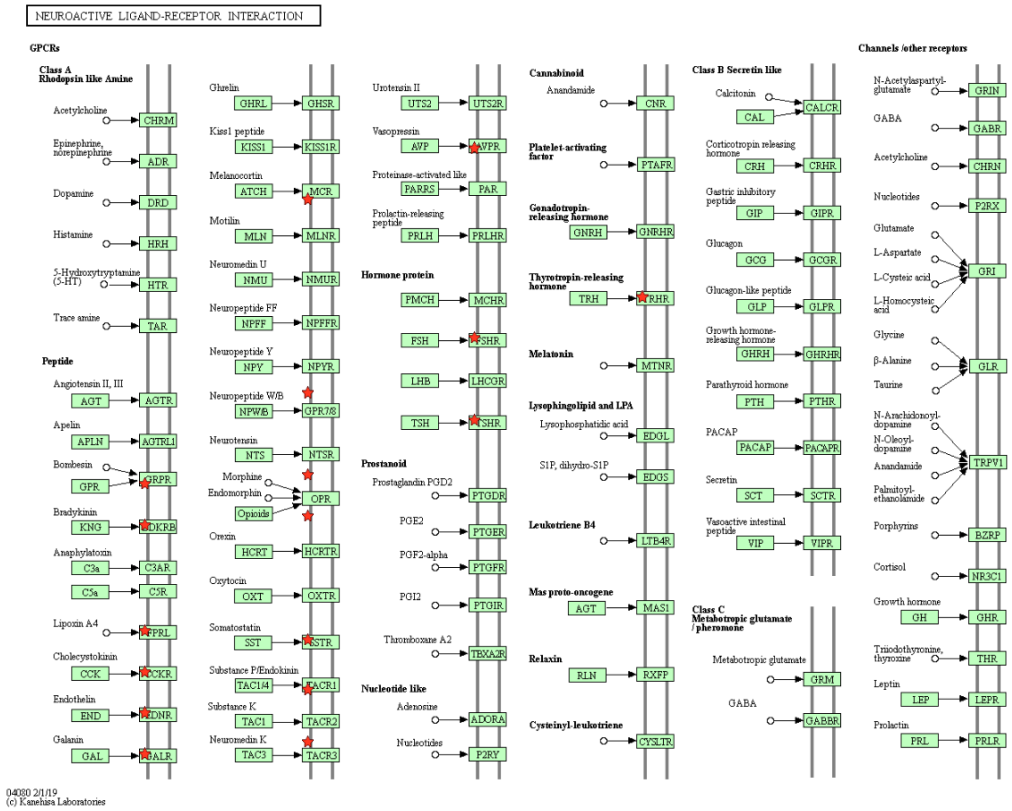


Figure 5-20: Peptide GPCRS pathway

5.1.2.1.4 Prostate Cancer Pathway

In Table 5-8 the genes associated with the Prostate Cancer pathway are listed.

Gene Name	SNP ID	-log(P-value)	Chrom #	Gene Start	Gene End
TCF7L2	rs290475	3.1709538	10	114710009	114927437
CREB5	rs42695	2.6987529	7	28338940	28865511
CCNE1	rs11672342	2.5104635	19	30302901	30315216
CREB1	rs17203016	2.4873157	2	208394461	208468155
FGFR2	rs10510097	2.4474535	10	123237848	123357972
TGFA	rs11466212	2.184555	2	70674412	70781325
CTNNB1	rs442115	2.100946	3	41236328	41301587
LEF1	rs1291490	2.073092	4	108968701	109089578
PIK3R1	rs16897333	2.0296066	5	67522462	67597649
AKT3	rs12048930	2.0193605	1	243651535	244013430
EGFR	rs11773818	1.9892762	7	55086714	55324313
EGF	rs17041230	1.9775716	4	110834047	110933422
IGF1R	rs4966035	1.975925	15	99192200	99507759
CREB3L2	rs273945	1.871924	7	137559725	137686813
BCL2	rs2849379	1.8288589	18	60790579	60987361
PIK3CD	rs7518793	1.8215986	1	9711803	9788977
E2F2	rs2075993	1.7986028	1	23832922	23857712
KRAS	rs10842514	1.7767637	12	25358182	25403854
CHUK	rs11190421	1.7039933	10	101948055	101989376
CCND1	rs587230	1.6343249	11	69455873	69469241
PDGFRA	rs6857523	1.5221555	4	55095457	55164414
NFKB1	rs13116385	1.4845235	4	103422486	103538459
PTEN	rs17322612	1.4076011	10	89622870	89731687
FOXO1	rs7327621	1.3900856	13	41129817	41240734
FGFR1	rs4739561	1.3682519	8	38268656	38326352
RAF1	rs3730269	1.3585259	3	12625100	12705725
PIK3R5	rs9895992	1.3391346	17	8782228	8869024
PDGFB	rs879180	1.3368647	22	39619364	39640756
CREB3L3	rs350885	1.3213906	19	4153629	4173050
CASP9	rs6685648	1.3139002	1	15814735	15853029

Table 5-8: Significant genes in the Prostate Cancer pathway

Figure 5-21 shows the pathway diagram for the Prostate Cancer pathway with the associated genes highlighted (see red stars).

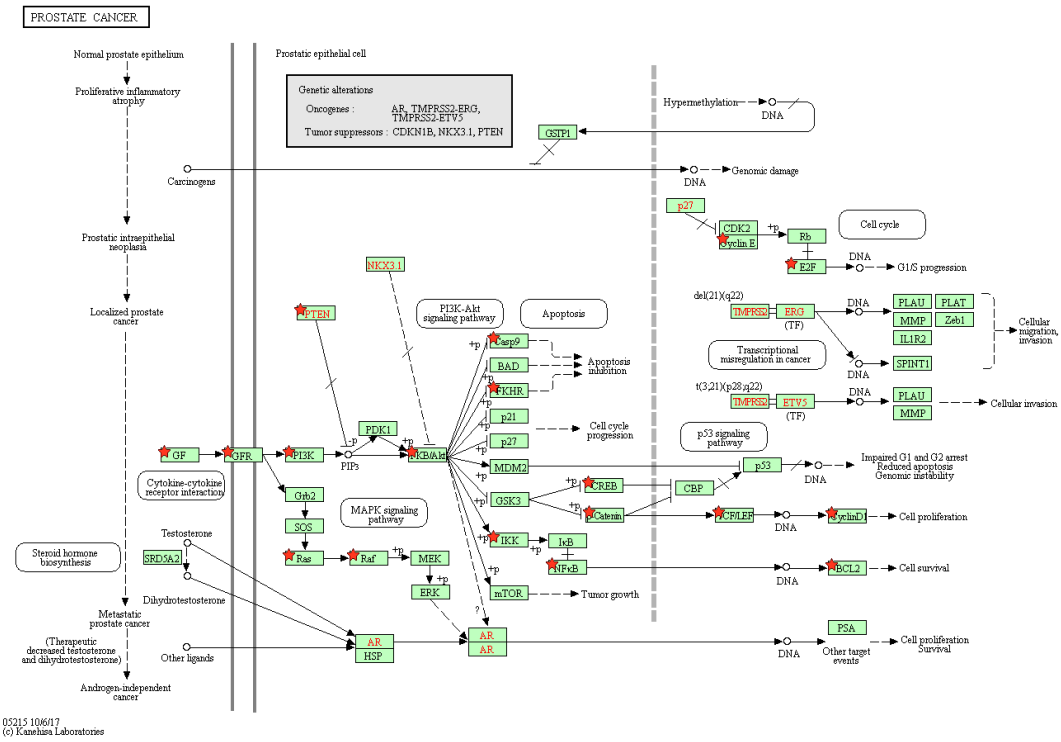


Figure 5-21: Prostate Cancer pathway

5.1.2.2 Association Rule Mining analysis of significant pathways

The association rules presented in this section are driven by biological knowledge using gene set enrichment analysis. This has the potential to identify epistasis in biologically functional genes.

The following table (see Table 5-9) provides a summary of the number of rules identified by the *Apriori* algorithm for each pathway in cases and controls, and how many of these were significant.

Pathway	# Rules CA	# Rules CAO	Sig. Rules CA	Sig. Rules CO
WNT Signaling	40	58	13	17
ECM Receptor Interaction	214	328	58	47
Peptide GPCRS	23	16	12	4
Prostate Cancer	111	161	31	44
All Pathways	4,574	6,048	1,371	1,720

Table 5-9: ARM summary for canonical pathways

Next, ARM results for each canonical pathway are presented. Moreover, results for the union of genetic variants within the four pathways are also reported.

5.1.2.2.1 ARM for WNT Signaling Pathway

In Table 5-10 the most significant rules identified in cases are listed. It can be noted that lift values are very close to 1 while chi-square values are lower than 3.84. This indicates that elements in the rules are independent as previously discussed in Chapter 3.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs2255146_A_D} => {rs12450493_A_D}	0.612	0.865	1.015	2.786
2	{rs7307707_T_D} => {rs6990501_G_D}	0.625	0.827	1.015	2.695
3	{rs2255146_A_D} => {rs2722279_C_D}	0.609	0.860	1.015	2.604
4	{rs2722279_C_D,rs593241_T_D} => {rs6990501_G_D}	0.605	0.826	1.014	2.118
5	{rs12450493_A_D,rs593241_T_D} => {rs6990501_G_D}	0.605	0.822	1.009	0.961
6	{rs17753394_G_D} => {rs12450493_A_D}	0.629	0.860	1.009	1.194
7	{rs6990501_G_D} => {rs12450493_A_D}	0.699	0.858	1.006	0.906
8	{rs6990501_G_D} => {rs2722279_C_D}	0.694	0.852	1.005	0.579
9	{rs7307707_T_D} => {rs593241_T_D}	0.656	0.869	1.005	0.440
10	{rs6990501_G_D} => {rs593241_T_D}	0.708	0.869	1.005	0.554
11	{rs2722279_C_D} => {rs2569538_A_D}	0.688	0.812	1.003	0.136
12	{rs2255146_A_D} => {rs593241_T_D}	0.613	0.867	1.002	0.068
13	{rs7307707_T_D} => {rs2722279_C_D}	0.642	0.849	1.002	0.071

Table 5-10: Rules identified in cases for the Wnt signalling pathway

The most significant rules identified in controls are reported in Table 5-11. In this occasion, although lift is not much larger than 1, the chi-square value for the top rule (rule 1) is higher than 3.84. This supports dependency of the rule {rs4733616_T_D} => {rs2569538_A_D}.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs4733616_T_D} => {rs2569538_A_D}	0.622	0.861	1.021	6.918
2	{rs12450493_A_D,rs2722279_C_D} => {rs593241_T_D}	0.638	0.823	1.014	3.137
3	{rs7439032_T_D} => {rs2722279_C_D}	0.601	0.889	1.010	1.693
4	{rs2230493_C_D} => {rs12450493_A_D}	0.630	0.892	1.009	1.539
5	{rs593241_T_D} => {rs2722279_C_D}	0.721	0.888	1.009	2.593
6	{rs4733616_T_D} => {rs12450493_A_D}	0.643	0.891	1.007	1.142
7	{rs2722279_C_D,rs7307707_T_D} => {rs2569538_A_D}	0.600	0.849	1.007	0.709
8	{rs6990501_G_D} => {rs12450493_A_D}	0.683	0.890	1.007	1.228
9	{rs593241_T_D} => {rs12450493_A_D}	0.723	0.890	1.006	1.307
10	{rs2230493_C_D} => {rs2722279_C_D}	0.624	0.885	1.005	0.520
11	{rs7307707_T_D} => {rs2569538_A_D}	0.682	0.848	1.005	0.595
12	{rs587230_A_D} => {rs2569538_A_D}	0.605	0.847	1.004	0.281
13	{rs17753394_G_D} => {rs12450493_A_D}	0.609	0.888	1.004	0.254
14	{rs7439032_T_D} => {rs12450493_A_D}	0.600	0.888	1.003	0.199
15	{rs6990501_G_D} => {rs2569538_A_D}	0.649	0.846	1.003	0.201
16	{rs587230_A_D} => {rs12450493_A_D}	0.632	0.886	1.002	0.050
17	{rs6725257_A_D} => {rs12450493_A_D}	0.602	0.886	1.001	0.019

Table 5-11: Rules identified in controls for the Wnt pathway

The network visualisation plots for the most significant rules identified in cases and controls for the Wnt signalling pathway are depicted in Figure 5-22 and Figure 5-23 respectively. The visualisation indicates, in a visual way, how the genetic variants in the rules interact, which rules are more frequent, and their level of correlation based on graph-based visualization as detailed in Section 3.5.4.

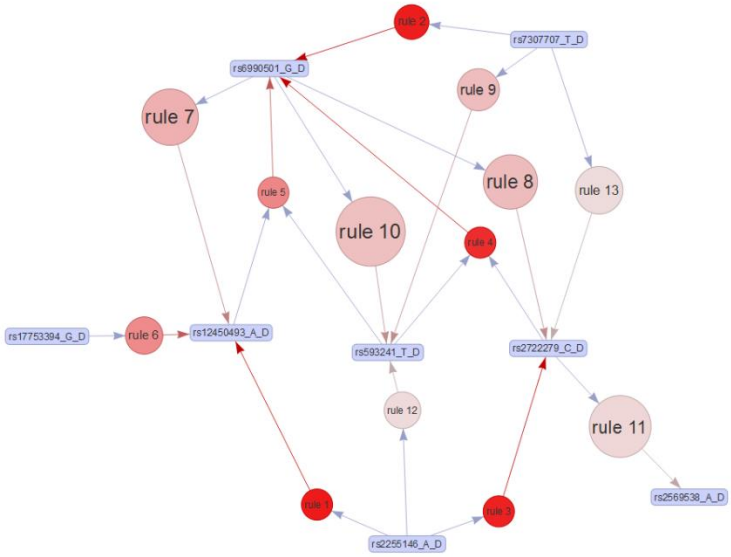


Figure 5-22: Rule visualisation network for the Wnt signalling pathway in cases

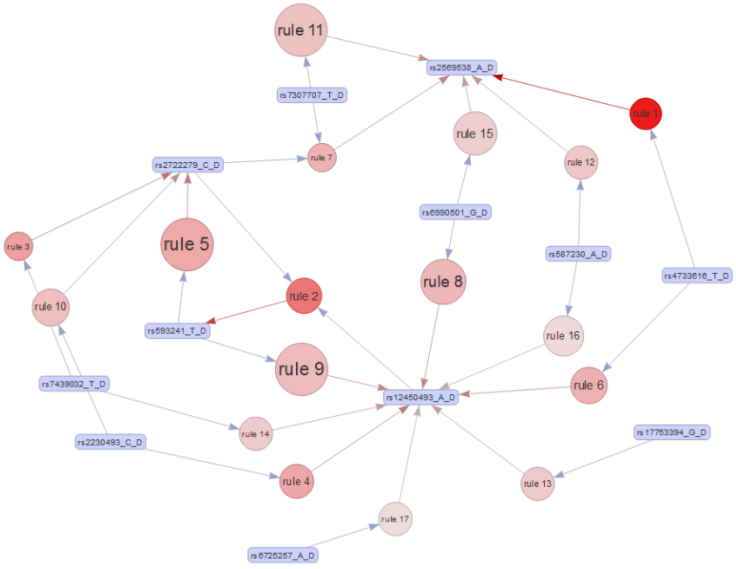


Figure 5-23: Rule visualisation network for the Wnt signalling pathway in controls

5.1.2.2.2 ARM for ECM Receptor Interaction Pathway

In this case, the number of significant rules identified in cases was 58 while 47 rules were identified in controls (see Table 5-9). Since the number of rules is large, only the top 10 were listed below. For the whole list of rules identified for the ECM receptor interaction pathway please refer to Appendix G.

Therefore, the top 10 most significant rules identified for ECM receptor interaction pathway in cases are listed in Table 5-12.

Based on ARM dependence framework, only rules 1, 2, 3, 4, 5, 19 and 21 in cases (see Appendix G) are showing dependency (although weak) since lift > 1 and χ^2 is > 3.84.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs3212578_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.654	0.906	1.023	9.375
2	{rs17057233_C_D,rs7537288_G_D} => {rs3212578_A_D}	0.671	0.885	1.020	7.478
3	{rs7182678_G_D} => {rs12328617_A_D}	0.613	0.889	1.018	4.370
4	{rs12328617_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.667	0.877	1.017	5.206
5	{rs12328617_A_D,rs17057233_C_D} => {rs3212578_A_D}	0.678	0.882	1.017	5.528
6	{rs747546_T_D} => {rs4865534_G_D}	0.644	0.814	1.017	3.783
7	{rs17479287_A_D,rs7537288_G_D} => {rs3212578_A_D}	0.622	0.881	1.016	3.565
8	{rs17479287_A_D,rs7537288_G_D} => {rs1627354_A_D}	0.622	0.881	1.016	3.565
9	{rs9975613_C_D} => {rs1627354_A_D}	0.618	0.880	1.015	3.112
10	{rs12328617_A_D,rs1627354_A_D} => {rs17479287_A_D}	0.639	0.843	1.015	2.871

Table 5-12: Top 10 rules identified in cases for the ECM receptor interaction pathway

Similarly, the top 10 most significant rules identified for ECM receptor interaction pathway in controls are listed in Table 5-13. As observed from the table, identified rules are independent.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs1627354_A_D,rs7537288_G_D} => {rs362777_C_D}	0.631	0.848	1.014	3.440
2	{rs1627354_A_D,rs17057233_C_D} => {rs8041354_T_D}	0.610	0.802	1.014	2.489
3	{rs12328617_A_D,rs17057233_C_D} => {rs8041354_T_D}	0.611	0.801	1.012	1.811
4	{rs6983702_T_D} => {rs362777_C_D}	0.618	0.846	1.011	1.985
5	{rs7537288_G_D} => {rs8041354_T_D}	0.655	0.800	1.011	2.159
6	{rs1627354_A_D,rs8041354_T_D} => {rs362777_C_D}	0.610	0.845	1.011	1.638
7	{rs17564993_T_D} => {rs12328617_A_D}	0.627	0.920	1.010	2.592
8	{rs3212578_A_D} => {rs7537288_G_D}	0.679	0.826	1.009	1.940
9	{rs8041354_T_D} => {rs17057233_C_D}	0.670	0.846	1.009	1.689
10	{rs9890077_C_D} => {rs17057233_C_D}	0.645	0.846	1.009	1.392

Table 5-13: Top 10 rules identified in controls for the ECM receptor interaction pathway

The network visualisation plots for the most significant rules identified in cases and controls for the ECM receptor interaction pathway are depicted in Figure 5-24 and Figure 5-25 respectively. These plots include the total number of significant rules identified. As the number of rules depicted increases, the network plot become less readable.

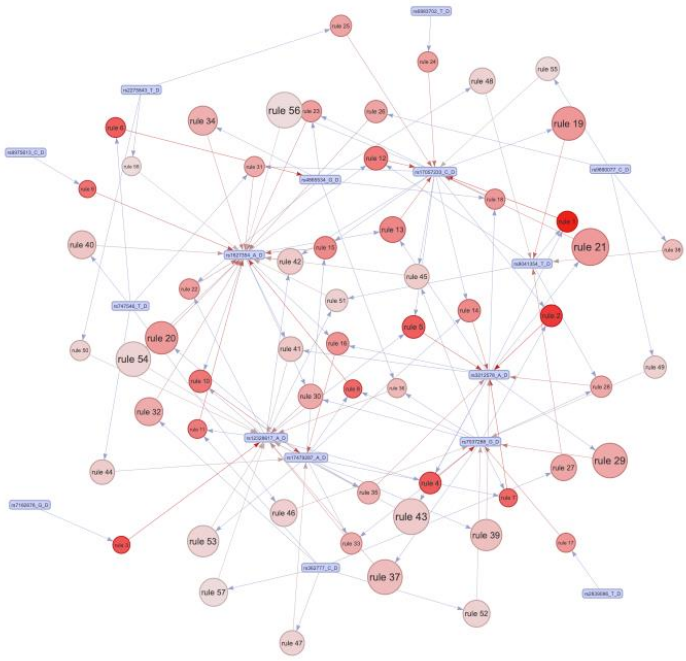


Figure 5-24: Rule visualisation network for the ECM receptor interaction pathway in cases

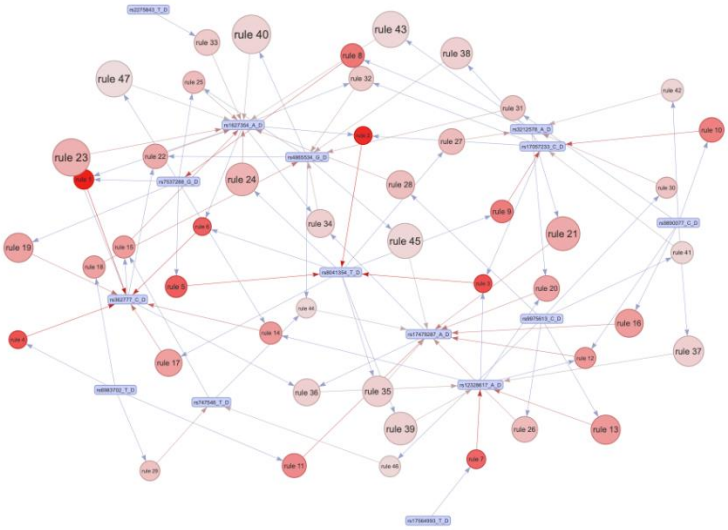


Figure 5-25: Rule visualisation network for the ECM receptor interaction pathway in controls

5.1.2.2.3 ARM for Peptide GPCRS Pathway

The most significant rules identified in cases for the peptide GPCRS pathway are listed in Table 5-14. For this pathway, 12 rules were identified with lift values slightly higher than 1, although only rules 1, 2, 3 and 4 had χ^2 values higher than 3.84. These four rules are therefore dependent.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs8080832_C_D} => {rs6469232_A_D}	0.612	0.821	1.018	3.597
2	{rs12883673_C_D,rs3136667_C_D} => {rs11628551_T_D}	0.625	0.935	1.017	6.121
3	{rs12883673_C_D} => {rs11628551_T_D}	0.721	0.932	1.014	6.970
4	{rs8080832_C_D} => {rs11628551_T_D}	0.693	0.930	1.012	3.849
5	{rs3798315_A_D} => {rs11628551_T_D}	0.697	0.929	1.010	3.262
6	{rs11628551_T_D,rs3136667_C_D} => {rs6469232_A_D}	0.644	0.814	1.010	1.292
7	{rs3798315_A_D} => {rs6469232_A_D}	0.610	0.812	1.007	0.518
8	{rs6469232_A_D} => {rs3136667_C_D}	0.697	0.865	1.005	0.625
9	{rs12883673_C_D} => {rs3136667_C_D}	0.668	0.863	1.004	0.250
10	{rs6469232_A_D} => {rs11628551_T_D}	0.744	0.922	1.004	0.505
11	{rs3798315_A_D} => {rs3136667_C_D}	0.647	0.862	1.002	0.093
12	{rs3136667_C_D} => {rs11628551_T_D}	0.791	0.919	1.000	0.001

Table 5-14: Rules identified in cases for the peptide GPCRS pathway

Conversely, only one rule was identified as correlated (rule 1) in controls as can be observed in Table 5-15. Additionally, a small number of significant rules were identified in controls for this pathway.

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs6469232_A_D} => {rs12883673_C_D}	0.644	0.841	1.017	4.836
2	{rs6469232_A_D} => {rs11628551_T_D}	0.682	0.890	1.004	0.519
3	{rs3136667_C_D} => {rs11628551_T_D}	0.734	0.887	1.001	0.057
4	{rs8080832_C_D} => {rs11628551_T_D}	0.619	0.887	1.001	0.015

Table 5-15: Rules identified in controls for the peptide GPCRS pathway

The network visualisation plots for the most significant rules identified in cases and controls for the peptide GPCRS pathway are depicted in Figure 5-26 and Figure 5-27 respectively.

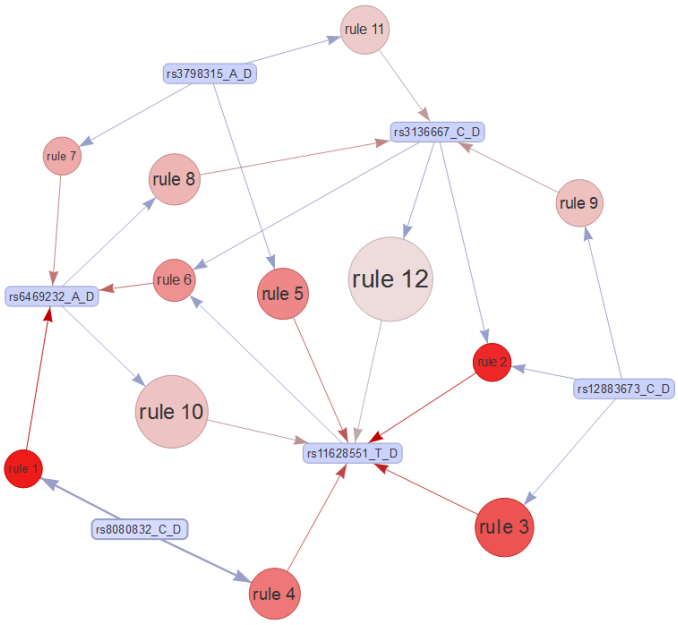


Figure 5-26: Rule visualisation network for the peptide GPCRS pathway in cases

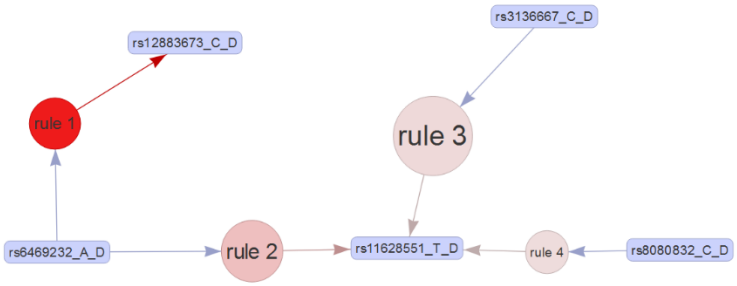


Figure 5-27: Rule visualisation network for the peptide GPCRS pathway in controls

5.1.2.2.4 ARM for Prostate Cancer Pathway

Only the top 10 most significant rules were reported for the prostate cancer pathway as the number of rules generated was large (see Table 5-9). The full list of rules identified for cases and controls is provided in Appendix G. No rules were identified as dependent in cases and controls as observed in Table 5-16 and Table 5-17 respectively.

Rank	**Rule**	**Supp.**	**Conf.**	**Lift**	χ^2
1	{rs10510097_A_D} => {rs42695_A_D}	0.631	0.838	1.014	2.391
2	{rs11672342_T_D} => {rs16897333_G_D}	0.626	0.816	1.012	1.646
3	{rs2849379_T_D} => {rs11466212_G_D}	0.662	0.862	1.009	1.438
4	{rs16897333_G_D,rs4739561_T_D} => {rs11190421_A_D}	0.610	0.854	1.008	0.850
5	{rs6857523_A_D} => {rs4739561_T_D}	0.638	0.898	1.008	1.033
6	{rs11466212_G_D,rs4739561_T_D} => {rs17041230_A_D}	0.667	0.883	1.006	0.734
7	{rs42695_A_D,rs4739561_T_D} => {rs11190421_A_D}	0.626	0.851	1.006	0.464
8	{rs2849379_T_D} => {rs13116385_T_D}	0.618	0.804	1.006	0.396
9	{rs17041230_A_D,rs4739561_T_D} => {rs13116385_T_D}	0.628	0.804	1.005	0.272
10	{rs17041230_A_D,rs2849379_T_D} => {rs4739561_T_D}	0.601	0.895	1.005	0.315

Table 5-16: Top 10 rules identified in cases for the prostate cancer pathway

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs587230_A_D} => {rs17041230_A_D}	0.605	0.847	1.014	2.775
2	{rs17041230_A_D,rs42695_A_D} => {rs13116385_T_D}	0.621	0.849	1.014	2.879
3	{rs13116385_T_D,rs4739561_T_D} => {rs17041230_A_D}	0.603	0.846	1.012	2.122
4	{rs11190421_A_D,rs11466212_G_D} => {rs17041230_A_D}	0.602	0.846	1.012	2.033
5	{rs10510097_A_D,rs13116385_T_D} => {rs42695_A_D}	0.606	0.886	1.012	2.422
6	{rs11672342_T_D} => {rs2849379_T_D}	0.666	0.819	1.011	2.554
7	{rs11190421_A_D,rs17041230_A_D} => {rs11672342_T_D}	0.609	0.822	1.010	1.493
8	{rs17041230_A_D} => {rs13116385_T_D}	0.707	0.846	1.009	2.493
9	{rs13116385_T_D,rs4739561_T_D} => {rs42695_A_D}	0.630	0.883	1.009	1.489
10	{rs16897333_G_D} => {rs11466212_G_D}	0.617	0.816	1.009	1.106

Table 5-17: Top 10 rules identified in controls for the prostate cancer pathway

The network visualisation plots for the most significant rules identified in cases and controls for the prostate cancer pathway are depicted in Figure 5-28 and Figure 5-29 respectively.

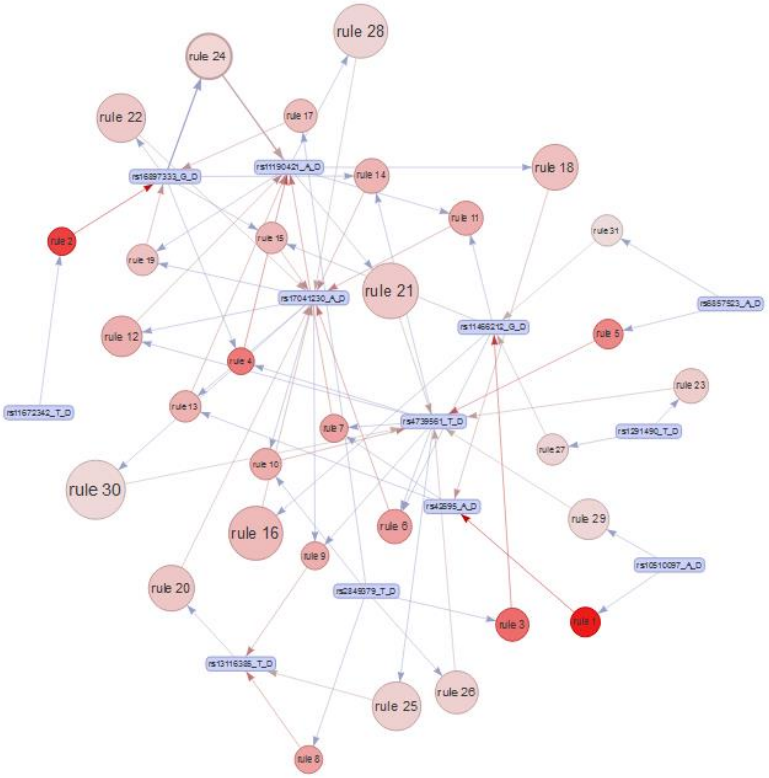


Figure 5-28: Rule visualisation network for the prostate cancer pathway in cases

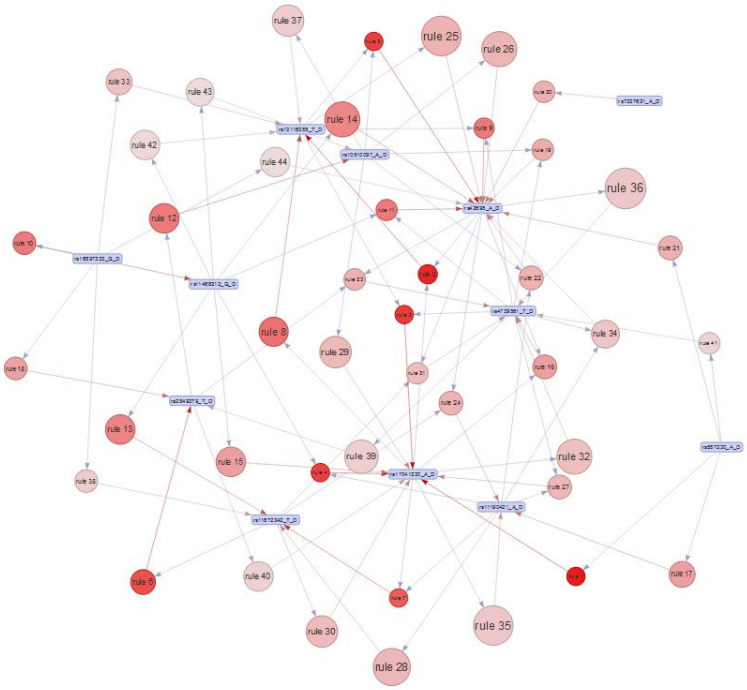


Figure 5-29: Rule visualisation network for the prostate cancer pathway in controls

5.1.2.2.5 ARM for the Union of All Canonical Pathways

In this section, the rules identified for the union of the four most significant canonical pathways are reported. A total of 191 and 112 correlated rules were identified for cases and controls respectively (see Appendix G). Since the number of generated rules is large, only the top 10 rules are listed in Table 5-18 (cases) and Table 5-19 (controls).

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs10510097_A_D,rs17041230_A_D} => {rs17057233_C_D}	0.601	0.913	1.032	13.385
2	{rs3212578_A_D,rs4739561_T_D} => {rs6990501_G_D}	0.647	0.839	1.030	11.947
3	{rs10510097_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.605	0.911	1.029	11.444
4	{rs11628551_T_D,rs16897333_G_D} => {rs2722279_C_D}	0.645	0.872	1.029	11.832
5	{rs10510097_A_D,rs11628551_T_D} => {rs17057233_C_D}	0.626	0.911	1.029	12.343
6	{rs11628551_T_D,rs12328617_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.615	0.887	1.028	10.113
7	{rs17057233_C_D,rs2722279_C_D,rs7537288_G_D} => {rs11628551_T_D}	0.604	0.945	1.028	13.769
8	{rs362777_C_D,rs593241_T_D} => {rs17041230_A_D}	0.612	0.901	1.027	9.978
9	{rs11190421_A_D,rs11628551_T_D} => {rs747546_T_D}	0.629	0.812	1.027	8.342
10	{rs17041230_A_D,rs3212578_A_D,rs4739561_T_D} => {rs17057233_C_D}	0.611	0.909	1.027	9.822

Table 5-18: Top 10 rules identified in cases for the union of all canonical pathways

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs12450493_A_D,rs7537288_G_D} => {rs4739561_T_D}	0.634	0.875	1.024	9.692
2	{rs11628551_T_D,rs8041354_T_D} => {rs2569538_A_D}	0.602	0.863	1.023	7.317
3	{rs11190421_A_D,rs11628551_T_D,rs4739561_T_D} => {rs12450493_A_D}	0.602	0.905	1.023	8.676
4	{rs12450493_A_D,rs4739561_T_D} => {rs8041354_T_D}	0.615	0.809	1.023	6.825
5	{rs17479287_A_D,rs362777_C_D} => {rs10510097_A_D}	0.608	0.831	1.021	6.054
6	{rs6469232_A_D} => {rs593241_T_D}	0.635	0.829	1.021	7.145
7	{rs11190421_A_D,rs17041230_A_D} => {rs3798315_A_D}	0.606	0.817	1.021	5.788
8	{rs10510097_A_D,rs11628551_T_D} => {rs593241_T_D}	0.600	0.829	1.021	5.712
9	{rs4733616_T_D} => {rs2569538_A_D}	0.622	0.861	1.021	6.918
10	{rs12328617_A_D,rs7537288_G_D} => {rs2849379_T_D}	0.614	0.827	1.021	6.055

Table 5-19: Top 10 rules identified in controls for the union of all canonical pathways

The network visualisation plots for the most significant rules identified in cases and controls for the union of all canonical pathways are depicted in Figure 5-30 and Figure 5-31 respectively. For each visualisation, the *arulesViz* package only plots a maximum of 100 rules (top 100) as the network visualisation becomes unreadable.

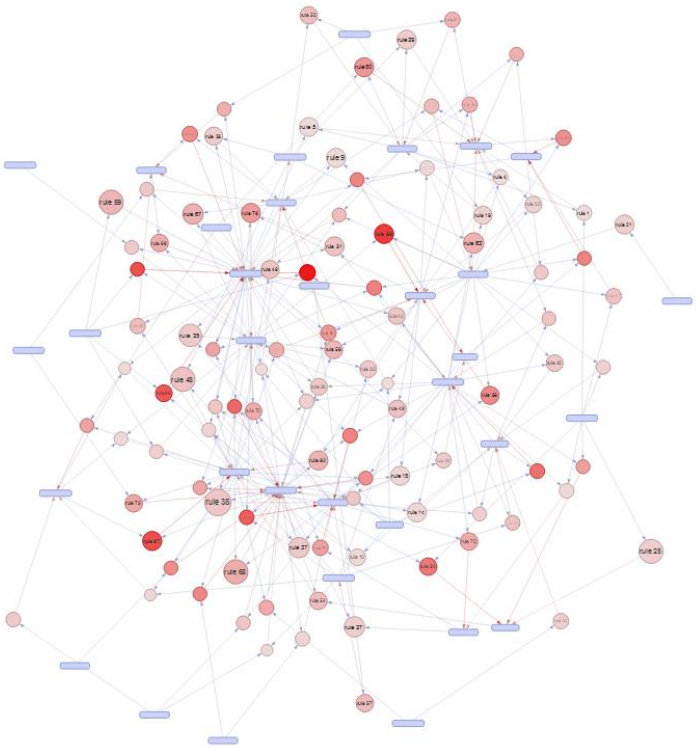


Figure 5-30: Rule visualisation network for the union of all canonical pathways

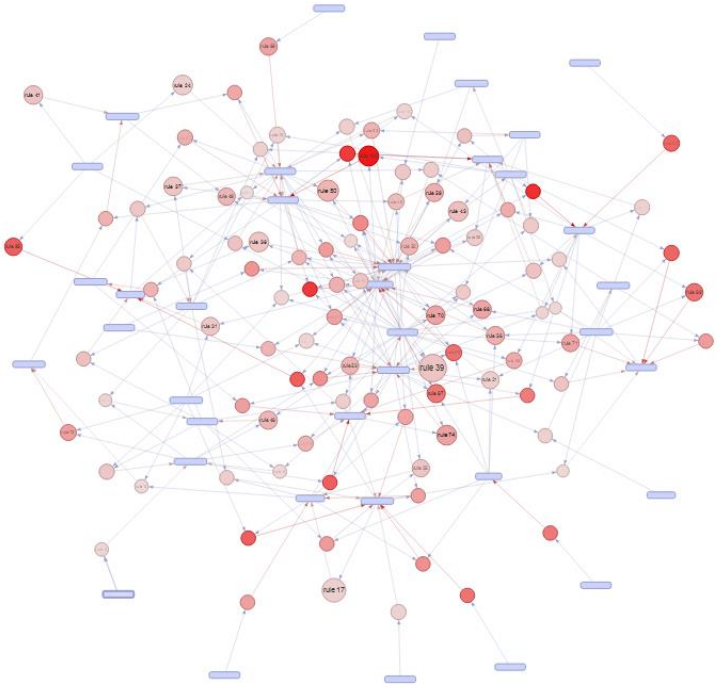


Figure 5-31: Rule visualisation network for the union of all canonical pathways

5.1.2.3 Classification analysis using pathway features

In this section, classification analyses are conducted using MLP only. Stacked autoencoders are no longer needed since our feature selection is based on biological pathways. The idea is to demonstrate whether features based on biological knowledge provide better classification results than a statistical filtering approach or not.

Based on methods described in Chapter 3, the performance of the different classifiers is reported using the SNPs within the most significant rules in each pathway as similarly conducted in the experiments with SAERMA. The network architecture and the regularization parameters were tuned as described in previous experiments using random search, while early stopping was adopted to avoid overfitting. Once more, ADADELTA was used for stochastic gradient descent optimisation, with parameters *rho* and *epsilon* set to 0.99 and $1x10^{-8}$

respectively. For each classifier, the name of the pathway, the number of features (SNPs) used as features and the performance metrics are provided.

Pathway	SNPs	Set	SE	SP	Gini	LogLoss	AUC	MSE
WNT	13	V	0.94	0.18	0.18	0.68	0.59	0.24
		T	0.95	0.11	0.17	0.68	0.59	0.24
EMC	17	V	0.98	0.07	0.24	0.67	0.62	0.24
		T	1.00	0.02	0.18	0.68	0.59	0.24
GPCRS	6	V	0.91	0.19	0.22	0.67	0.61	0.24
		T	0.99	0.01	0.10	0.69	0.55	0.25
Prostate Cancer	14	V	1.00	0.035	0.15	0.68	0.58	0.25
		T	0.88	0.30	0.23	0.68	0.62	0.24
All Pathways	47	V	0.81	0.47	0.40	0.63	0.70	0.22
		T	0.92	0.28	0.30	0.65	0.65	0.23

Table 5-20: Performance for classification analysis using the different canonical pathways and the union

The best model performance was achieved by the combination of all pathways with 65% AUC in the test set. It can be noted that the sensitivities in all the classifiers were very low, indicating that models have difficulties to correctly recognise non-obese individuals. Conversely, sensitivities for all models are very high. In Figure 5-32, the combined ROC curves for the test set using trained models with the SNPs of the most significant rules from each pathway is depicted. In the figure, the orange ROC curve (all canonical pathways) represents the best performance of all classifiers.

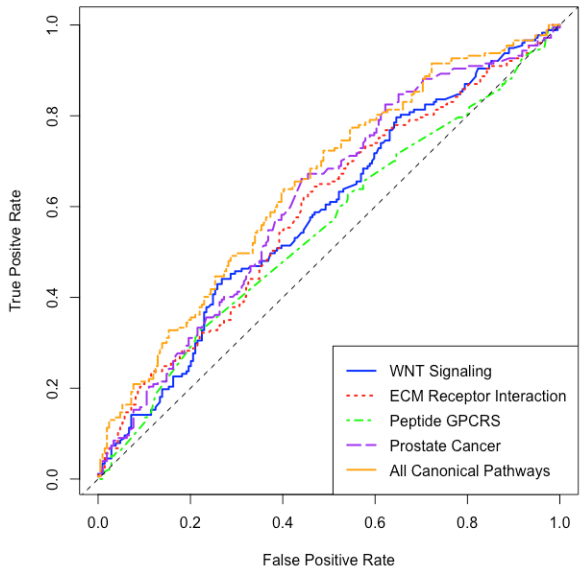


Figure 5-32: Combined ROC curves for the test set using trained models with the SNPs of the most significant rules

5.2 Chapter summary

In this chapter, association rules identified by SAERMA were biologically validated via pathway analysis using DAVID and functional data obtained using the KEGG and the Reactome pathway knowledgebase. The results confirmed the effectiveness of ARM for the discovery of epistasis in complex diseases using GWAS data as several rules were identified in pathways relevant to obesity. To demonstrate this, three main pathways were reported: metabolism pathway, metabolism of lipids and FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes pathway.

Additionally, a proof of concept experiment was performed (therefore future work needs to be conducted) to filter genetic variants based on GSEA instead of statistical filtering. This experiment was conducted using the *improved* gene-set enrichment analysis for GWAS (*i*-GSEA4GWAS) web tool. Results obtained are intended to be compared with those achieved by SAERMA in order

to validate the effectiveness of using different filtering methods to reduce genome-wide data dimensionality before conducting epistatic analysis in complex diseases.

Chapter 6. DISCUSSION

In this study, bioinformatics tools and techniques were combined with deep learning and frequent pattern mining, as a framework to capture nonlinear dependencies and epistatic interactions between SNPs. GWAS has proven to be useful for identifying common genetic variants. However, SNPs are independently tested for association with a phenotype of interest, ignoring the possible relationships that may exist between genetic variants. Single SNP scan is still the most extensively used approach. Hence, a multi-stage procedure has been considered in this research, in which a subset of SNPs following quality-control (QC) and association analysis was selected to explore the epistatic interactions between genetic variants using SAE and ARM. The information extracted was then used for classification analysis in various experiments and the rules validated using gene set enrichment analysis.

Following QC, 1,997 individuals (879 cases and 1,118 controls) and 240,950 genetic variants remained for subsequent analysis. This corresponds to 353,084 SNPs, 127 cases and 146 controls pruned from the original data set. Although it is normal for samples and SNPs to be removed after QC filtering procedures, an elevated number of genetic variants are removed due to missing genotype and MAF as indicated in Table 4-2. Missing call rate is an indicator of data completeness, but it is also an indicator of genotype quality (Laurie et al. 2010). Thus, using a dataset genotyped with more up to date technology would have produced data with higher call rates.

Genomic control was applied in association analysis to control for population stratification. Figure 4-1 indicates that population structure is appropriately

managed since there is no early deviation from $y = x$. This is supported by a genomic inflation factor value close to one (λ = 1.0384). A suggestive association between extreme obesity and the CDH13 protein-coding gene was observed in association analysis, with the most significantly suggestive marker being rs763727 (P-value = 1.821×10^{-6}, allele A). The second strongest suggestive association was achieved by the marker rs726553 (P-value = 7.330×10^{-6}, allele G), situated in an intergenic region of the genome. The third suggestive SNP listed was rs10817737 (P-value = 8.319×10^{-6}, allele A), situated in the TMOD1 protein-coding gene. The suggestive variant rs3050 (P-value = 9.061×10^{-6}, allele A) was located in the protein-coding gene PLEKHG1 and rs1278895 (P-value = 9.979×10^{-6}, allele T) in an intergenic region. Using SNPnexus the closest genes for the intergenic variants were obtained. Therefore, rs726553 is located 109,332 b upstream from the DOCK10 protein-coding gene, while rs1278895 is located 818 b upstream from RP11-159D23.2, which is a long intervening noncoding RNA (lincRNA) (Ulitsky & Bartel 2013). In Figure 4-3 these five SNPs are highlighted and labelled.

Three of the reported genes have previously been associated with obesity related traits. Particularly, gene CDH13 which is associated with Adiponectin levels (Chung et al. 2011), Hypertension (Org et al. 2009) and Coronary Artery Disease (CAD) (Nelson et al. 2017). PLEKHG1 has been reported to be associated with blood pressure in African-ancestry populations (Liang et al. 2017). While, DOCK10 is thought to be linked with the response of triglycerides (TG) during regular exercise, as reported in (Sarzynski et al. 2015) - elevated TG is strongly associated with increased risk of cardiovascular disease (CVD).

From a clinical perspective, these findings can be seen as relevant since they can be used as potential candidate variants for future work on obesity, with previous indications of being involved in obesity related traits. From a methodological point of view however, these SNPs have been studied in isolation, without taking into consideration how they interact with other SNPs/genes and cumulatively lead to obesity in humans.

6.1 Generalised Linear Model with P-values < 10^{-2}

The first experiment conducted following QC and association analysis tested the capacity of the filtered SNPs to discriminate between case and control samples using a GLM; the go-to method for binary classification. This experiment was used to establish a baseline performance for more advanced ML methods.

Results presented in Section 4.4 indicate that GLM can accurately identify case and control individuals using 2,465 features (SNPs) with an AUC of 94% (SE = 85%, SP = 90%, Gini = 87%, Logloss = 0.3288 and MSE = 0.0976) when using the test set, as shown in Table 4-8. These performance values were achieved with *alpha* and *lambda* regularisation values equal to 0.5 and 0.00151 respectively, which were introduced to avoid overfitting (see Table 4-6). Although AUC values remained high when 248 and 32 SNPs were used as input features (see Table 4-8), specificities started deteriorating when the number of SNPs was reduced; achieving the lowest value when only 5 SNPs were used. Hence, for GLM to be able to classify cases and controls in a balanced way, the model required an elevated number of SNPs (2,465 SNPs).

The major limitation however of using GLM is that it is not possible to fit models using nonlinear data as in the case of epistasis. Consequently, the use of non-linear machine learning algorithms was considered a viable alternative to study polygenic obesity in this thesis.

6.2 Classification using MLP and P-values < 10^{-2}

The second classification experiment modelled MLP NNs, which have been extensively used in bioinformatics. MLPs are non-parametric models that provide significant advantages over GLM, including its capacity to capture complex non-linear relationships between dependent and independent variables through hidden nodes.

Using an MLP classifier with rectifier as the activation function with dropout and genetic variants with P-value < $1x10^{-2}$ (2,465 SNPs) it was possible to obtain the results (SE = 0.9548, SP = 0.9761, Gini = 0.9878, LogLoss = 0.1061, AUC = 0.9938 and MSE = 0.0291). In contrast, using the 5 suggestive SNPs (P-value < $1x10^{-5}$) resulted in a significant performance drop (SE = 0.9943, SP = 0.0622, Gini = 0.2074, LogLoss = 0.6750, AUC = 0.6037 and MSE = 0.2410). The lowest specificity (SP = 0.0622) value achieved was also reported in the model with 5 SNPs (see Table 4-12), indicating that the model was unable to classify normal individuals correctly. In Figure 4-5 (e) and (f) signs of overfitting can be observed. In both cases, there is an early divergence between the training and validation curves. The effect of overfitting causes the model to perform well on the training data but not on the validation set. In other words, the model remembers the training samples but does not generalise well to new samples. In the classification experiments conducted with 32 SNPs (P-value <

$1x10^{-4}$), SP = 75% in the validation set, SP = 29% in the test set indicated in Table 4-11 and Table 4-12 respectively show that the model has been closely fitted to the training data; i.e. the model has been overfitted.

Acceptable results were obtained using MLPs with 2,465 and 248 SNPs as inputs, with high AUCs and relatively balanced SE and SP values as shown in Table 4-12. However, compared with the GLM classifier experiment, specificities started to deteriorate when the number of input features reached 32 or less. These results reveal that MLP achieve overall better results than GLM, probably due to the non-linear interactions that occur between SNPs.

The preference for MLP as a classifier is, thus, explained by the fact that nonlinearities are learnt which is not the case with logistic regression.

6.3 Epistatic interactions using Stacked Autoencoders

Following the previous experiments, an important limitation arises concerning the epistatic interactions between genetic variants. Despite the capacity of GLM and MLP models to classify case and control individuals using a maximum of 2,465 SNPs, GLM fails to capture the non-linear interactions present in SNP-to-SNP interactions. Whereas MLP can learn and capture epistatic information, yet a high number of features are required to achieve good performance. It is not clear to what extend those SNPs interact and what proportion of the data actually represents noise. Investigating this further, autoencoders were used here to determine if a low-dimensional representation of our input data (2,465 SNPs) could be achieved, while retaining all relevant information. This helps to remove any redundant features with a particular focus on epistasis.

In this experiment, a set of 2,465 SNPs (P-value < $1x10^{-2}$) and four single layer AEs were implemented to compress SNPs through 2,000-1,000-500-50 hidden units. AEs are stacked to enable greedy layer-wise learning following the network architecture defined in Figure 3-11. Each AE and associated compressed hidden layer force the network to only retain important information (features) in the data. The results produced by the SAE are utilised to pre-train the weights for the MLP classifier, rather than randomly initialising the weights to small values. This is an advantage of using greedy layer-wise pre-training, which helps the model initialise the parameters near to a good local minimum and transform the problem space to a better form of optimisation.

The best result using the test set was obtained using 2,000 hidden units (SE = 95%, SP = 93%, Gini = 95%, Logloss = 0.1956, AUC = 0.97497 and MSE = 0.054057) and a rectifier activation function with dropout. Conversely, the worst result was achieved when the features were compressed to 50 hidden units (SE = 78%, SP = 80%, Gini = 70%, Logloss = 0.476864, AUC = 85% and MSE = 0.156315), which are still encouraging.

Figure 4-7 shows that there is no significant overfitting between the training and validation datasets, except in the subfigures (c-d) and (e-f) where the validation logloss tends to increase after 2 epochs, thus deviating from the training logloss. While the validation AUCs for 1,000 and 500 compressed units plummet with respect to their training AUCs. On the other hand, Figure 4-8 from *(a)* to *(d)* shows a gradual deterioration in performance when the features are compressed into a smaller number of hidden units. However, the performance is still high even with 50 units, over 85% AUC with SE = 78% and

SP = 80% with no evidence of overfitting as shown in Figure 4-7. This supports our previous argument and shows that there is significant noise within the initial 2,465 SNPs. This, thus, demonstrates the potential of the proposed deep learning methodology to abstract large, complex and unstructured data into latent representations capable of capturing the epistatic effect between SNPs in GWAS.

In comparison with the experiments presented in Sections 4.4 and 4.5 , where SNPs were filtered using suggestive SNPs and Bonferroni levels of significance, the results are much better and highly encouraging using SAEs. Sensitivities and specificities are generally more balanced – for example, compare the results in Table 4-12, for 32 SNPs (P-value < $1x10^{-4}$), where SE = 95% and SP = 29% with those in Table 4-16, for 2000-1000-500-50, where SE = 79% and SP = 80%. In addition to SE, SP and AUC values, SAEs also improved Gini, Logloss and MSE values when compared with models using a similar number of input features. Furthermore, the results obtained using SAEs with 50 hidden nodes are close to those achieved with 248 SNPs using GLM and MLP. A summary of these results is shown in Table 6-1.

Model	Features	Set	SE	SP	Gini	LogLoss	AUC	MSE
GLM	248 SNPs	Test	0.9548	0.6315	0.7798	0.4119	0.8899	0.1350
MLP	248 SNPs	Test	0.9039	0.7942	0.8512	0.3475	0.9256	0.1094
SAE	50 units	Test	0.7853	0.7990	0.7036	0.4769	0.8518	0.1563

Table 6-1: Result comparison for GLM, MLP and SAE using 248, 248 and 50 features respectively

Utilising deep learning SAE provides a more effective approach than using direct features from statistical approaches such as logistic regression in GWAS

for classification tasks, since it improves overall model performance while reducing the dimensional space and overfitting.

Although deep learning is used in Eraslan et al., (2016), this approach differs from the approach presented in this section, in that, QC and GWAS are conducted using all the SNPs genotyped in the MyCode study dataset. Pre-SNP selection, based on functional regulatory effects from a public repository, is not applied since the aim of this thesis is to find epistatic interactions between SNPs. While DeepWAS concentrates more on biological outcomes (including regulatory mechanisms in GWAS), this thesis focuses on testing new algorithms for epistasis and classification analysis.

The experiment reported in this section represents a novel approach with emphasis on the feature extraction and classification phases, using latent information extracted from high-dimensional genomic data for the identification of individuals with higher predisposition to obesity. However, compressing the features using SAEs makes it difficult to identify which of the 2,465 SNPs contributes to the compressed hidden units. This is a well-known problem in neural networks where model interpretation is difficult to achieve (Manning et al. 2014). Consequently, there is a need to support the interpretation of any model used to fully understand the genetic influence SNPs have in the manifestation of phenotypes.

In order to address this issue, the final experiment combines the strength of SAE for the identification of epistasis and ARM via the *Apriori* algorithm, to provide an interpretation of the deep learning networks utilised in this research.

6.4 SAERMA

An advantage of ARM is that the underlying algorithms they use provide a more interpretable way to identify features and the interactions between them through the rules that ARM identifies. ARM is more transparent than other machine learning algorithms as it provides knowledge based explanative rules, serving therefore as a white-box model. Hence, this approach allows us to investigate relevant epistatic patterns and to determine the direction of associations between SNPs, while the use of SAE and MLP classification provides an objective performance measure for the models ARM produces. These are tightly correlated in that altering the interest measures (support and confidence) in ARM impacts on the performance metrics of the SAE and MLP models as discussed later.

Generating association rules using frequent itemset results in a large number of rules, many of which are redundant. Therefore, in the rule generation process, redundant rules were removed to alleviate the high number of rules being generated in the rule mining process and aid with computational efficiency. Although lift values for all the most significant rules in cases and controls were slightly higher than 1, the dependency of the rules was supported by very high values of χ^2. Furthermore, the frequency of appearance of the rules (support) in both cases and controls was ~60%. This means that even though exclusive patterns were identified in cases, they were not common in all individuals classified as extremely obese (only in ~60% of the samples). The inference made by an association rule does not necessarily imply causality. Counterwise, it suggests a strong relationship between SNPs in the antecedent and consequent

of the rule. Causality, in contrast, requires certain knowledge about the cause and effect of attributes in the data and typically involves relationships occurring over time. Hence, ARM results need to be carefully interpreted.

One of the possible reasons why obesity related variants within the genes FTO or MC4R were not identified in any stages of the proposed methodology, may be due to the effect of removing a very large number of variants by using stringent thresholds in the per-marker QC step. It is known that statistical power to detect an SNP of a given effect via GWAS increases with both sample size and the density of genetic variants across the genome (Spencer et al. 2009) and, in this study, the sample size is relatively small (1,997 individuals after QC) and the density of markers was also reduced considerably (240,950 SNPs after QC). While no variants were identified within these well-known genes, other genes with previously reported implications with obesity were identified, including CDH13, PLEKHG1 and DOCK10. These variants were identified utilising standard GWAS procedures with logistic regression, so, while the SNPs are considered individually suggestive of association, it is not clear that they intervene in epistatic interactions. However, epistasis was explored using the proposed method where several rules were identified as significant and their genes reported and tested for classification analysis.

Although reporting SNP to genes mapping might provide information about whether they have been previously reported to be associated with obesity or not, this does not validate the association rules from a biological perspective. That is, it is not possible to know if the rules truly represent epistasis or just random relationship between SNPs mined by the *Apriori* algorithm. Therefore, using

functional and computational follow-up analyses, the top 300 most significant association rules identified by SAERMA in cases and controls were validated. We gained insights into rules-implicated genes and the biological pathways through which they influence obesity. The results revealed a relationship between rules identified by SAERMA and biological pathways with roles in metabolism (including metabolism of lipids) and neurobiology in body weight regulation (FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes) which validates the potential of ARM in epistatic analysis. Particularly, the following rules (see Table 6-2) were identified for the main three pathways highlighted, although more were identified as reported in Table 5-3.

Path	Cases	Controls
Metabolism	{**ACAD10**}=>{**ALDH2**}	{ATXN2, **ALDH2**, **ACAD10**}=>{ATXN2}
	{**ALDH2,** RP3-462E2.3}=>{**ACAD10**}	{**ENTPD4**, ATXN2}=>{ATXN2}
	{**ALDH2,** RP3-462E2.3}=>{**ALDH2**}	{MAPKAPK5, ATXN2, **ACAD10**, **SLC10A2**}=>{ATXN2}
	{MAPKAPK5**, ACAD10**, HECTD4}=>{**ALDH2**}	{**AOX1**, DOCK4}=>{SGOL2}
	{NAA25**, ACAD10**}=>{**ACAD10**}	{ATXN2**, ACAD10**, HECTD4}=>{ATXN2}
	{SGOL2}=>{**AOX1**}	{**MTMR7**}=>{**MTMR7**}
		{**HK1,** GRIK1}=>{GRIK1}
Lipids	{**ACAD10**, RP3-462E2.3}=>{**ACAD10**}	{MAPKAP5, ATXN2, **ACAD10**, **SLC10A2**}=>{ATXN2}
		{**MTMR7**}=>{**MTMR7**}
FOXO	{**HDCA2**, ACAD10}=>{BRAP}	{ATXN2, **NPY**}=>{ATXN2}

Table 6-2: Identified rules within relevant obesity pathways

Several relevant genes were identified in the three pathways highlighted (two super pathways and a contained pathway), some of them with implications in the synthesis and breakdown of fatty acids (key in energy metabolism) and the control of appetite and food intake. As an example, NPY stimulates food intake

and weight gain, ACAD10 contribute to the beta-oxidation of fatty acids in the mitochondria, SLC10A2 plays a key role in cholesterol metabolism or HK1, which catalyses the conversion of glucose into glucose-6-phosphate, the first essential step of glucose metabolism. Furthermore, those genes within the identified rules mapped to the pathways (highlighted in Table 6-2), also interact with genes that have been previously linked with obesity. For example, the ATXN2 gene has been involved in severe early onset obesity in children (Figueroa et al. 2009). Moreover, loss of function mutations in this gene may be associated with disease susceptibility (type I diabetes, obesity and hypertension) as indicated by GWAS. The MAPKAPK5 gene, which has shown gender-dependent differences in anxiety-related processes and locomotor activity in mice (Gerits et al. 2007). A weak but positive association between anxiety and obesity in humans has been reported, although further studies were recommended by Gariepy et al. (2010). Furthermore, the GRIK1 gene has been reported as a novel obesity candidate gene that may contribute to highly penetrant forms of familial obesity (Serra-Juhé et al. 2017).

These identified genes represent, thus, potential targets which could lead to novel and more precise approaches for the treatment of obesity.

6.4.1 Classification analysis

After rule mining was applied to the filtered SNPs (2,465 SNPs), several classifiers were pre-trained with the compressed units extracted from the top 300, 200, 100 and 50 rules. For each set of rules, their SNPs (forming the rules) were used as input features for several SAEs. Then, the MLP classifiers were

initialised and fine-tuned with the final hidden layer of the SAE. The results are presented in Table 4-20 to Table 4-23.

Top 300 rules

In the first set of 300 rules with 204 SNPs, the best result in the test set was achieved when the input features were compressed to 100 units, with an AUC of 77%, SE = 77%, SP = 68%, Gini = 53%, Logloss = 0.5769 and MSE = 0.1968, as shown in Table 4-20. Although a higher AUC was achieved with a single AE and 150 hidden units (AUC = 78%, SE = 80%, SP = 63%, Gini = 56%, Logloss = 0.5770 and MSE = 0.1952), the specificity value was inferior. In these situations, it is up to the expert/clinician to decide whether it is more important to detect cases of obesity more accurately than normal individuals. However, in this thesis, the capacity of our proposed solution to detect cases and controls in a balanced manner has been prioritised. This means that results with a balanced sensitivity and specificity and high AUC were selected.

Top 200 rules

In the second experiment conducted using the top 200 rules (161 SNPs), and according to the above criteria, the best result in the test set was accomplished when 75 compressed units (using two AEs) were used as input for the MLP classifier (see Table 4-21). The classifier achieved an AUC = 69% with SE = 70% and SP = 67% (Gini = 38%, Logloss = 0.6553 and MSE = 0.2259). However, the MLP trained with 125 compressed units (using a single AE) achieved an AUC = 73% with SE = 74% and SP = 66% (Gini = 47%, Logloss = 0.6099 and MSE = 0.2104). While the sensitivity in this case is 0.01% lower

than using 75 hidden units, the overall classification performance achieved represents an improvement.

Top 100 rules

The best result observed in Table 4-22 (top 100 rules with 124 SNPs) was achieved by compressing 124 SNPs down to 90 units. Performance metrics for this model were 71% AUC with 69% sensitivity and 66% specificity (Gini = 42%, Logloss = 0.6231 and MSE = 0.2167).

Top 50 rules

Finally, the models trained with the lowest number of features (92 SNPs from the top 50 rules) achieved the best classification results using a 75-50 layer configuration (See Table 4-23). This model reached an AUC value of 73% with sensitivity and specificity values of 77% and 63% respectively (Gini = 45%, Logloss = 0.6178 and MSE = 0.2142).

In Figure 6-1 the AUC values for the different classifiers are depicted. The different colours in the plot correspond to the different AEs (compression layers) considered in the stack, where the first, second and third layers are represented in blue, orange and green respectively. These results demonstrate that the classifier is not randomly assigning labels to the samples (AUC > 50%), although it struggles to classify non-obese individuals.

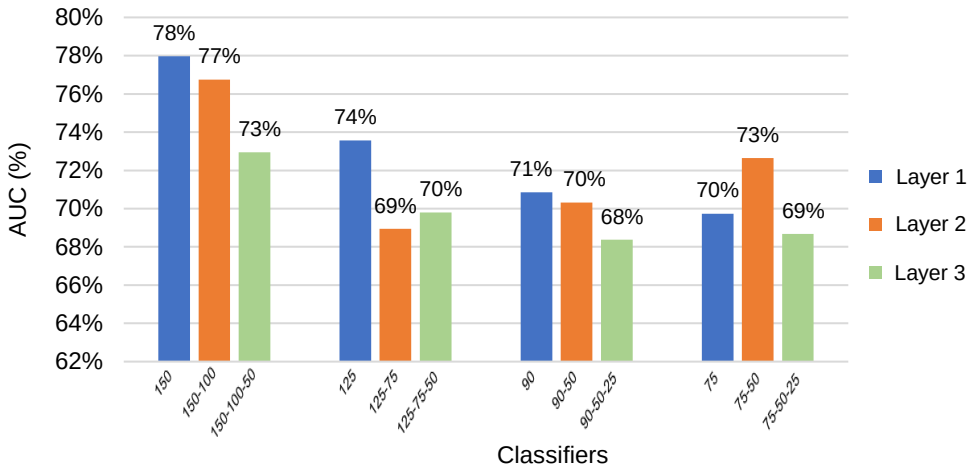


Figure 6-1: AUC values for the different classification analyses conducted for the top 300, 200, 100 and 50 rules

In all experiments conducted, classifiers' sensitivities were higher than specificities, showing that models were able to predict samples labelled as obese better than those labelled as non-obese. Generally, logloss values tend to increase when input features were compressed into a smaller number of units, with some exceptions. Even though the best results were achieved in the largest set of SNPs (300 rules), it can be observed that some of the models were able to compress the features down to 50 hidden units and get over 70% AUC.

A summary of the best results selected from each of above experiments is shown in Table 6-3. The results correspond to the total number of SNPs used as input and the compression layers utilised by the SAE.

Top rules	SNPs	Layers	SE	SP	Gini	LogLoss	AUC	MSE
300	204	150-100	0.7684	0.6794	0.5349	0.5769	0.7675	0.1968
200	161	125	0.7401	0.6651	0.4715	0.6099	0.7357	0.2104
100	124	90	0.6949	0.6603	0.4170	0.6231	0.7085	0.2167
50	92	75-50	0.7740	0.6268	0.4529	0.6178	0.7265	0.2142

Table 6-3: Best results selected from the different configurations with SAERMA using the test set

In Figure 6-2, the AUC, SE and SP values from Table 6-3 are depicted for an easy inspection. The dashed lines indicate that there is not much variation between the performance values (AUC, SE and SP) among classifiers despite the reduction in the number of SNPs and hidden units within the AEs.

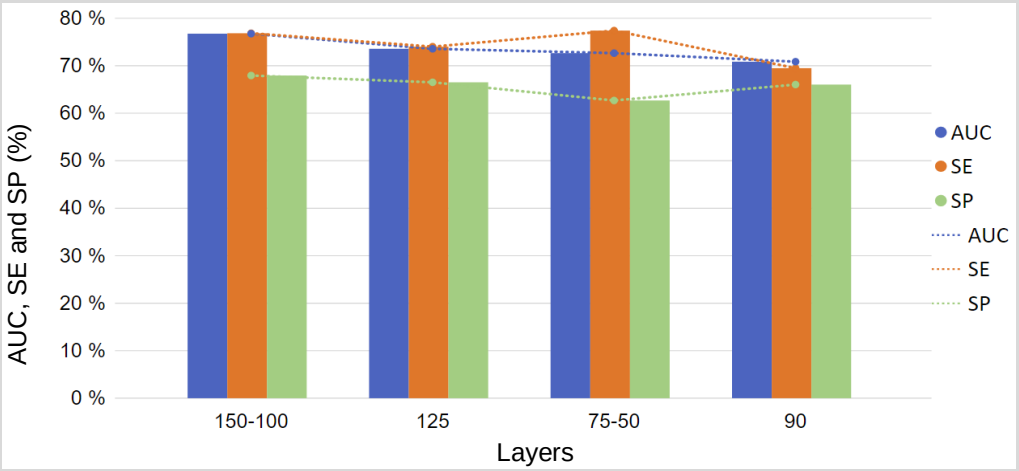


Figure 6-2: Best results AUC, SE and SP from SAERMA

Therefore, the best overall result from the different classifiers was AUC = 77%, attained by 100 compressed units from the top 300 rules as can be observed in Table 6-3. The classifier was able to classify obese individuals (SE = 77%) better than normal samples (SP = 68%). These results can be achieved with a maximum of 204 SNPs although the SAE is able to reduce noise and achieve that value (AUC = 77%) with 100 hidden neurons (this is a 50.99% reduction in the feature space). However, it is not possible to accurately determine which of those 204 SNPs correspond to the 100 compressed hidden neurons. For a more granular mapping of the interactions between SNPs, we can refer to the top 50 rules result (92 SNPs), where the input was compressed to 50 hidden units (see Table 6-3). Even though dimensionality reduction in this case affects the performance of the classifier with respect to the best result (using 204 SNPs), the SE value remains the same (77%), while SP is reduced

by 0.05 and AUC by 0.04. Thus, it is true to say that the 50 hidden nodes representing epistatic interactions can be interpreted using the 92 SNPs selected by ARM. Although this does not represent a full interpretation of the results obtained using SAEs, the approach presented in this thesis provides a close approximation of the epistatic interactions that likely occur in the MyCode data.

The best overall performance was achieved by the SAE using 204 SNPs. Hence, utilising SNPnexus it was possible to query the 204 SNPs and report the overlapped or closest genes according to the GRCh37 assembly. A table containing genomic annotations for the 204 SNPs reported in this thesis has been included in Appendix D. Additionally, genetic association data from complex diseases and disorders was also extracted for those SNPs using The Genetic Association Database via SNPnexus. A full list with phenotype and disease association is presented in Appendix E. To limit the scope of the disease associations from the GAD, a filter criterion was considered, and only SNPs under a metabolic and cardiovascular disease class were reported. It is expected that these findings will help future researchers to better understand how epistasis in obesity occurs using genome-wide data, providing candidate SNPs to investigate obesity further.

The effect of choosing the support and confidence threshold values in the rule generation stage of the methodology, determines the number of rules generated by the *Apriori* algorithm. In this thesis, the values 0.6 and 0.8 were selected for support and confidence respectively, which are the lowest thresholds to allow computationally feasible ARM experiments to be conducted. This allowed us to generate rules that represent interactions between

frequently occurring SNPs. To test the classification performance of these SNPs, several classifiers were tested and evaluated. In this approach we know that the SNPs are important due to ARM and these can be used in the SAE. However, there is no way to extract the epistatic interactions (rules) and use these in the SAE. Nonetheless, the assumption is that if ARM has identified the most significant SNPs then by using these SNPs in the SAE, the SAE should be able to find the epistatic interactions by pushing the SNPs through the autoencoder layers. While we cannot fully argue that what the SAE produces fully maps onto the rules generated by ARM, we can infer that there is some similarity between the two. It is on this basis that we are arguing in this thesis that we provide loose interpretation of the deep SAE layers within the neural network architecture.

The results are perhaps not as good as previous experiments. However, all previous experimental approaches in this thesis apart from SAERMA cannot identify important SNPs related to case and control and show the interactions that exist between them. SAERMA can achieve this. ARM selects the most important SNPs and their associations, and the effectiveness of the models that ARM generates can be objectively measured using the SAE. The results show that SE and SP are approximately 70% each and this is not chance. The model produced by ARM and the epistatic interactions extracted using SAE can distinguish between case and controls and get this correct 70% of the time. This is encouraging and grounds for more in-depth research in this area. It is on these grounds that we feel this work is highly novel. The approach is foundational and to the best of our knowledge has never been presented before in the academic literature.

The *Apriori* algorithm has been considered in the context of GWAS as AprioriGWAS, to study epistasis in age-related macular degeneration and bipolar disorder (Zhang et al. 2014). The authors used FIM to identify patterns in cases and controls by adjusting several interest measures (i.e. minimum support). However, the approach proposed in this final experiment differs from that one presented by Zhang et al. (2014) in that, logistic regression was used as a statistical filtering to reduce the SNP dimension prior to conducting machine learning experiments. Furthermore, SAE was used in combination with ARM, where not only frequent itemsets were identified using the *Apriori* algorithm, but rules generated from them. This allowed for a visual inspection of the most significant rules generated and ranked based on several interest measures such as support, confidence and lift. In addition to SAE and ARM, the top rules served as input features for binary classification analysis, which is also another fundamental difference from the work conducted in Zhang et al., (2014). While AprioriGWAS is applied to AMD and BD, the study presented in this thesis concentrates on obesity.

6.4.1.1 Randomly selected samples for classification

Since the best results achieved by SAERMA were obtained using 204 SNPs mined by ARM (top 300 rules), several additional classification experiments were conducted with subsets of 204 features (SNPs) randomly selected from the total 2,465 SNPs filtered from association analysis. These analysis were performed to demonstrate the convenience of using ARM with the proposed method. Therefore, this does not represent part of the proposed method but it is discussed here for validation purposes.

To validate this experiment, three random samples sets of size 204 were subsetted from the filtered 2,465 SNPs and used as input features in different MLP classifiers. As similarly conducted in previous classification experiments, the network architecture and the regularization parameters were tuned. To achieve this, random search was used, while early stopping was adopted to avoid overfitting. Again, adaptive learning rate ADADELTA was used for stochastic gradient descent optimisation, with parameters *rho* and *epsilon* set to 0.99 and $1x10^{-8}$ respectively, to balance the global and local search efficiencies. Results for random samples 1, 2 and 3 are reported in Table 6-4, Table 6-5 and Table 6-6 respectively.

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
150	Validation	0.8032	0.5683	0.5100	0.5869	0.7549	0.2006
	Test	0.8531	0.5598	0.5513	0.5640	0.7756	0.1929
150-100	Validation	0.8138	0.4053	0.3235	0.6469	0.6617	0.2277
	Test	0.8701	0.3684	0.2462	0.6844	0.623	0.2447
150-100-50	Validation	0.8723	0.2643	0.2357	0.6692	0.6178	0.2382
	Test	0.9435	0.1579	0.1673	0.6868	0.5836	0.2467

Table 6-4: MLP classifier performance for random sample set 1

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
150	Validation	0.8511	0.4846	0.5060	0.5881	0.753	0.2005
	Test	0.8136	0.5598	0.4655	0.6064	0.7327	0.2088
150-100	Validation	0.9415	0.1322	0.2442	0.6671	0.6221	0.2372
	Test	0.9831	0.1148	0.3056	0.6506	0.6528	0.2296
150-100-50	Validation	1.0000	0.0088	0.1299	0.6824	0.5649	0.2446
	Test	1.0000	0.0048	0.1135	0.6854	0.5567	0.2461

Table 6-5: MLP classifier performance for random sample set 2

Layer	Set	SE	SP	Gini	LogLoss	AUC	MSE
150	Validation	0.8298	0.4802	0.4620	0.6076	0.731	0.2093
	Test	0.8531	0.4880	0.4958	0.5939	0.7479	0.2038
150-100	Validation	0.8989	0.2819	0.3004	0.6504	0.6501	0.2298
	Test	0.9887	0.0766	0.1853	0.6808	0.5926	0.2439
150-100-50	Validation	0.9628	0.1454	0.2589	0.6633	0.6294	0.2354
	Test	0.9266	0.1770	0.1272	0.6910	0.5636	0.2487

Table 6-6: MLP classifier performance for random sample set 3

Empirical analysis conducted using the random subsets of SNPs revealed lower performances than those achieved by the features (204 SNPs) mined by ARM, reported in Table 4-20. More specifically, the performance of the classifiers worsened when the number of input features was gradually compressed into smaller hidden units in the SAE. Although this behaviour is somehow expected (as observed in previous experiments with SAEs), using random SNPs as inputs produced lower AUC values and very low specificities. Hence, using ARM to preselect SNPs before using SAEs improve performance metrics and allows to select overall better classification models with better AUCs and balanced SE and SP, as well as better Gini, LogLoss and MSE values.

6.4.2 Proof of concept experiment using *i*-GSEA4GWAS

In addition to the biological validation of the rules, a proof of concept experiment was conducted using GSEA tools and classification analysis to test whether using a biological filtering strategy outperformed the statistical filtering approach used in SAERMA.

Results revealed enrichment of GWAS results with four canonical pathways (Wnt signalling, ECM Receptor, Peptide GPCRS and Prostate cancer pathways) and numerous GO terms listed in Appendix G. The best classification

performance using MLP was achieved when features from all canonical pathways were used. Therefore, with 47 genetic variants (features) it was possible to achieve 65% AUC with 92% SE and 28% SP. However, these results are not better than those achieved by SAERMA as can be observed in Table 6-3. Although the best results achieved by SAERMA were obtained with 100 compressed units (204 input SNPs), it was possible to achieve 73% AUC, 77% SE and 63% SP with 50 units (92 input SNPs). This clearly demonstrates that using a statistical filtering approach within the proposed method SAERMA allows to achieve overall good results while improving specificity, without previous biological knowledge.

Even though better results are obtained using SAERMA, results achieved using the biological filtering via GSEA are still significant and need to be studied further. Several genes considered less significant in association analysis (those with large P-values) were enriched by the *i*-GSEA4GWAS tool and revealed pathways with implications in obesity. For example, an important gene involved in the peptide GPCRS pathway, the MC4R gene, has been highlighted as a key obesity-related gene in many studies.

6.5 SAERMA Limitations

While the results reported by SAERMA are encouraging, a number of limitations remain. Although we did provide a close interpretation of the features extracted by the SAE, this is not a full interpretable model. Therefore, the solution provided in this research is not final and it is open to further investigation.

Furthermore, ARM is computationally expensive when dealing with genome-wide data. Despite several filtering strategies, the minimum value for support and confidence in the rule generation process was empirically limited to 0.6 and 0.8 respectively. This constraint affected the number of rules produced and thus, influenced the classification performance of the SAERMA algorithm. This highlighted the difficulty in conducting epistatic analysis using high order SNP-SNP interactions.

Last but not least, data quality has played an important role in the genetic variants identified. In the QC phase, a high number of SNPs were removed. This SNP pruning limited the number of variants being explored in the proposed methodology and this also hindered novel discoveries.

6.6 Chapter Summary

The aim of this chapter is to describe the results derived from the framework adopted in this thesis for the analysis and interpretation of epistasis in obesity, using a genome-wide dataset. The choice of the final solution, SAERMA, has been driven by the limitations of GLM, MLP and SAEs when they are used independently since non-linearity is neglected, a high number of features are required, and interpretability of the results is an issue. Combined, ARM, SAEs and MLP as well as QC and association analysis provide a more complete solution to study epistasis in obesity as a complex disease.

The current study has important implications for interventions focused on identifying SNP interactions in complex disorders such as obesity. By combining omics with bioinformatics and functional studies, novel diagnostic

markers and therapeutic targets were identified following the emerging principles of systems medicine. It was found that none of the most statistically significance SNPs after association analysis with logistic regression (see Table 4-4) were identified as part of the top rules generated by the *Apriori* algorithm. This indicates that GWAS does not consider the collaborative effect between SNPs identified, whereas ARM assumes no hierarchy of SNP risk and creates simple association rules between two or more SNPs.

This thesis presents a novel approach with emphasis on the feature extraction and classification phases, using latent information extracted from high-dimensional genomic data for the identification of individuals with a high predisposition to obesity. However, while SNP-phenotype associations can be obtained using logistic regression analysis, SAEs maintain performance and reduce overfitting while minimising the dimensional space. ARM offers more interpretable risk patterns in the form of rules with support, confidence, lift and χ^2 as measures, while SNPs within the most significant rules also provide better classification performance than using random samples subsetted from the filtered 2,465 SNPs. Most importantly, this approach allowed us to investigate relevant epistatic patterns, validated via functional analysis, and to determine the direction of associations between SNPs. It also enabled us to identify candidate hidden risk SNP interactions which to the best of the author's knowledge have not been reported previously in the literature.

The results show the validity of SAERMA to detect a subset of attributes representing epistasis that are closely interpretable by the SNPs with the top rules identified by ARM. This is performed by comparing the results obtained

from the different classification experiments which provide an objective measure of the SNPs within the most relevant rules. Association rules were biologically validated and the results revealed a relationship between rules identified by SAERMA and biological pathways with roles in metabolism (including metabolism of lipids) and neurobiology in body weight regulation (FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes) which validates the potential of ARM in epistatic analysis. The extended plots generated from the rules provide the user with a visual tool to see which genetic variants take part in epistasis. The findings reported in this study are not limited to the genes identified in the top 10 rules in cases and controls. An extended list of genetic variants utilised to achieve the best results with the proposed algorithm SAERMA is presented in Appendix D.

Although the utilization of SAE, a multilayer feedforward ANN (MLP) and ARM have been previously considered separately in many areas of research, this thesis claims that this is the first time that they have been combined to study epistatic interactions between SNPs in GWAS of polygenic obesity and the results statistically and biologically validated using classification and functional analysis respectively.

Chapter 7. CONCLUSION AND FUTURE WORK

Obesity is complex with numerous compounding factors, some of which have been highlighted in this thesis. Research is providing a deeper understanding of the way risk factors interact with each other to help find potential solutions to the obesity epidemic, which are as multileveled and complex as its causes.

Continuous advances in sequencing technology have facilitated genetic analysis studies by sequencing the entire genome of individuals. This has resulted in a dramatic growth in the amount of data available for analysis where data mining and machine learning methods have become increasingly more important.

Advances in molecular technology have allowed us to conduct coarse genomic examinations, high resolution linkage analysis, and more recently GWAS. Methodological advances in GWAS contribute enormously to the amount of results generated and the number of experiments and samples collected. However, generating more data is not going to solve the current limitations found in association analysis. Development and application of innovative analytic approaches and study designs that allow the detection of epistasis are needed.

This thesis provides researchers with an approach for data mining in the application of case-control analysis for epistasis. The techniques applied were used to find associations between SNPs and the disease/phenotype, and subsequently to investigate epistatic interactions between associated SNPs. Although the results were not fully mapped to an SNP set in the validation

process, it is still possible to detect a reduced number of SNPs that are likely to play an important role in epistasis in the MyCode dataset. We argue that an interpretation (of which there are many) is provided for the proposed network architecture used to generate the classification results.

The methodology presented combines existing techniques for genetic analysis and machine learning models to identify SNPs and the interactions that exist between them, providing strong interpretation of the results, statistically validated through case-control classification tasks and biologically validated via functional annotation of gene sets. The proposed SAERMA algorithm has revealed combinations between SNPs that have not been previously considered and may be used as candidates or hypothesis generators in future studies. These candidate genetic variants need to be studied further by doctors and experts in complex diseases, to discover new therapies that may help to identify obesity susceptibility/predisposition from early stages of life. Translating computational models to etiological inferences represents one of the most difficult challenges in the study of complex disorders. In addition to epistasis discovery, identified SNPs also reported reasonably good classification performance when discriminating between extremely obese and normal samples.

Overall, the results in this thesis highlight the benefits of using deep learning stacked autoencoders to detect epistatic interactions between SNPs in genomic data and how these can be used to model MLPs to classify obese and non-obese observations from the eMERGE MyCode dataset. This contributes to the computational biology and bioinformatics field and provides new insights into

the use of deep learning algorithms when analysing GWAS that warrants further investigation. However, the minute non-linear transformations of the input space occur in the autoencoders, it is very difficult to trace the amount of variance they contribute from case-control data. This is a common problem in neural network modelling that seriously hinders genomic analysis. To aid with this issue, association rule mining was used in combination with stacked autoencoders. This allowed us to identify patterns in the form of rules which represent interactions between a filtered subset of SNPs. The benefits of incorporating rule mining to the proposed pipeline were twofold. First, it allowed us to generate significant rules and plot their interactions. Second, feeding the stacked autoencoders with the most significant rules allowed us to obtain dynamic classification performances by adjusting the number of rules generated in the rule mining process, serving thus as a validation and interpretation technique for the epistatic feature extraction in the neural network utilised in the study. Adjusting support and confidence coefficients to increase the number of rules also requires more computational complexity. Therefore, in this study only rules generated with support and confidence values of 0.6 and 0.8 respectively were presented. This allowed us to empirically produce the best results without reaching computational overload with the resources available. Conversely, using higher values of support or confidence would have resulted in information not being captured by the rules.

While work exists in biological analysis of variants that alter functional regulatory elements (i.e. elements that control gene expression and DNA) using deep learning methods (Eraslan et al. 2016) and epistasis analysis based on frequent itemset mining using the *Apriori* algorithm (Zhang et al. 2014), to the

best of our knowledge this thesis is the first comprehensive study of its kind that combines GWAS quality control and logistic regression with association rule mining and deep learning stacked autoencoders for epistatic-drive GWAS analysis and case-control classification.

7.1 Future Work

Several novel contributions have been provided using the proposed methodology. However, there are still areas for improvement. In this section, these are outlined.

Association analysis with logistic regression helped to reduce data dimensionality before machine learning experiments and this aided us with computational complexity. However, by assuming that epistasis only occurs between markers that independently have some effect on the phenotype, potential discoveries could be missed. Thus, in future work, statistical and biological filtering approaches should be replaced by more efficient techniques capable of dealing with the curse of dimensionality present in high dimensional genetic data. For instance, SAEs can be applied directly to the genotype data after QC without using filter approaches. We intend to investigate this using an advanced high-performance computing platform, such as a NVIDIA DGX-2[5] and their RAPIDS[6] development environment.

Only genetic features were used to evaluate an individual's risk to obesity in this study. In future work, non-genetic features such as physical activity and

[5] https://www.nvidia.com/en-gb/data-center/dgx-2/

[6] https://developer.nvidia.com/rapids

clinical data such as blood pressure may be considered to increase power in classification analysis.

To gain more insights into the possible role of identified rules in obesity, SNP-gene-pathway was analysed. By using pathway and functional enrichment analysis we were able to provide enough evidences to validate our results from a biological point of view. However, in future work the incorporation of eQTL analysis can be seen as a more robust approach to be used to functionally annotate GWAS results and prioritise the most significant SNPs, especially in the SNP selection stage prior to epistasis analysis with SAERMA. SNPs that are eQTLs in obesity-specific tissues can be used for the annotation based on expression data. Several tools for functional mapping and annotation of GWAS including FUMA (Watanabe et al. 2017) and Sherlock (He et al. 2013) are available and could be used for this purpose.

For model validation, a three-way data split procedure was considered. Other popular resampling-based methods for model validation such as cross validation (CV) are available. However, it was not possible to perform this due to the inherent computational overheads required in neural network CV tasks. In future work, resampling-based methods such as CV will be considered and explored.

The classifier in this research is not benchmarked against other well-known models since empirical analysis in previous work (Fergus, Curbelo et al. 2018) reported the MLP as the best classifier. However, in future work a classifier comparison may be considered as an extension of this research, particularly

using more advanced neural networking architectures such as Convolutional Neural Networks.

Replication of results using a different and larger dataset is also necessary in future work. Appropriate validation is needed for this method to be considered in clinical practice (Ritchie & Van Steen 2018). The same SNPs associated in at least two independent datasets extracted from the same population also need to be observed, preferably with the same study design.

REFERENCES

Agrawal, R., Imieliński, T. & Swami, A., 1993. Mining association rules between sets of items in large databases. *Proceedings of the 1993 ACM SIGMOD International Conference on Management of Data*, 22(2), pp.207–216. Available at: http://portal.acm.org/citation.cfm?doid=170036.170072.

Agrawal, R. & Srikant, R., 1994. Fast algorithms for mining association rules. In Z. C. Bocca JB, Jarke M, ed. *Proceedings of the 20th International Conference on Very Large Data Bases*. Santiago de Chile: Morgan Kaufman, pp. 487–499. Available at: http://dl.acm.org/citation.cfm?id=672836.

Ahmad, S. et al., 2013. Gene × Physical Activity Interactions in Obesity: Combined Analysis of 111,421 Individuals of European Ancestry D. B. Allison, ed. *PLoS Genetics*, 9(7), p.e1003607. Available at: http://dx.plos.org/10.1371/journal.pgen.1003607.

Ahn, K.-I., 2012. Effective product assignment based on association rule mining in retail. *Expert Systems with Applications*, 39(16), pp.12551–12556. Available at: http://dx.doi.org/10.1016/j.eswa.2012.04.086.

Alberts, B. et al., 2014. *Essential Cell Biology* Fourth edi. J. Clayton, ed., New York: Garland Science, Taylor & Francis Group, LLC, an informa business. Available at: http://books.google.com/books?hl=en&lr=&id=Cg4WAgAAQBAJ&oi=fnd&pg=PR4&dq=Essential+Cell+Biology&ots=yd2R4M_fI3&sig=ZLfYzYDLI6G6o_NAi4IjoAVuKCU.

Alberts, B. et al., 2015. *Molecular Biology of the Cell* Sixth Edit. Garland Science, ed., New York: Garland Science, Taylor & Francis Group.

Altman, N. & Krzywinski, M., 2018. The curse(s) of dimensionality. *Nature Methods*, 15(6), pp.399–400. Available at: https://doi.org/10.1038/s41592-018-0019-x.

Anderson, C.A. et al., 2010. Data quality control in genetic case-control association studies. *Nature protocols*, 5(9), pp.1564–73. Available at: http://www.nature.com/doifinder/10.1038/nprot.2010.116.

Angermueller, C. et al., 2016. Deep learning for computational biology. *Molecular Systems Biology*, 12(7), p.878. Available at: http://msb.embopress.org/lookup/doi/10.15252/msb.20156651.

Anon, 2010. On beyond GWAS. *Nature genetics*, 42(7), p.551. Available at: https://www.nature.com/articles/ng0710-551.pdf.

Apweiler, R. et al., 2018. Whither systems medicine? *Experimental & Molecular Medicine*, 50(3), p.e453. Available at: http://www.nature.com/doifinder/10.1038/emm.2017.290.

Ashburner, M. et al., 2000. Gene Ontology: tool for the unification of biology. *Nature Genetics*, 25(1), pp.25–29. Available at: http://www.nature.com/doifinder/10.1038/75556.

Bailey-Wilson, J.E. & Wilson, A.F., 2011. Linkage Analysis in the Next-Generation Sequencing Era. *Human Heredity*, 72(4), pp.228–236. Available at: www.karger.com.

Ball, M.P. et al., 2012. A public resource facilitating clinical use of genomes. *Proceedings of the National Academy of Sciences*, 109(30), pp.11920–11927. Available at: http://www.pnas.org/lookup/doi/10.1073/pnas.1201904109.

Ball, M.P. et al., 2014. Harvard Personal Genome Project: lessons from participatory public research. *Genome medicine*, 6(2), p.10. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3978420&tool=pmcentrez&rendertype=abstract.

Bao, F., Deng, Y. & Dai, Q., 2016. ACID: Association Correction for Imbalanced Data in GWAS. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 15(1), pp.316–322. Available at: http://ieeexplore.ieee.org/document/7565545/.

Becker, K.G. et al., 2004. The Genetic Association Database. *Nature Genetics*, 36(5), pp.431–432. Available at: http://www.nature.com/articles/ng0504-431.

Bello, A. et al., 2013. Using linked administrative data to study periprocedural mortality in obesity and chronic kidney disease (CKD). *Nephrology Dialysis Transplantation*, 28(suppl 4), pp.iv57-iv64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24084324.

Below, J.E. et al., 2016. Meta-analysis of lipid-traits in Hispanics identifies novel loci, population-specific effects and tissue-specific enrichment of eQTLs. *Scientific Reports*, 6(1), p.19429. Available at: http://www.nature.com/articles/srep19429.

Belsky, D.W. et al., 2013. Development and Evaluation of a Genetic Risk Score for Obesity. *Biodemography and Social Biology*, 59(1), pp.85–100. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23701538.

Bengio, Y. et al., 2007. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems*. MIT Press, pp. 153–160. Available at: http://dl.acm.org/citation.cfm?id=2976456.2976476.

Bergstra, J. et al., 2011. Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems (NIPS)*, 24, pp.2546–2554. Available at: https://hal.inria.fr/hal-00642998.

Bergstra, J. & Bengio, Y., 2012. Random Search for Hyper-Parameter Optimization. *Journal of Machine Learning Research*, 13, pp.281–305. Available at: http://www.jmlr.org/papers/volume13/bergstra12a/bergstra12a.pdf.

Bhaskaran, K. et al., 2014. Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5·24 million UK adults. *The Lancet*, 384(9945), pp.755–765. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25129328.

Bishop, C., 2006. *Pattern Recognition and Machine Learning* 1st ed., New

York: Springer-Verlag New York.

Bjorntorp, P. et al., 2000. Obesity : preventing and managing the global epidemic. *WHO Technical Report Series*, p.253.

Blank, P. & Gutzwiller, F., 2014. Current challenges in handling genetic data. *Swiss Medical Weekly*, 144(August), p.w13998. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25144861.

Bomba, L., Walter, K. & Soranzo, N., 2017. The impact of rare and low-frequency genetic variants in common disease. *Genome Biology*, 18(1), p.77. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408830/pdf/13059_2017_Article_1212.pdf.

Borrell, L.N. & Samuel, L., 2014. Body Mass Index Categories and Mortality Risk in US Adults: The Effect of Overweight and Obesity on Advancing Death. *American Journal of Public Health*, 104(3), pp.512–519. Available at: http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.2013.301597.

Braga, I. et al., 2013. A Note on Parameter Selection for Support Vector Machines. In F. Castro, A. Gelbukh, & M. González, eds. *Advances in Soft Computing and Its Applications. MICAI 2013. Lecture Notes in Computer Science*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, pp. 233–244. Available at: http://link.springer.com/10.1007/978-3-642-45111-9_21.

Brin, S. et al., 1997. Dynamic itemset counting and implication rules for market basket data. *ACM SIGMOD Record*, 26(2), pp.255–264. Available at: http://portal.acm.org/citation.cfm?doid=253262.253325.

Brown, M.S. & Goldstein, J.L., 2009. Cholesterol feedback: from Schoenheimer's bottle to Scap's MELADL. *Journal of Lipid Research*, 50(Supplement), pp.S15–S27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18974038.

Bumgarner, R., 2013. Overview of DNA Microarrays: Types, Applications, and Their Future. In *Current Protocols in Molecular Biology*. Hoboken, NJ,

USA: John Wiley & Sons, Inc., p. Unit 22.1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23288464.

Burton, P.R. et al., 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145), pp.661–678. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17554300.

Bush, W.S. & Moore, J.H., 2012. Chapter 11: Genome-Wide Association Studies F. Lewitter & M. Kann, eds. *PLoS Computational Biology*, 8(12), p.e1002822. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23300413.

Butland, B. et al., 2007. *Foresight Tackling Obesities: Future Choices – Project report*, Available at: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/287937/07-1184x-tackling-obesities-future-choices-report.pdf.

Candel, A. & LeDell, E., 2018. Deep Learning With H2O. , p.55. Available at: http://h2o.ai/resources/.

Cantor, R.M., 2013. Analysis of Genetic Linkage. In *Emery and Rimoin's Principles and Practice of Medical Genetics*. Elsevier, pp. 1–9. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780123838346000100.

Cardon, L.R. & Palmer, L.J., 2003. Population stratification and spurious allelic association. *The Lancet*, 361(9357), pp.598–604. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0140673603125202.

Carey, D.J. et al., 2016. The Geisinger MyCode community health initiative: an electronic health record–linked biobank for precision medicine research. *Genetics in Medicine*, 18(9), pp.906–913. Available at: http://www.nature.com/doifinder/10.1038/gim.2015.187.

Carl Baker, 2018. Obesity Statistics Breefing Paper. *House of Commons Library*, (3336), pp.1–34. Available at: www.noo.org.uk/NOO_about_obesity/mortality.

Chakravarti, A., Clark, A.G. & Mootha, V.K., 2013. Distilling pathophysiology from complex disease genetics. *Cell*, 155(1), pp.21–6. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24074858.

Chambers, J.C. et al., 2008. Common genetic variation near MC4R is associated with waist circumference and insulin resistance. *Nature Genetics*, 40(6), pp.716–718. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18454146.

Chen, K. & Kurgan, L.A., 2012. Neural Networks in Bioinformatics. In *Handbook of Natural Computing*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 565–583. Available at: http://link.springer.com/10.1007/978-3-540-92910-9_18.

CHEN, K.Y. et al., 2012. Redefining the Roles of Sensors in Objective Physical Activity Monitoring. *Medicine & Science in Sports & Exercise*, 44(301), pp.S13–S23. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/pmc3245644/.

Chen, S.-C. et al., 2015. Dynamic association rules for gene expression data analysis. *BMC Genomics*, 16(1), p.786. Available at: http://www.biomedcentral.com/1471-2164/16/786%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/26467206.

Chen, S.-H. et al., 2008. A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology*, 32(2), pp.152–167. Available at: http://doi.wiley.com/10.1002/gepi.20272.

Christensen, K. & Murray, J.C., 2007. What Genome-wide Association Studies Can Do for Medicine. *New England Journal of Medicine*, 356(11), pp.1094–1097. Available at: http://www.nejm.org/doi/abs/10.1056/NEJMp068126.

Chung, C.-M. et al., 2011. A Genome-Wide Association Study Reveals a Quantitative Trait Locus of Adiponectin on CDH13 That Predicts Cardiometabolic Outcomes. *Diabetes*, 60(9), pp.2417–2423. Available at:

http://diabetes.diabetesjournals.org/cgi/doi/10.2337/db10-1321.

Chung, W.K., 2012. An overview of mongenic and syndromic obesities in humans. *Pediatric Blood & Cancer*, 58(1), pp.122–128. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21994130.

Clarke, G.M. et al., 2011. Basic statistical analysis in genetic case-control studies. *Nature Protocols*, 6(2), pp.121–133. Available at: http://www.nature.com/doifinder/10.1038/nprot.2010.182.

Clayton, D.G. et al., 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nature Genetics*, 37(11), pp.1243–1246. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=pubmed&cmd=Retrieve&dopt=AbstractPlus&list_uids=16228001.

Cole, B.S. et al., 2017. Analysis of Gene-Gene Interactions. In *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 1.14.1-1.14.10. Available at: http://doi.wiley.com/10.1002/0471142905.hg0114s70.

Cook, D., 2016. *Practical Machine Learning with H2O: Powerful, Scalable Techniques for Deep Learning and AI* First Edit., O'Reilly Media, Inc. Available at: https://www.safaribooksonline.com/library/view/practical-machine-learning/9781491964590/.

Cooke Bailey, J.N. & Igo, R.P., 2016. Genetic Risk Scores. In *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. 1.29.1-1.29.9. Available at: http://doi.wiley.com/10.1002/cphg.20.

Cordell, H.J., 2002. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20), pp.2463–2468. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12351582.

Corella, D. et al., 2012. Statistical and Biological Gene-Lifestyle Interactions of MC4R and FTO with Diet and Physical Activity on Obesity: New Effects

on Alcohol Consumption K. Dasgupta, ed. *PLoS ONE*, 7(12), p.e52344. Available at: http://dx.plos.org/10.1371/journal.pone.0052344.

Croft, D. et al., 2014. The Reactome pathway knowledgebase. *Nucleic Acids Research*, 42(D1), pp.D472–D477. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24243840.

Cummings, C., 2000. Using DNA Microarrays to Study Host-Microbe Interactions. *Emerging Infectious Diseases*, 6(5), pp.513–525. Available at: http://genome-www4.stanford.edu/Mi-.

Cummings, D.E. & Schwartz, M.W., 2003. Genetics and Pathophysiology of Human Obesity. *Annual Review of Medicine*, 54(1), pp.453–471. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12414915.

Curbelo, C., Fergus, P., Chalmers, C., et al., 2018. Analysis of Extremely Obese Individuals Using Deep Learning Stacked Autoencoders and Genome-Wide Genetic Data. In *15th International Conference on Computational Intelligence methods for Bioinformatics and Biostatistics*. Lisbon, Portugal, pp. 1–13. Available at: http://arxiv.org/abs/1804.06262.

Curbelo, C., Fergus, P., Curbelo, A., et al., 2018. Deep Learning Classification of Polygenic Obesity using Genome Wide Association Study SNPs. In *2018 International Joint Conference on Neural Networks (IJCNN)*. Rio de Janeiro, Brasil: IEEE, pp. 1–8. Available at: https://ieeexplore.ieee.org/document/8489048/.

Curbelo, C., Fergus, P., Hussain, A., Al-Jumeily, D., Dorak, M.T., et al., 2017. Evaluation of Phenotype Classification Methods for Obesity Using Direct to Consumer Genetic Data. In D.-S. Huang, K.-H. Jo, & J. C. Figueroa-García, eds. *Intelligent Computing Theories and Application: 13th International Conference, ICIC 2017*. Liverpool: Springer International Publishing, pp. 350–362. Available at: https://doi.org/10.1007/978-3-319-63312-1_31.

Curbelo, C., Fergus, P., Hussain, A., Al-Jumeily, D., Abdulaimma, B., et al.,

2017. Machine learning approaches for the prediction of obesity using publicly available genetic profiles. In *2017 International Joint Conference on Neural Networks (IJCNN)*. Anchorage, Alaska: IEEE, pp. 2743–2750. Available at: http://ieeexplore.ieee.org/document/7966194/.

D'Eustachio, P. & Schmidt, E., 2003. *Metabolism of carbohydrates R-HSA-71387*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-71387.

Danaee, P., Ghaeini, R. & Hendrix, D.A., 2017. A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 22(4), pp.219–229. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27896977%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5177447.

Dashty, M., 2013. A quick look at biochemistry: Carbohydrate metabolism. *Clinical Biochemistry*, 46(15), pp.1339–1352. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0009912013001677.

Dawn Teare, M. & Barrett, J.H., 2005. Genetic linkage studies. *The Lancet*, 366(9490), pp.1036–1044. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0140673605673825.

Dayem Ullah, A.Z. et al., 2018. SNPnexus: assessing the functional relevance of genetic variation to facilitate the promise of precision medicine. *Nucleic Acids Research*, 46(W1), pp.W109–W113. Available at: www.snp-nexus.org.

Dayem Ullah, A.Z., Lemoine, N.R. & Chelala, C., 2013. A practical guide for the functional annotation of genetic variations using SNPnexus. *Briefings in Bioinformatics*, 14(4), pp.437–447. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt004.

Dayem Ullah, A.Z., Lemoine, N.R. & Chelala, C., 2012. SNPnexus: a web server for functional annotation of novel and publicly known genetic

variants (2012 update). *Nucleic Acids Research*, 40(W1), pp.W65–W70. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gks364.

De, R. et al., 2015. Characterizing gene-gene interactions in a statistical epistasis network of twelve candidate genes for obesity. *BioData mining*, 8(1), p.45. Available at: http://biodatamining.biomedcentral.com/articles/10.1186/s13040-015-0077-x.

Deloukas, P. et al., 2013. Large-scale association analysis identifies new risk loci for coronary artery disease. *Nature Genetics*, 45(1), pp.25–33. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3679547/pdf/nihms-468575.pdf.

Dennis, G. et al., 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome biology*, 4(5), p.P3. Available at: http://dot.ped.med.umich.edu:2000/.

Deo, R.C., 2015. Machine Learning in Medicine. *Circulation*, 132(20), pp.1920–1930. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5831252/pdf/nihms729905.pdf.

Department of Health, 2016. *Health Survey for England 2015 Adult overweight and obesity*,

Van Dijk, E.L. et al., 2014. Ten years of next-generation sequencing technology. *Trends in Genetics*, 30(9).

Disse, E. et al., 2018. An artificial neural network to predict resting energy expenditure in obesity. *Clinical Nutrition*, 37(5), pp.1661–1669. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28893410.

Domingos, P., 2012. A few useful things to know about machine learning. *Communications of the ACM*, 55(10), p.78. Available at:

https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf.

Dougherty, G., 2013. Estimating and Comparing Classifiers. In *Pattern Recognition and Classification - An Introduction*. New York, NY: Springer New York, pp. 157–176. Available at: http://link.springer.com/10.1007/978-1-4614-5323-9_9.

Dowhan, W., 1997. Molecular basis for membrane phospholipid diversity: why are there so many lipids? *Annual review of biochemistry*, 66(1), pp.199–232. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9242906.

Duan, J. et al., 2016. The Next Generation Sequencing and Applications in Clinical Research. In X. Wang et al., eds. *Application of Clinical Bioinformatics*. Dordrecht: Springer Netherlands. Available at: http://link.springer.com/10.1007/978-94-017-7543-4.

Dudbridge, F. & Gusnanto, A., 2008. Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3), pp.227–234. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2573032/pdf/gepi0032-0227.pdf.

Dudoit, S. & van der Laan, M.J., 2008. *Multiple Testing Procedures with Applications to Genomics*, New York, NY: Springer New York. Available at: http://link.springer.com/10.1007/978-0-387-98135-2.

Dunnen, J.T. den & Antonarakis, S.E., 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: A discussion. *Human Mutation*, 15(1), pp.7–12. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10612815.

Durbin, R.M. et al., 2010. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319), pp.1061–1073. Available at: http://www.nature.com/doifinder/10.1038/nature09534.

Ebbert, M.T.W., Ridge, P.G. & Kauwe, J.S.K., 2015. Bridging the gap between statistical and biological epistasis in Alzheimer's disease. *BioMed*

research international, 2015, p.870123. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26075270.

Edwards, S.L. et al., 2013. Beyond GWASs: illuminating the dark road from association to function. *American journal of human genetics*, 93(5), pp.779–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24210251.

Eichler, E.E. et al., 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Reviews Genetics*, 11(6), pp.446–450. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2942068/pdf/nihms218486.pdf.

El-Sayed Moustafa, J.S. & Froguel, P., 2013. From obesity genetics to the future of personalized obesity therapy. *Nature Reviews Endocrinology*, 9(7), pp.402–413. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23529041.

Elston, R.C., Satagopan, J.M. & Sun, S., 2012. Genetic terminology. *Methods in molecular biology (Clifton, N.J.)*, 850, pp.1–9. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4450815/pdf/nihms-689203.pdf.

ENCODE Project Consortium, 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74. Available at: http://www.nature.com/articles/nature11247.

Eraslan, G. et al., 2016. DeepWAS: Directly integrating regulatory information into GWAS using deep learning supports master regulator MEF2C as risk factor for major depressive disorder. *bioRxiv*, p.069096. Available at: http://biorxiv.org/lookup/doi/10.1101/069096.

Esteller, M., 2011. Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12), pp.861–874. Available at: http://dx.doi.org/10.1038/nrg3074.

Evangelou, E. & Ioannidis, J.P.A., 2013. Meta-analysis methods for genome-

wide association studies and beyond. *Nature Reviews Genetics*, 14(6), pp.379–389. Available at: http://www.nature.com/articles/nrg3472.

Fabregat, A., Korninger, F., et al., 2018. Reactome graph database: Efficient access to complex pathway data T. Poisot, ed. *PLOS Computational Biology*, 14(1), p.e1005968. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29377902.

Fabregat, A. et al., 2017. Reactome pathway analysis: A high-performance in-memory approach. *BMC Bioinformatics*.

Fabregat, A., Jupe, S., et al., 2018. The Reactome Pathway Knowledgebase. *Nucleic Acids Research*, 46(D1), pp.D649–D655. Available at: http://academic.oup.com/nar/article/46/D1/D649/4626770.

Fadista, J. et al., 2016. The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants. *European Journal of Human Genetics*, 24(8), pp.1202–1205. Available at: http://www.nature.com/doifinder/10.1038/ejhg.2015.269.

Fakoor, R. et al., 2013. Using deep learning to enhance cancer diagnosis and classification. In *Proceedings of the 30th International Conference on Machine Learning*. Atlanta, Georgia, USA. Available at: http://dl.matlabproject.ir/form/files/392437.pdf.

Fall, T. & Ingelsson, E., 2014. Genome-wide association studies of obesity and metabolic syndrome. *Molecular and Cellular Endocrinology*, 382(1), pp.740–757. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22963884.

Farooqi, I.S. et al., 2002. Beneficial effects of leptin on obesity, T cell hyporesponsiveness, and neuroendocrine/metabolic dysfunction of human congenital leptin deficiency. *Journal of Clinical Investigation*, 110(8), pp.1093–1103. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12393845.

Farooqi, I.S., 2008. Monogenic Human Obesity. In *Obesity and Metabolism*.

Basel: KARGER, pp. 1–11. Available at: https://www.karger.com/Article/FullText/115333.

Fawcett, T., 2004. ROC Graphs : Notes and Practical Considerations for Researchers. *ReCALL*, 31(HPL-2003-4), pp.1–38. Available at: http://www.hpl.hp.com/techreports/2003/HPL-2003-4.pdf.

Fergus, P. et al., 2018. Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp.1–1. Available at: http://arxiv.org/abs/1801.02977.

Fernald, G.H. et al., 2011. Bioinformatics challenges for personalized medicine. *Bioinformatics (Oxford, England)*, 27(13), pp.1741–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21596790.

Ferrari, R. et al., 2015. A genome-wide screening and SNPs-to-genes approach to identify novel genetic risk factors associated with frontotemporal dementia. *Neurobiology of Aging*, 36(10), p.2904.e13-2904.e26. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26154020.

Figueroa, K.P. et al., 2009. Genetic Variance in the Spinocerebellar Ataxia Type 2 (ATXN2) Gene in Children with Severe Early Onset Obesity H. Ulrich, ed. *PLoS ONE*, 4(12), p.e8280. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20016785.

Flegal, K.M. et al., 2007. Cause-Specific Excess Deaths Associated With Underweight, Overweight, and Obesity. *JAMA*, 298(17), p.2028. Available at: http://jama.jamanetwork.com/article.aspx?doi=10.1001/jama.298.17.2028.

Frayling, T.M. et al., 2007. A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science*, 316(5826), pp.889–894. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1141634.

Frazer, K.A. et al., 2009. Human genetic variation and its contribution to complex traits. *Nature Reviews Genetics*, 10(4), pp.241–251. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19293820.

Gamazon, E.R. et al., 2015. A gene-based association method for mapping traits using reference transcriptome data. *Nature genetics*, 47(9), pp.1091–8. Available at: http://www.nature.com/articles/ng.3367.

García-Campos, M.A., Espinal-Enríquez, J. & Hernández-Lemus, E., 2015. Pathway Analysis: State of the Art. *Frontiers in Physiology*, 6(DEC), p.383. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26733877.

Gariepy, G., Nitka, D. & Schmitz, N., 2010. The association between obesity and anxiety disorders in the population: a systematic review and meta-analysis. *International Journal of Obesity*, 34(3), pp.407–419. Available at: http://www.nature.com/articles/ijo2009252.

Gasperskaja, E. & Kučinskas, V., 2017. The most common technologies and tools for functional genome analysis. *Acta medica Lituanica*, 24(1), pp.1–11. Available at: http://www.ncbi.nlm.nih.gov/pubmed/28630587.

Gautron, L., Elmquist, J.K. & Williams, K.W., 2015. Neural Control of Energy Balance: Translating Circuits to Therapies. *Cell*, 161(1), pp.133–145. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4392840/pdf/nihms-673794.pdf.

Gerits, N., Van Belle, W. & Moens, U., 2007. Transgenic mice expressing constitutive active MAPKAPK5 display gender-dependent differences in exploration and activity. *Behavioral and Brain Functions*, 3(1), p.58. Available at: http://www.ncbi.nlm.nih.gov/pubmed/17997833.

Gibbs, R.A. et al., 2015. A global reference for human genetic variation. *Nature*, 526(7571), pp.68–74. Available at: https://www.nature.com/articles/nature15393.pdf.

Gibbs, R.A. et al., 2003. The International HapMap Project. *Nature*, 426(6968),

pp.789–796. Available at: http://www.nature.com/articles/nature02168.

Gilbert-Diamond, D. & Moore, J.H., 2011. Analysis of Gene-Gene Interactions. In *Current Protocols in Human Genetics*. Hoboken, NJ, USA: John Wiley & Sons, Inc., p. Unit1.14. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21735376.

Glorot, X., Bordes, A. & Bengio, Y., 2011. Deep Sparse Rectifier Neural Networks. In Geoffrey Gordon and David Dunson and Miroslav Dudík, ed. *Proceedings of the 14th International Conference on Artificial Intelligence and Statistics (AISTATS)*. Fort Lauderdale, FL, USA: PMLR, pp. 315--323. Available at: http://proceedings.mlr.press/v15/glorot11a/glorot11a.pdf.

GNU Project, R. Available at: http://www.r-project.org.

Gola, D. et al., 2016. A roadmap to multifactor dimensionality reduction methods. *Briefings in Bioinformatics*, 17(2), pp.293–308. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26108231.

Gomez-Cabrero, D., Lluch-Ariet, M., et al., 2014. Synergy-COPD: a systems approach for understanding and managing chronic diseases. *Journal of Translational Medicine*, 12(Suppl 2), p.S2. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25472826.

Gomez-Cabrero, D., Menche, J., et al., 2014. Systems Medicine: from molecular features and models to the clinic in COPD. *Journal of Translational Medicine*, 12(Suppl 2), p.S4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25471042.

Gondro, C. et al., 2013. Quality control for genome-wide association studies. *Methods in molecular biology (Clifton, N.J.)*, 1019(3), pp.129–47. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23756889.

Gonzaga-Jauregui, C., Lupski, J.R. & Gibbs, R.A., 2012. Human Genome Sequencing in Health and Disease. *Annual Review of Medicine*, 63(1), pp.35–61. Available at:

http://www.annualreviews.org/doi/abs/10.1146/annurev-med-051010-162644.

Goodfellow, I., Bengio, Y. & Courville, A., 2016. Deep Learning. *MIT press*, p.1. Available at: http://files.sig2d.org/sig2d14.pdf#page=5.

Gormley, P. et al., 2016. Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine. *Nature Genetics*, 48(8), pp.856–866. Available at: http://www.nature.com/doifinder/10.1038/ng.3598.

Gortmaker, S.L. et al., 2011. Changing the future of obesity: science, policy, and action. *The Lancet*, 378(9793), pp.838–847. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21872752.

Gottesman, O. et al., 2013. The Electronic Medical Records and Genomics (eMERGE) Network: past, present, and future. *Genetics in medicine : official journal of the American College of Medical Genetics*, 15(10), pp.761–71. Available at: http://www.nature.com/doifinder/10.1038/gim.2013.72.

Grada, A. & Weinbrecht, K., 2013. Next-Generation Sequencing: Methodology and Application. *Journal of Investigative Dermatology*, 133(8), pp.1–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23856935.

Gray, I.C., 2000. Single nucleotide polymorphisms as tools in human genetics. *Human Molecular Genetics*, 9(16), pp.2403–2408. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11005795.

Green, E.D., Watson, J.D. & Collins, F.S., 2015. Human Genome Project: Twenty-five years of big biology. *Nature*, 526(7571), pp.29–31. Available at: http://www.nature.com/doifinder/10.1038/526029a.

Gretarsdottir, S. et al., 2010. Genome-wide association study identifies a sequence variant within the DAB2IP gene conferring susceptibility to abdominal aortic aneurysm. *Nature Genetics*, 42(8), pp.692–697. Available at: http://www.nature.com/doifinder/10.1038/ng.622.

Griffin, B.H., Chitty, L.S. & Bitner-Glindzicz, M., 2017. The 100 000 Genomes Project: What it means for paediatrics. *Archives of disease in childhood - Education & practice edition*, 102(2), pp.105–107. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27940446.

Grundberg, E., Kwan, T. & Pastinen, T.M., 2010. Analysis of the impact of genetic variation on human gene expression. *Methods in molecular biology (Clifton, N.J.)*, 628(3), pp.321–39. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20238090.

Gül, H., Aydin Son, Y. & Açikel, C., 2014. Discovering missing heritability and early risk prediction for type 2 diabetes: a new perspective for genome-wide association study analysis with the Nurses' Health Study and the Health Professionals' Follow-Up Study. *Turkish journal of medical sciences*, 44(6), pp.946–54. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25552146.

Gunderson, K.L. et al., 2005. A genome-wide scalable SNP genotyping assay using microarray technology. *Nature Genetics*, 37(5), pp.549–554. Available at: http://www.nature.com/articles/ng1547.

Günther, F., Wawro, N. & Bammann, K., 2009. Neural networks for modeling gene-gene interactions in association studies. *BMC genetics*, 10(1), p.87. Available at: http://www.biomedcentral.com/1471-2156/10/87.

Guo, X. et al., 2018. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. *The American Journal of Human Genetics*, 102(5), pp.890–903. Available at: https://www.sciencedirect.com/science/article/pii/S0002929718301058.

Hahsler, M. et al., 2018. arules: Mining Association Rules and Frequent Itemsets. Available at: https://cran.r-project.org/package=arules.

Hahsler, M., 2017. arulesViz: Interactive Visualization of Association Rules with R. *The R Journal*, 9/2, pp.1–13. Available at: https://journal.r-

project.org/archive/2017/RJ-2017-047/RJ-2017-047.pdf.

Hahsler, M. et al., 2011. The arules R-Package Ecosystem: Analyzing Interesting Patterns from Large Transaction Data Sets. *Journal of Machine Learning Research*, 12, pp.2021–2025. Available at: http://cran.r-project.org.

Hahsler, M. & Chelluboina, S., 2011. Visualizing Association Rules: Introduction to the R-extension Package arulesViz. *R project module*, pp.1–24. Available at: http://www.comp.nus.edu.sg/~zhanghao/project/visualization/[2010]arulesViz.pdf.

Hahsler, M., Grün, B. & Hornik, K., 2005. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. *Journal of Statistical Software*, 14(15). Available at: http://www.jstatsoft.org/v14/i15/.

Haibo He & Garcia, E.A., 2009. Learning from Imbalanced Data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9), pp.1263–1284. Available at: http://ieeexplore.ieee.org/document/5128907/.

Hall, K.D., 2018. Did the Food Environment Cause the Obesity Epidemic? *Obesity*, 26(1), pp.11–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29265772.

Harbron, J. et al., 2014. Fat Mass and Obesity-Associated (FTO) Gene Polymorphisms Are Associated with Physical Activity, Food Intake, Eating Behaviors, Psychological Health, and Modeled Change in Body Mass Index in Overweight/Obese Caucasian Adults. *Nutrients*, 6(8), pp.3130–3152. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25102252.

Hardy, J. & Singleton, A., 2009. Genomewide Association Studies and Human Disease. *New England Journal of Medicine*, 360(17), pp.1759–1768. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19369657.

He, H. et al., 2016. Biostatistics, Data Mining and Computational Modeling. In B. J. Wang X., Baumgartner C., Shields D., Deng HW., ed. *Application of Clinical Bioinformatics*. Dordrecht: Springer, Dordrecht, pp. 23–57. Available at: http://link.springer.com/10.1007/978-94-017-7543-4_2.

He, L. & Hannon, G.J., 2004. MicroRNAs: Small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, 5(7), pp.522–531.

He, X. et al., 2013. Sherlock: Detecting Gene-Disease Associations by Matching Patterns of Expression QTL and GWAS. *The American Journal of Human Genetics*, 92(5), pp.667–680. Available at: https://www.sciencedirect.com/science/article/pii/S0002929713001596.

Heard-Costa, N.L. et al., 2009. NRXN3 is a novel locus for waist circumference: a genome-wide association study from the CHARGE Consortium. *PLoS genetics*, 5(6), p.e1000539. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19557197.

Heid, I.M., 2005. Association of the 103I MC4R allele with decreased body mass in 7937 participants of two population based surveys. *Journal of Medical Genetics*, 42(4), pp.e21–e21. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15805150.

Herrera, B.M. & Lindgren, C.M., 2010. The Genetics of Obesity. *Current Diabetes Reports*, 10(6), pp.498–505. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20931363.

Higdon, R. et al., 2013. Unraveling the Complexities of Life Sciences Data. *Big Data*, 1(1), pp.42–50. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27447037.

Hilton, S., Patterson, C. & Teyhan, A., 2012. Escalating Coverage of Obesity in UK Newspapers: The Evolution and Framing of the “Obesity Epidemic” From 1996 to 2010. *Obesity*, 20(8), pp.1688–1695. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22318314.

Himes, B.E. et al., 2013. ITGB5 and AGFG1 variants are associated with

severity of airway responsiveness. *BMC Medical Genetics*, 14(1), p.86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23984888.

Hinney, A. et al., 2007. Genome wide association (GWA) study for early onset extreme obesity supports the role of fat mass and obesity associated gene (FTO) variants. *PloS one*, 2(12), p.e1361. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18159244.

Hinney, A. & Giuranna, J., 2018. Polygenic Obesity. In *Pediatric Obesity*. Cham: Springer International Publishing, pp. 183–202. Available at: http://link.springer.com/10.1007/978-3-319-68192-4_10.

Hinrichs, A.S. et al., 2016. UCSC Data Integrator and Variant Annotation Integrator. *Bioinformatics*, 32(9), pp.1430–1432. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26740527.

Hinton, G.E. & Salakhutdinov, R.R., 2006. Reducing the dimensionality of data with neural networks. *Science (New York, N.Y.)*, 313(5786), pp.504–7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16873662.

Hipp, J., Güntzer, U. & Nakhaeizadeh, G., 2000. Algorithms for association rule mining - a general survey and comparison. *ACM SIGKDD Explorations Newsletter*, 2(1), pp.58–64. Available at: http://portal.acm.org/citation.cfm?doid=360402.360421.

Hirko, K.A. et al., 2015. Body Mass Index in Young Adulthood, Obesity Trajectory, and Premature Mortality. *American Journal of Epidemiology*, 182(5), pp.441–450. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25977515.

Hoens, T.R. & Chawla, N. V, 2013. Imbalanced Datasets: From Sampling to Classifiers. In T. I. of E. and E. Engineers, ed. *Imbalanced Learning*. Hoboken, NJ, USA: John Wiley & Sons, Inc., pp. 43–59. Available at: http://doi.wiley.com/10.1002/9781118646106.ch3.

Hoh, J. et al., 2000. Selecting SNPs in two-stage analysis of disease association data: a model-free approach. *Annals of human genetics*, 64(Pt 5), pp.413–

7. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11281279.

Hong, S.-H. et al., 2012. Minibrain/Dyrk1a Regulates Food Intake through the Sir2-FOXO-sNPF/NPY Pathway in Drosophila and Mammals P. Kapahi, ed. *PLoS Genetics*, 8(8), p.e1002857. Available at: www.plosgenetics.org.

Hoo-Chang Shin et al., 2013. Stacked Autoencoders for Unsupervised Feature Learning and Multiple Organ Detection in a Pilot Study Using 4D Patient Data. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8), pp.1930–1943. Available at: http://ieeexplore.ieee.org/document/6399478/.

Howard, R., Carriquiry, A.L. & Beavis, W.D., 2014. Parametric and nonparametric statistical methods for genomic selection of traits with additive and epistatic genetic architectures. *G3 (Bethesda, Md.)*, 4(6), pp.1027–46. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24727289.

Hruby, A. & Hu, F.B., 2015. The Epidemiology of Obesity: A Big Picture. *PharmacoEconomics*, 33(7), pp.673–689. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25471927.

Huang, D.W., Sherman, B.T. & Lempicki, R.A., 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols*, 4(1), pp.44–57. Available at: http://www.nature.com/articles/nprot.2008.211.

Hung, C.-F. et al., 2015. A genetic risk score combining 32 SNPs is associated with body mass index and improves obesity prediction in people with major depressive disorder. *BMC Medicine*, 13(1), p.86. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25903154.

Illumina, 2010. *Genome-Wide DNA Analysis BeadChips*, San Diego, CA. Available at: https://www.illumina.com/content/dam/illumina-marketing/documents/products/brochures/datasheet_omni_whole-genome_arrays.pdf.

Iniesta, R., Stahl, D. & McGuffin, P., 2016. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine*, 46(12), pp.2455–2465. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27406289.

Jain, M. et al., 2018. Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, 36(4), pp.338–345. Available at: https://www.nature.com/articles/nbt.4060.pdf.

James, W.P.T., 2008. WHO recognition of the global obesity epidemic. *International Journal of Obesity*, 32(S7), pp.S120–S126. Available at: http://www.nature.com/articles/ijo2008247.

Jassal, B., 2008. *Biological oxidations R-HSA-211859*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-211859.

Jassal, B., Gillespie, M.E., Gopinathrao, G. & D'Eustachio, P., 2007. *Fatty acid metabolism R-HSA-8978868*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-8978868.

Jassal, B., Gillespie, M.E., Gopinathrao, G. & D'Eustachio, P., 2007. *Metabolism of lipids R-HSA-556833*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-556833.

Jassal, B., 2003. *Metabolism of nucleotides R-HSA-15869*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-15869.

Jassal, B., 2007a. *Metabolism of steroids R-HSA-8957322*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-8957322.

Jassal, B., 2007b. *Metabolism of vitamins and cofactors R-HSA-196854*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-196854.

Jassal, B., 2011. *Metabolism Pathway R-HSA-1430728*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-1430728.

Jensen, M.D. et al., 2014. 2013 AHA/ACC/TOS Guideline for the Management

of Overweight and Obesity in Adults. *Circulation*, 129(25 suppl 2), pp.S102–S138. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24222017.

Jiang, X. et al., 2011. Learning genetic epistasis using Bayesian network scoring criteria. *BMC Bioinformatics*, 12(1), p.89. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21453508.

Jiao, X. et al., 2012. DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinformatics*, 28(13), pp.1805–1806. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22543366.

Jin, L. et al., 2014. Pathway-based analysis tools for complex diseases: a review. *Genomics, proteomics & bioinformatics*, 12(5), pp.210–20. Available at: http://www.sciencedirect.com/science/article/pii/S1672022914001065.

Jones, A. et al., 2007. Foresight Tackling Obesities : Future Choices – Obesogenic Environments – Evidence Review. , p.52. Available at: www.foresight.gov.uk.

Joshi-Tope, G., 2004. Reactome: a knowledgebase of biological pathways. *Nucleic Acids Research*, 33(Database issue), pp.D428–D432. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gki072.

Jou, C., 2014. The Biology and Genetics of Obesity — A Century of Inquiries. *New England Journal of Medicine*, 370(20), pp.1874–1877. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24827033.

Kamitsuji, S. et al., 2015. Japan PGx Data Science Consortium Database: SNPs and HLA genotype data from 2994 Japanese healthy individuals for pharmacogenomics studies. *Journal of Human Genetics*, 60(6), pp.319–326. Available at: http://www.nature.com/doifinder/10.1038/jhg.2015.23.

Kanehisa, M. & Goto, S., 2000. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1), pp.27–30. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10592173.

Karki, R. et al., 2015. Defining “mutation” and “polymorphism” in the era of personal genomics. *BMC Medical Genomics*, 8(1), p.37. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26173390.

Khatri, P., Sirota, M. & Butte, A.J., 2012. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges C. A. Ouzounis, ed. *PLoS Computational Biology*, 8(2), p.e1002375. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22383865.

Kidd, J.M. et al., 2008. Mapping and sequencing of structural variation from eight human genomes. *Nature*, 453(7191), pp.56–64. Available at: http://www.nature.com/articles/nature06862.

Kilpeläinen, T.O. et al., 2011. Physical Activity Attenuates the Influence of FTO Variants on Obesity Risk: A Meta-Analysis of 218,166 Adults and 19,268 Children C. Lewis, ed. *PLoS Medicine*, 8(11), p.e1001116. Available at: www.plosmedicine.org.

Kim, M.-S. et al., 2006. Role of hypothalamic Foxo1 in the regulation of food intake and energy homeostasis. *Nature Neuroscience*, 9(7), pp.901–906. Available at: http://www.nature.com/articles/nn1731.

Kirk, S.F.L. & Penney, T.L., 2013. The Role of Health Systems in Obesity Management and Prevention: Problems and Paradigm Shifts. *Current Obesity Reports*, 2(4), pp.315–319. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24273700.

Klein, R.J., 2005. Complement Factor H Polymorphism in Age-Related Macular Degeneration. *Science*, 308(5720), pp.385–389. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15761122.

Klemettinen, M. et al., 1994. Finding interesting rules from large sets of discovered association rules. In *Proceedings of the third international conference on Information and knowledge management - CIKM ’94*. New York, New York, USA: ACM Press, pp. 401–407. Available at: http://dl.acm.org/citation.cfm?id=191314.

Komoda, T. & Matsunaga, T., 2015. Metabolic Pathways in the Human Body. In *Biochemistry for Medical Professionals*. Elsevier, pp. 25–63. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780128019184000049.

Koo, C.L. et al., 2013. A Review for Detecting Gene-Gene Interactions Using Machine Learning Methods in Genetic Epidemiology. *BioMed Research International*, 2013, pp.1–13. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24228248.

Koonin, E. V, 2012. Does the central dogma still stand? *Biology Direct*, 7(1), p.27. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22913395.

Kourou, K. et al., 2015. Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, 13, pp.8–17. Available at: http://www.ncbi.nlm.nih.gov/pubmed/25750696.

Kraljevic, T., 2018. H2O: R Interface for "H2O." Available at: https://cran.r-project.org/package=h2o.

Kramer, F., Just, S. & Zeller, T., 2018. New perspectives: systems medicine in cardiovascular disease. *BMC Systems Biology*, 12(1), p.57. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29699591.

Kruppa, J., Ziegler, A. & König, I.R., 2012. Risk estimation and risk prediction using machine-learning methods. *Human Genetics*, 131(10), pp.1639–1654. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22752090.

Kundaje, A. et al., 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), pp.317–330. Available at: http://www.nature.com/articles/nature14248.

Lamy, P., Grove, J. & Wiuf, C., 2011. A review of software for microarray genotyping. *Human genomics*, 5(4), pp.304–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21712191.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921. Available at:

https://www.nature.com/articles/35057062.pdf.

Langlet, F. et al., 2017. Selective Inhibition of FOXO1 Activator/Repressor Balance Modulates Hepatic Glucose Handling. *Cell*, 171(4), p.824–835.e18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29056338.

Laurie, C.C. et al., 2010. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genetic Epidemiology*, 34(6), pp.591–602.

Le, Q. V, 2015. *A Tutorial on Deep Learning Part 2: Autoencoders, Convolutional Neural Networks and Recurrent Neural Networks*, Mountain View, CA.

Lee, P.H. et al., 2012. INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics*, 28(13), pp.1797–1799. Available at: http://atgu.mgh.harvard.edu/inrich/.

Lee, S., Kwon, M.-S. & Park, T., 2012. Network Graph Analysis of Gene-Gene Interactions in Genome-Wide Association Study Data. *Genomics & Informatics*, 10(4), p.256. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23346039.

Lever, J., Krzywinski, M. & Altman, N., 2016. Model selection and overfitting. *Nature Methods*, 13(9), pp.703–704. Available at: https://www.nature.com/articles/nmeth.3968.pdf.

Lewis, C.M., 2002. Genetic association studies: Design, analysis and interpretation. *Briefings in Bioinformatics*, 3(2), pp.146–153. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12139434.

Lewis, C.M. et al., 2010. Genome-Wide Association Study of Major Recurrent Depression in the U.K. Population. *American Journal of Psychiatry*, 167(8), pp.949–957. Available at: http://psychiatryonline.org/doi/abs/10.1176/appi.ajp.2010.09091380.

Li, D. et al., 2013. Using association rule mining for phenotype extraction from

electronic health records. *AMIA Joint Summits on Translational Science proceedings. AMIA Joint Summits on Translational Science*, 2013, pp.142–6. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3845788/pdf/AMIA_CRI_2013_142.pdf.

Li, S., Zhao, J.H., Luan, J., Luben, R.N., et al., 2010. Cumulative effects and predictive value of common obesity-susceptibility variants identified by genome-wide association studies. *The American Journal of Clinical Nutrition*, 91(1), pp.184–190. Available at: http://www.hapmap.

Li, S., Zhao, J.H., Luan, J., Ekelund, U., et al., 2010. Physical activity attenuates the genetic predisposition to obesity in 20,000 men and women from EPIC-Norfolk prospective population study. *PLoS Medicine*, 7(8), pp.1–9.

Li, W., 2007. Three lectures on case control genetic association analysis. *Briefings in Bioinformatics*, 9(1), pp.1–13. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbm058.

Liang, J. et al., 2017. Single-trait and multi-trait genome-wide association analyses identify novel loci for blood pressure in African-ancestry populations G. Gibson, ed. *PLOS Genetics*, 13(5), p.e1006728. Available at: http://dx.plos.org/10.1371/journal.pgen.1006728.

Liberzon, A. et al., 2015. The Molecular Signatures Database Hallmark Gene Set Collection. *Cell Systems*, 1(6), pp.417–425. Available at: http://dx.doi.org/10.1016/j.cels.2015.12.004http://dx.doi.org/10.1016/j.cels.2015.12.004.

Lifestyles statistics team. Health and Social Care Information Centre, 2014. *Statistics on Obesity, Physical Activity and Diet: England 2014*, England. Available at: http://www.hscic.gov.uk/catalogue/PUB13648/Obes-phys-acti-diet-eng-2014-rep.pdf.

Lindgren, C.M. et al., 2009. Genome-Wide Association Scan Meta-Analysis Identifies Three Loci Influencing Adiposity and Fat Distribution D. B.

Allison, ed. *PLoS Genetics*, 5(6), p.e1000508. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19557161.

Ling, H. et al., 2015. Junk DNA and the long non-coding RNA twist in cancer genetics. *Oncogene*, 34(39), pp.5003–5011. Available at: http://www.nature.com/articles/onc2014456.

Lisboa, P.J. & Taktak, A.F.G., 2006. The use of artificial neural networks in decision support in cancer: A systematic review. *Neural Networks*, 19(4), pp.408–415. Available at: www.elsevier.com/locate/neunet.

Liu, B., Hsu, W. & Ma, Y., 1999. Pruning and summarizing the discovered associations. In *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '99*. New York, New York, USA: ACM Press, pp. 125–134. Available at: http://portal.acm.org/citation.cfm?doid=312129.312216.

Lobstein, T., Baur, L. & Uauy, R., 2004. Obesity in children and young people: a crisis in public health. *Obesity Reviews*, 5(s1), pp.4–85. Available at: http://doi.wiley.com/10.1111/j.1467-789X.2004.00133.x.

Locke, A.E. et al., 2015. Genetic studies of body mass index yield new insights for obesity biology. *Nature*, 518(7538), pp.197–206. Available at: http://www.nature.com/doifinder/10.1038/nature14177.

Lonsdale, J. et al., 2013. The Genotype-Tissue Expression (GTEx) project. *Nature Genetics*, 45(6), pp.580–585. Available at: http://www.nature.com/articles/ng.2653.

Loos, R.J.F. et al., 2008. Common variants near MC4R are associated with fat mass, weight and risk of obesity. *Nature Genetics*, 40(6), pp.768–775. Available at: http://www.nature.com/articles/ng.140.

Loos, R.J.F., 2012. Genetic determinants of common obesity and their value in prediction. *Best Practice & Research Clinical Endocrinology & Metabolism*, 26(2), pp.211–226. Available at: http://www.ncbi.nlm.nih.gov/pubmed/22498250.

Loos, R.J.F. & Yeo, G.S.H., 2014. The bigger picture of FTO—the first GWAS-identified obesity gene. *Nature Reviews Endocrinology*, 10(1), pp.51–61. Available at: http://www.nature.com/articles/nrendo.2013.227.

De los Campos, G. et al., 2010. Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics research*, 92(4), pp.295–308. Available at: http://www.journals.cambridge.org/abstract_S0016672310000285.

Louhelainen, J., 2016. SNP Arrays. *Microarrays (Basel, Switzerland)*, 5(4). Available at: http://www.ncbi.nlm.nih.gov/pubmed/27792140.

Lunetta, K.L. et al., 2004. Screening large-scale association study data: exploiting interactions using random forests. *BMC genetics*, 5, p.32. Available at: http://www.biomedcentral.com/1471-2156/5/32.

Lyons, J. et al., 2014. Predicting backbone Cα angles and dihedrals from protein sequences by stacked sparse auto-encoder deep neural network. *Journal of Computational Chemistry*, 35(28), pp.2040–2046. Available at: http://doi.wiley.com/10.1002/jcc.23718.

Ma, L. et al., 2010. An “almost exhaustive” search-based sequential permutation method for detecting epistasis in disease association studies. *Genetic epidemiology*, 34(5), pp.434–43. Available at: http://doi.wiley.com/10.1002/gepi.20496.

MacArthur, J. et al., 2017. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45(D1), pp.D896–D901. Available at: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1133.

Maes, H.H., Neale, M.C. & Eaves, L.J., 1997. Genetic and environmental factors in relative body weight and human adiposity. *Behavior genetics*, 27(4), pp.325–51. Available at: http://www.ncbi.nlm.nih.gov/pubmed/9519560.

Maher, B., 2008. Personal genomes: The case of the missing heritability.

Nature, 456(7218), pp.18–21. Available at: https://www.nature.com/news/2008/081105/pdf/456018a.pdf.

Majewski, J. et al., 2011. What can exome sequencing do for you? *Journal of medical genetics*, 48(9), pp.580–589.

Malis, C. et al., 2005. Total and regional fat distribution is strongly influenced by genetic factors in young and elderly twins. *Obesity research*, 13(12), pp.2139–45. Available at: http://doi.wiley.com/10.1038/oby.2005.265.

Man-Hon Wong et al., 2015. Discovering Binding Cores in Protein-DNA Binding Using Association Rule Mining with Statistical Measures. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 12(1), pp.142–154. Available at: http://ieeexplore.ieee.org/document/6867331/.

Manning, T., Sleator, R.D. & Walsh, P., 2014. Biologically inspired intelligent decision making. *Bioengineered*, 5(2), pp.80–95. Available at: http://www.tandfonline.com/action/journalInformation?journalCode=kbie20.

Manolio, T.A. et al., 2009. Finding the missing heritability of complex diseases. *Nature*, 461(7265), pp.747–753. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2831613/pdf/nihms175346.pdf.

Manolio, T.A., 2010. Genomewide Association Studies and Assessment of the Risk of Disease W. G. Feero & A. E. Guttmacher, eds. *New England Journal of Medicine*, 363(2), pp.166–176. Available at: http://www.nejm.org/doi/10.1056/NEJMra0905980.

Mantovani, R.G. et al., 2015. Effectiveness of Random Search in SVM hyper-parameter tuning. In *2015 International Joint Conference on Neural Networks (IJCNN)*. IEEE, pp. 1–8. Available at: http://ieeexplore.ieee.org/document/7280664/.

Marchini, J., Donnelly, P. & Cardon, L.R., 2005. Genome-wide strategies for

detecting multiple loci that influence complex diseases. *Nature Genetics*, 37(4), pp.413–417. Available at: http://www.nature.com/doifinder/10.1038/ng1537.

Mardis, E.R., 2008. The impact of next-generation sequencing technology on genetics. *Trends in Genetics*, 24(3), pp.133–141. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168952508000231.

Marques, F.Z., Booth, S.A. & Charchar, F.J., 2015. The emerging role of non-coding RNA in essential hypertension and blood pressure regulation. *Journal of Human Hypertension*, 29(8), pp.459–467. Available at: http://www.nature.com/articles/jhh201499.

Marx, V., 2013. Biology: The big challenges of big data. *Nature*, 498(7453), pp.255–260. Available at: http://www.nature.com.gate1.inist.fr/nature/journal/v498/n7453/full/498255a.html%5Cnhttp://www.nature.com.gate1.inist.fr/nature/journal/v498/n7453/pdf/498255a.pdf.

Mattson, D.L. & Liang, M., 2017. From GWAS to functional genomics-based precision medicine. *Nature Reviews Nephrology*, 13(4), pp.195–196. Available at: http://dx.doi.org/10.1038/nrneph.2017.21.

McCarthy, M.I. et al., 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5), pp.356–369. Available at: http://dx.doi.org/10.1038/nrg2344%5Cnhttp://www.nature.com/nrg/journal/v9/n5/abs/nrg2344.html%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/18398418.

McCarty, C.A. et al., 2011. The eMERGE Network: A consortium of biorepositories linked to electronic medical records data for conducting genomic studies. *BMC Medical Genomics*, 4(1), p.13. Available at: http://bmcmedgenomics.biomedcentral.com/articles/10.1186/1755-8794-4-13.

McCullagh, P., 1984. Generalized linear models. *European Journal of Operational Research*, 16(3), pp.285–292. Available at: http://www.jstor.org/stable/2344614.

McKinney, B.A. et al., 2006. Machine Learning for Detecting Gene-Gene Interactions. *Applied Bioinformatics*, 5(2), pp.77–88. Available at: http://link.springer.com/article/10.2165/00822942-200605020-00002%5Cnhttp://link.springer.com/article/10.2165/00822942-200605020-00002.

Mehlhorn, H., 2010. *Pediatric Obesity* M. Freemark, ed., New York, NY: Springer New York. Available at: http://link.springer.com/10.1007/978-1-60327-874-4.

Meyre, D. et al., 2009. Genome-wide association study for early-onset and morbid adult obesity identifies three new risk loci in European populations. *Nature Genetics*, 41(2), pp.157–159. Available at: http://www.nature.com/doifinder/10.1038/ng.301.

Milani, D. et al., 2014. Syndromic obesity: clinical implications of a correct diagnosis. *Italian Journal of Pediatrics*, 40(1), p.33. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4230022/pdf/1824-7288-40-33.pdf.

Min, J., Chiu, D.T. & Wang, Y., 2013. Variation in the heritability of body mass index based on diverse twin studies: a systematic review. *Obesity Reviews*, 14(11), pp.871–882. Available at: http://doi.wiley.com/10.1111/obr.12065.

Montague, C.T. et al., 1997. Congenital leptin deficiency is associated with severe early-onset obesity in humans. *Nature*, 387(6636), pp.903–908. Available at: http://www.nature.com/articles/43185.

Moore, J.H. et al., 2006. A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility. *Journal of Theoretical Biology*,

241(2), pp.252–261. Available at: http://www.epistasis.org.

Moore, J.H. & Andrews, P.C., 2015. Epistasis Analysis Using Multifactor Dimensionality Reduction. In *Epistasis. Methods in Molecular Biology*. New York: Humana Press, New York, NY, pp. 301–314. Available at: http://link.springer.com/10.1007/978-1-4939-2155-3_16.

Moore, J.H. & White, B.C., 2007. Tuning ReliefF for Genome-Wide Genetic Analysis. In Marchiori E., M. J.H., & R. J.C., eds. *Evolutionary Computation,Machine Learning and Data Mining in Bioinformatics*. Berlin, Heidelberg: Springer, Berlin, Heidelberg, pp. 166–175. Available at: http://link.springer.com/10.1007/978-3-540-71783-6_16.

Moore, J.H. & Williams, S.M., 2009. Epistasis and Its Implications for Personal Genetics. *The American Journal of Human Genetics*, 85(3), pp.309–320. Available at: http://dx.doi.org/10.1016/j.ajhg.2009.08.006.

Moore, J.H. & Williams, S.M., 2005. Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. *BioEssays*, 27(6), pp.637–646. Available at: http://doi.wiley.com/10.1002/bies.20236.

Morandi, A. et al., 2012. Estimation of Newborn Risk for Child or Adolescent Obesity: Lessons from Longitudinal Birth Cohorts M. Manco, ed. *PLoS ONE*, 7(11), p.e49919. Available at: www.plosone.org.

Morris, A.P. & Zeggini, E., 2010. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic Epidemiology*, 34(2), pp.188–193. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19810025.

Muñoz Yáñez, C., García Vargas, G.G. & Pérez-Morales, R., 2017. Monogenic , Polygenic and Multifactorial Obesity in Children : Genetic and Environmental Factors. *Austin Journal of Nutrition & Metabolism*, 4(3), pp.1–12. Available at: https://www.researchgate.net/publication/321137176.

Mutch, D.M. & Clément, K., 2006. Genetics of human obesity. *Best Practice & Research Clinical Endocrinology & Metabolism*, 20(4), pp.647–664. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1521690X06000777.

Nakka, P., Raphael, B.J. & Ramachandran, S., 2016. Gene and network analysis of common variants reveals novel associations in multiple complex diseases. *Genetics*, 204(2), pp.783–798. Available at: http://www.genetics.org/cgi/doi/10.1534/genetics.116.188391.

Namjou, B. et al., 2015. A GWAS Study on Liver Function Test Using eMERGE Network Participants L. Prokunina-Olsson, ed. *PLOS ONE*, 10(9), p.e0138677. Available at: http://dx.plos.org/10.1371/journal.pone.0138677.

National Clinical Guideline Centre, 2014. *Obesity: Identification, Assessment and Management of Overweight and Obesity in Children, Young People and Adults: Partial Update of CG43*, Available at: https://www.ncbi.nlm.nih.gov/books/NBK264165/pdf/Bookshelf_NBK264165.pdf.

Naulaerts, S. et al., 2015. A primer to frequent itemset mining for bioinformatics. *Briefings in Bioinformatics*, 16(2), pp.216–231. Available at: https://academic.oup.com/bib/article-lookup/doi/10.1093/bib/bbt074.

NCD Risk Factor Collaboration (NCD-RisC), 2016. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19·2 million participants. *The Lancet*, 387(10026), pp.1377–1396. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27115820.

Neel, J. V, 1962. Diabetes mellitus: a “thrifty” genotype rendered detrimental by “progress”? *American journal of human genetics*, 14(8), pp.353–62. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10516792.

Nelson, C.P. et al., 2017. Association analyses based on false discovery rate

implicate new loci for coronary artery disease. *Nature Genetics*, 49(9), pp.1385–1391. Available at: http://www.nature.com/doifinder/10.1038/ng.3913.

Ng, A., 2011. Sparse Autoencoder. In *CS294A Lecture notes*. pp. 1–19. Available at: http://www.stanford.edu/class/cs294a/.

Nica, A.C. & Dermitzakis, E.T., 2008. Using gene expression to investigate the genetic basis of complex disorders. *Human Molecular Genetics*, 17(R2), pp.R129–R134. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18852201.

Nicolae, D.L. et al., 2010. Trait-Associated SNPs Are More Likely to Be eQTLs: Annotation to Enhance Discovery from GWAS G. Gibson, ed. *PLoS Genetics*, 6(4), p.e1000888. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20369019.

Niel, C. et al., 2015. A survey about methods dedicated to epistasis detection. *Frontiers in Genetics*, 6(SEP). Available at: http://journal.frontiersin.org/Article/10.3389/fgene.2015.00285/abstract.

Noble, W.S., 2009. How does multiple testing correction work? *Nature Biotechnology*, 27(12), pp.1135–1137. Available at: https://www.nature.com/articles/nbt1209-1135.pdf.

Nordang, G.B.N. et al., 2017. Next-generation sequencing of the monogenic obesity genes LEP , LEPR , MC4R , PCSK1 and POMC in a Norwegian cohort of patients with morbid obesity and normal weight controls. *Molecular Genetics and Metabolism*, 121(1), pp.51–56. Available at: http://linkinghub.elsevier.com/retrieve/pii/S1096719216308319.

Nykodym, T. et al., 2018. Generalized Linear Modeling with H2O. Available at: http://h2o.ai/resources/.

O'Rahilly, S., 2009. Human genetics illuminates the paths to metabolic disease. *Nature*, 462(7271), pp.307–314. Available at: http://dx.doi.org/10.1038/nature08532.

de Onis, M., Blössner, M. & Borghi, E., 2010. Global prevalence and trends of overweight and obesity among preschool children. *The American journal of clinical nutrition*, 92(5), pp.1257–64. Available at: https://academic.oup.com/ajcn/article/92/5/1257/4597558.

Org, E. et al., 2009. Genome-wide scan identifies CDH13 as a novel susceptibility locus contributing to blood pressure determination in two European populations. *Human Molecular Genetics*, 18(12), pp.2288–2296. Available at: https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddp135.

Orlic-Milacic, M., 2018a. *FOXO-mediated transcription of oxidative stress, metabolic and neuronal genes R-HSA-9615017*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-9615017.

Orlic-Milacic, M., 2018b. *FOXO1 and SIN3A: HDAC complex bind GCK gene promoter R-HSA-9625101*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-9625101.

Orlic-Milacic, M., 2018c. *FOXO1 binds NPY gene promoter R-HSA-9622980*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-9622980.

Orlic-Milacic, M., 2018d. *GCK gene expression is inhibited by FOXO1, SIN3A and HDACs R-HSA-9625124*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-9625124.

Orlic-Milacic, M., 2018e. *NPY gene expression is stimulated by FOXO1 R-HSA-9622981*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-9622981.

Owen, A.B., 2007. Infinitely Imbalanced Logistic Regression. *Journal of Machine Learning Research*, 8, pp.761–773. Available at: http://www.jmlr.org/papers/volume8/owen07a/owen07a.pdf.

Paliwal, M. & Kumar, U.A., 2009. Neural networks and statistical techniques: A review of applications. *Expert Systems with Applications*, 36(1), pp.2–

17. Available at: www.elsevier.com/locate/eswa.

Panagiotou, O.A. & Ioannidis, J.P.A., 2012. What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *International Journal of Epidemiology*, 41(1), pp.273–286. Available at: http://www.ije.oxfordjournals.org/lookup/doi/10.1093/ije/dyr178.

Pang, G. et al., 2014. Energy intake, metabolic homeostasis, and human health. *Food Science and Human Wellness*, 3(3–4), pp.89–103. Available at: https://ac.els-cdn.com/S221345301500004X/1-s2.0-S221345301500004X-main.pdf?_tid=b55c4dae-036f-4d7a-8717-1a0eadf3e65a&acdnat=1520011183_a130d9fbaa6dbf559ae5049a97a36d41.

Peeters, A. et al., 2003. Obesity in Adulthood and Its Consequences for Life Expectancy: A Life-Table Analysis. *Annals of Internal Medicine*, 138(1), p.24.

Pereira, D.M. et al., 2015. "Omics" Technologies. In *Principles of Translational Science in Medicine*. Elsevier, pp. 25–39. Available at: http://dx.doi.org/10.1016/B978-0-12-800687-0.00003-7.

Perneger, T. V, 1998. What's wrong with Bonferroni adjustments. *BMJ*, 316(7139), pp.1236–1238. Available at: www.bmj.com.

Peschansky, V.J. & Wahlestedt, C., 2014. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics*, 9(1), pp.3–12. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3928183/.

Peter Armitage, Geoffrey Berry, J.N.S.M., 2001. *Statistical Methods in Medical Research* 4th Editio., John Wiley and Sons.

Phillips, P.C., 2008. Epistasis — the essential role of gene interactions in the structure and evolution of genetic systems. *Nature Reviews Genetics*, 9(11), pp.855–867. Available at: http://www.nature.com/doifinder/10.1038/nrg2452.

Phillips, P.C., 1998. The language of gene interaction. *Genetics*, 149(3), pp.1167–71. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1460246/pdf/9649511.pdf.

Pigeyre, M. et al., 2016. Recent progress in genetics, epigenetics and metagenomics unveils the pathophysiology of human obesity. *Clinical Science*, 130(12), pp.943–986. Available at: http://clinsci.org/cgi/doi/10.1042/CS20160136.

Piotrowski, A.P. & Napiorkowski, J.J., 2013. A comparison of methods to avoid overfitting in neural networks training in the case of catchment runoff modelling. *Journal of Hydrology*, 476, pp.97–111. Available at: http://dx.doi.org/10.1016/j.jhydrol.2012.10.019.

Pirinen, M., Donnelly, P. & Spencer, C.C.A., 2012. Including known covariates can reduce power to detect genetic effects in case-control studies. *Nature Genetics*, 44(8), pp.848–851. Available at: http://www.nature.com/articles/ng.2346.

Pirmohamed, M., 2011. Pharmacogenetics: Past, present and future. *Drug Discovery Today*, 16(19–20), pp.852–861. Available at: http://dx.doi.org/10.1016/j.drudis.2011.08.006.

Polderman, T.J.C. et al., 2015. Meta-analysis of the heritability of human traits based on fifty years of twin studies. *Nature Genetics*, 47(7), pp.702–709.

Price, A.L. et al., 2010. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7), pp.459–463. Available at: http://www.nature.com/nrg/journal/v11/n7/abs/nrg2813.html.

Price, A.L. et al., 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genetics*, 38(8), pp.904–909. Available at: http://www.nature.com/doifinder/10.1038/ng1847.

Purcell, S. et al., 2007. PLINK: A Tool Set for Whole-Genome Association and

Population-Based Linkage Analyses. *The American Journal of Human Genetics*, 81(3), pp.559–575. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524.

Purcell, S., 2009. PLINK: Whole genome data analysis toolset. Available at: http://zzz.bwh.harvard.edu/plink/ [Accessed September 11, 2018].

Purcell, S. & Chang, C., 2018. PLINK 1.9. *PLINK 1.9*. Available at: https://www.cog-genomics.org/plink2 [Accessed September 11, 2018].

R Development Core Team, 2008. R: A language and environment for statistical computing. Available at: http://mirror.fcaglp.unlp.edu.ar/CRAN/doc/manuals/r-patched/R-lang.pdf.

Rabbani, B., Tekin, M. & Mahdieh, N., 2014. The promise of whole-exome sequencing in medical genetics. *Journal of human genetics*, 59(1), pp.5–15. Available at: http://dx.doi.org/10.1038/jhg.2013.114%5Cnhttp://www.ncbi.nlm.nih.gov/pubmed/24196381.

Ramachandrappa, S. & Farooqi, I.S., 2011. Genetic approaches to understanding human obesity. *Journal of Clinical Investigation*, 121(6), pp.2080–2086. Available at: http://www.jci.org/articles/view/46044.

Ravi, D. et al., 2017. Deep Learning for Health Informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1), pp.4–21. Available at: http://ieeexplore.ieee.org/document/7801947/.

Reed, E. et al., 2015. A guide to genome-wide association analysis and post-analytic interrogation. *Statistics in Medicine*, 34(28), pp.3769–3792. Available at: http://doi.wiley.com/10.1002/sim.6605.

Reich, D.E. & Lander, E.S., 2001. On the allelic spectrum of human disease. *Trends in Genetics*, 17(9), pp.502–510. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168952501024106.

Renehan, A.G. et al., 2008. Body-mass index and incidence of cancer: a systematic review and meta-analysis of prospective observational studies. *The Lancet*, 371(9612), pp.569–578. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18280327.

Rentería, M.E., Cortes, A. & Medland, S.E., 2013. Using PLINK for Genome-Wide Association Studies (GWAS) and Data Analysis. In C. Gondro, J. van der Werf, & B. Hayes, eds. *Methods in molecular biology (Clifton, N.J.)*. Methods in Molecular Biology. Totowa, NJ: Humana Press, pp. 193–213. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23756904.

de Ridder, D., de Ridder, J. & Reinders, M.J.T., 2013. Pattern recognition in bioinformatics. *Briefings in bioinformatics*, 14(5), pp.633–47. Available at: https://academic.oup.com/bib/article-abstract/14/5/633/218439.

Ritchie, H. & Roser, M., 2019. Obesity & BMI. *Our World in Data*. Available at: https://ourworldindata.org/obesity [Accessed February 14, 2019].

Ritchie, M.D., 2015. Finding the Epistasis Needles in the Genome-Wide Haystack. In J. H. Moore & S. M. Williams, eds. *Epistasis. Methods in Molecular Biology*. Springer New York, pp. 19–33. Available at: http://link.springer.com/10.1007/978-1-4939-2155-3_2.

Ritchie, M.D. et al., 2001. Multifactor-Dimensionality Reduction Reveals High-Order Interactions among Estrogen-Metabolism Genes in Sporadic Breast Cancer. *The American Journal of Human Genetics*, 69(1), pp.138–147. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1226028/pdf/AJHGv69p138.pdf.

Ritchie, M.D. & Van Steen, K., 2018. The search for gene-gene interactions in genome-wide association studies: challenges in abundance of methods, practical considerations, and biological interpretation. *Annals of Translational Medicine*, 6(8), pp.157–157. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29862246.

Rudin, C. et al., 2014. Discovery with Data: Leveraging Statistics with Computer Science to Transform Science and Society. , pp.1–26.

Rumelhart, D.E., Hinton, G.E. & Williams, R.J., 1986. Learning representations by back-propagating errors. *Nature*, 323(6088), pp.533–536. Available at: http://www.nature.com/doifinder/10.1038/323533a0.

Saeys, Y., Inza, I. & Larranaga, P., 2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19), pp.2507–2517. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btm344.

Salari, N. et al., 2014. A Novel Hybrid Classification Model of Genetic Algorithms, Modified k-Nearest Neighbor and Developed Backpropagation Neural Network S. Gómez, ed. *PLoS ONE*, 9(11), p.e112987. Available at: http://dx.plos.org/10.1371/journal.pone.0112987.

Samish, I., Bourne, P.E. & Najmanovich, R.J., 2014. Achievements and challenges in structural bioinformatics and computational biophysics. *Bioinformatics*, 31(1), pp.146–150. Available at: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu769.

Sánchez-Pla, A., 2014. DNA Microarrays Technology. In pp. 1–23. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780444626516000015.

Sarzynski, M.A. et al., 2015. Genomic and transcriptomic predictors of triglyceride response to regular exercise. *British Journal of Sports Medicine*, 49(23), pp.1524–1531. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26491034.

Scarborough, P. et al., 2011. The economic burden of ill health due to diet, physical inactivity, smoking, alcohol and obesity in the UK: an update to 2006-07 NHS costs. *Journal of public health (Oxford, England)*, 33(4),

pp.527–35.

Schadt, E.E., Turner, S. & Kasarskis, A., 2010. A window into third-generation sequencing. *Human Molecular Genetics*, 19(R2), pp.R227–R240. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20858600.

Scherag, A. et al., 2010. Two new Loci for body-weight regulation identified in a joint analysis of genome-wide association studies for early-onset extreme obesity in French and german study groups. *PLoS genetics*, 6(4), p.e1000916. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20421936.

Schloss, J.A., 2008. How to get genomes at one ten-thousandth the cost. *Nature Biotechnology*, 26(10), pp.1113–1115. Available at: http://www.nature.com/articles/nbt1008-1113.

Schneider, V. & Church, D., 2013. Genome Reference Consortium. In Bethesda (MD), ed. *The NCBI Handbook*. National Center for Biotechnology Information (US). Available at: https://www.ncbi.nlm.nih.gov/books/NBK153600/pdf/Bookshelf_NBK153600.pdf.

Schork, N.J. et al., 2009. Common vs. rare allele hypotheses for complex diseases. *Current Opinion in Genetics & Development*, 19(3), pp.212–219. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2914559/pdf/nihms214356.pdf.

Schousboe, K. et al., 2003. Sex Differences in Heritability of BMI: A Comparative Study of Results from Twin Studies in Eight Countries. *Twin Research*, 6(5), pp.409–421. Available at: http://dx.doi.org/10.1375/136905203770326411.

Schwarze, K. et al., 2018. Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*, 20(10), pp.1122–1130. Available at:

https://www.nature.com/articles/gim2017247.pdf.

Scuteri, A. et al., 2007. Genome-Wide Association Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genetics*, 3(7), p.e115. Available at: http://dx.plos.org/10.1371/journal.pgen.0030115.

Sebastiani, P. & Solovieff, N., 2010a. Genome Wide Association Studies. In L. S. Heath & N. Ramakrishnan, eds. *Problem Solving Handbook in Computational Biology and Bioinformatics*. Boston, MA: Springer US, pp. 159–175. Available at: http://adsabs.harvard.edu/abs/2011pshi.book.....H.

Sebastiani, P. & Solovieff, N., 2010b. Genome Wide Association Studies. In L. S. Heath & N. Ramakrishnan, eds. *Problem Solving Handbook in Computational Biology and Bioinformatics*. Boston, MA: Springer US, pp. 159–175. Available at: http://www.springerlink.com/index/10.1007/978-0-387-09760-2.

Sedgwick, P., 2014. Multiple hypothesis testing and Bonferroni's correction. *BMJ*, 349(oct20 3), pp.g6284–g6284. Available at: http://www.bmj.com/permissionsSubscribe:http://www.bmj.com/subscribeBMJ2014;349:g6284doi:10.1136/bmj.g6284.

Segula, D., 2014. Complications of obesity in adults: a short review of the literature. *Malawi medical journal : the journal of Medical Association of Malawi*, 26(1), pp.20–4. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24959321.

Serra-Juhé, C. et al., 2017. Novel genes involved in severe early-onset obesity revealed by rare copy number and sequence variants X. Estivill, ed. *PLOS Genetics*, 13(5), p.e1006657. Available at: http://dx.plos.org/10.1371/journal.pgen.1006657.

Shawky, R.M. & Sadik, D.I., 2012. Genetics of obesity. *The Egyptian Journal of Medical Human Genetics*, 13(1), pp.11–17. Available at: http://dx.doi.org/10.1016/j.ejmhg.2011.08.005.

Sheehan, S. & Song, Y.S., 2016. Deep Learning for Population Genetic Inference K. Chen, ed. *PLOS Computational Biology*, 12(3), p.e1004845. Available at: http://biorxiv.org/content/early/2015/10/02/028175.abstract.

Shendure, J. & Ji, H., 2008. Next-generation DNA sequencing. *Nature Biotechnology*, 26(10), pp.1135–1145. Available at: http://www.nature.com/doifinder/10.1038/nbt1486.

Shi, M. & Weinberg, C.R., 2011. How Much Are We Missing in SNP-by-SNP Analyses of Genome-wide Association Studies? *Epidemiology*, p.1. Available at: www.niehs.nih.gov/research/atniehs/labs/bb/staff/weinberg/index.cfm#downloads.

Shu, L. et al., 2016. Mergeomics: multidimensional data integration to identify pathogenic perturbations to biological systems. *BMC Genomics*, 17(1), p.874. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27814671.

Sidiropoulos, K. et al., 2017. Reactome enhanced pathway visualization J. Kelso, ed. *Bioinformatics*, 33(21), pp.3461–3467. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29077811.

Smith, K.B. & Smith, M.S., 2016. Obesity Statistics. *Primary Care: Clinics in Office Practice*, 43(1), pp.121–135. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26896205.

Snyder, E.E. et al., 2004. The Human Obesity Gene Map: The 2003 Update. *Obesity Research*, 12(3), pp.369–439. Available at: http://doi.wiley.com/10.1038/oby.2004.47.

Speakman, J.R., 2007. A nonadaptive scenario explaining the genetic predisposition to obesity: the "predation release" hypothesis. *Cell metabolism*, 6(1), pp.5–12. Available at: https://www.cell.com/cell-metabolism/pdf/S1550-4131(07)00160-X.pdf.

Speakman, J.R., 2015. The 'Fat Mass and Obesity Related' (FTO) gene: Mechanisms of Impact on Obesity and Energy Balance. *Current Obesity*

Reports, 4(1), pp.73–91. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26627093.

Speliotes, E.K. et al., 2010. Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genetics*, 42(11), pp.937–948. Available at: http://www.nature.com/doifinder/10.1038/ng.686.

Spencer, C.C.A. et al., 2009. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip J. D. Storey, ed. *PLoS Genetics*, 5(5), p.e1000477. Available at: www.plosgenetics.org.

Srivastava, N. et al., 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), pp.1929–1958. Available at: http://jmlr.org/papers/volume15/srivastava14a.old/srivastava14a.pdf.

Statistics Team NHS Digital, 2017. *Statistics on Obesity, Physical Activity and Diet. England: 2017*, UK. Available at: https://digital.nhs.uk/data-and-information/publications/statistical/statistics-on-obesity-physical-activity-and-diet/statistics-on-obesity-physical-activity-and-diet-england-2017.

Strogatz, S.H., 2001. Exploring complex networks. *Nature*, 410(6825), pp.268–276. Available at: http://dx.doi.org/10.1038/35065725.

Stunkard, A.J. et al., 1986. An adoption study of human obesity. *The New England journal of medicine*, 314(4), pp.193–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/3941707.

Stunkard, A.J. et al., 1990. The Body-Mass Index of Twins Who Have Been Reared Apart. *New England Journal of Medicine*, 322(21), pp.1483–1487. Available at: http://www.ncbi.nlm.nih.gov/pubmed/2336075.

Su, P., 2013. Direct-to-consumer genetic testing: a comprehensive view. *The Yale journal of biology and medicine*, 86(3), pp.359–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24058310.

Subramanian, A. et al., 2005. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43), pp.15545–15550. Available at: http://www.ncbi.nlm.nih.gov/pubmed/16199517.

Sung, W.K., 2012. Bioinformatics applications in genomics. *Computer*, 45(6), pp.57–63. Available at: http://www.cs.wustl.edu/~schmidt/PDF/GEI.pdf.

Tan, P.-N. & Kumar, V., 2000. Interestingness Measures for Association Patters: A Perspective. *KDD Workshop on Postprocessing in Machine Learning and Data Mining*, 6(0), pp.1–9. Available at: http://cs.fit.edu/~pkc/ml/related/tan-postkdd00.pdf.

Tan, P.-N., Steinbach, M. & Kumar, V., 2005. Chapter 6. Association Analysis: Basic Concepts and Algorithms. In *Introduction to Data mining*. Boston: Addison-Wesley Longman Publishing Co., pp. 327–414. Available at: http://www-users.cs.umn.edu/~kumar/dmbook/index.php.

Teo, Y.Y., 2008. Common statistical issues in genome-wide association studies: a review on power, data quality control, genotype calling and population structure. *Current Opinion in Lipidology*, 19(2), pp.133–143. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=18388693.

Terada, A. et al., 2016. LAMPLINK: detection of statistically significant SNP combinations from GWAS data. *Bioinformatics (Oxford, England)*, 32(22), pp.3513–3515. Available at: http://bioinformatics.oxfordjournals.org/lookup/doi/10.1093/bioinformatics/btw418.

Terada, A. et al., 2013. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences*, 110(32), pp.12996–13001. Available at: http://www.pnas.org/cgi/doi/10.1073/pnas.1302233110.

The NHS Information Centre Lifestyle Statistics, 2011. *Statistics on obesity* ,

physical activity and diet: England, 2011,

Thorleifsson, G. et al., 2009. Genome-wide association yields new sequence variants at seven loci that associate with measures of obesity. *Nature Genetics*, 41(1), pp.18–24. Available at: http://www.nature.com/articles/ng.274.

Tryka, K.A. et al., 2014. NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Research*, 42(D1), pp.D975–D979. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1211.

Turner, S. et al., 2011. Quality Control Procedures for Genome-Wide Association Studies. *Current Protocols in Human Genetics*, 68(1), p.1.19.1-1.19.18. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21234875.

Turner, S.D., 2014. *qqman: an R package for visualizing GWAS results using Q-Q and manhattan plots*, Available at: http://linkinghub.elsevier.com/retrieve/pii/S0002929707613524.

U.S. Department of Health and Human Services, 1998. *Clinical Guidelines on the Identification, Evaluation, and Treatment of Overweight and Obesity in Adults: The Evidence Report.*, Bethesda (MD). Available at: https://hearttruth.gov/health/public/heart/obesity/wecan/portion/documents/CORESET1.pdf.

Ulitsky, I. & Bartel, D.P., 2013. lincRNAs: genomics, evolution, and mechanisms. *Cell*, 154(1), pp.26–46. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3924787/pdf/nihms-501071.pdf.

Uppu, S., Krishna, A. & Gopalan, R., 2017. A review on methods for detecting SNP interactions in high-dimensional genomic data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 5963(c), pp.1–1. Available at: http://www.ncbi.nlm.nih.gov/pubmed/27925593%5Cnhttp://ieeexplore.ie

ee.org/document/7765022/.

Valavanis, I.K. et al., 2010. A multifactorial analysis of obesity as CVD risk factor: use of neural network based methods in a nutrigenetics context. *BMC bioinformatics*, 11, p.453.

Vallgårda, S. et al., 2015. Backward- and forward-looking responsibility for obesity: policies from WHO, the EU and England. *The European Journal of Public Health*, 25(5), pp.845–848. Available at: http://eurpub.oxfordjournals.org/cgi/doi/10.1093/eurpub/ckv076.

Value, T.H.E. et al., 2016. The benefits of body mass index and waist circumference in the assessment of health risk. *ACSM's Health & Fitness Journal*, 20(4), pp.15–20.

Vella, F., 2008. *Biochemistry of Lipids, Lipoproteins and Membranes* Fifth., Elsevier. Available at: https://linkinghub.elsevier.com/retrieve/pii/B9780444532190X50016.

Venter, J.C. et al., 2001. The Sequence of the Human Genome. *Science*, 291(5507), pp.1304–1351. Available at: http://science.sciencemag.org/.

Vimaleswaran, K.S. & Loos, R.J.F., 2010. Progress in the genetics of common obesity and type 2 diabetes. *Expert Reviews in Molecular Medicine*, 12(February), p.e7. Available at: http://www.journals.cambridge.org/abstract_S1462399410001389.

Visscher, P.M. et al., 2012. Five Years of GWAS Discovery. *The American Journal of Human Genetics*, 90(1), pp.7–24. Available at: http://dx.doi.org/10.1016/j.ajhg.2011.11.029.

Vogler, G.P. et al., 1995. Influences of genes and shared family environment on adult body mass index assessed in an adoption study by a comprehensive path model. *International journal of obesity and related metabolic disorders : journal of the International Association for the Study of Obesity*, 19(1), pp.40–5. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7719389.

Walley, A.J., Blakemore, A.I.F. & Froguel, P., 2006. Genetics of obesity and the prediction of risk for health. *Human Molecular Genetics*, 15(SUPPL. 2), pp.124–130.

Wang, K. et al., 2011. A Genome-Wide Association Study on Obesity and Obesity-Related Traits Z. Zhao, ed. *PLoS ONE*, 6(4), p.e18939. Available at: http://dx.plos.org/10.1371/journal.pone.0018939.

Wang, R. et al., 2016. *Application of Clinical Bioinformatics* X. Wang et al., eds., Dordrecht: Springer Netherlands. Available at: http://link.springer.com/10.1007/978-94-017-7543-4.

Wang, Y. & Lobstein, T., 2006. Worldwide trends in childhood overweight and obesity. *International Journal of Pediatric Obesity*, 1(1), pp.11–25. Available at: http://informahealthcare.com/doi/abs/10.1080/17477160600586747.

Wardle, J. et al., 2008. Evidence for a strong genetic influence on childhood adiposity despite the force of the obesogenic environment. *American Journal of Clinical Nutrition*, 87(2), pp.398–404.

Watanabe, K. et al., 2017. Functional mapping and annotation of genetic associations with FUMA. *Nature communications*, 8(1), p.1826. Available at: http://www.ncbi.nlm.nih.gov/pubmed/29184056%0Ahttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC5705698.

Waterland, R.A. & Michels, K.B., 2007. Epigenetic Epidemiology of the Developmental Origins Hypothesis. *Annual Review of Nutrition*, 27(1), pp.363–388. Available at: http://nutr.annualreviews.org.

Watson, J.D. et al., 2014. *Molecular Biology of the Gene* Seventh Ed. Pearson, ed., New York: Cold Spring Harbor Laboratory Press.

Weale, M.E., 2010. Quality Control for Genome-Wide Association Studies. In *Genetic Variation: Methods and Protocols, Methods in Molecular Biology*. pp. 341–372. Available at:

http://www.springerlink.com/index/10.1007/978-1-60327-367-1.

Wei, W.-H. et al., 2012. Genome-wide analysis of epistasis in body mass index using multiple human populations. *European Journal of Human Genetics*, 20(8), pp.857–862. Available at: http://www.nature.com/doifinder/10.1038/ejhg.2012.17.

Wei, W.-H., Hemani, G. & Haley, C.S., 2014. Detecting epistasis in human complex traits. *Nature Reviews Genetics*, 15(11), pp.722–733. Available at: http://dx.doi.org/10.1038/nrg3747.

Welin, M. & Nordlund, P., 2010. Understanding specificity in metabolic pathways—Structural biology of human nucleotide metabolism. *Biochemical and Biophysical Research Communications*, 396(1), pp.157–163. Available at: https://linkinghub.elsevier.com/retrieve/pii/S0006291X10007230.

Welter, D. et al., 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Research*, 42(D1), pp.D1001–D1006. Available at: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1229.

Wen, X., Pique-Regi, R. & Luca, F., 2017. Integrating molecular QTL data into genome-wide genetic association analysis: Probabilistic assessment of enrichment and colocalization B. Li, ed. *PLOS Genetics*, 13(3), p.e1006646. Available at: https://dx.plos.org/10.1371/journal.pgen.1006646.

Westra, H.-J. & Franke, L., 2014. From genome to function by studying eQTLs. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, 1842(10), pp.1896–1902. Available at: http://dx.doi.org/10.1016/j.bbadis.2014.04.024.

Wetterstrand KA, 2018. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: https://www.genome.gov/27541954/dna-sequencing-costs-data/.

Whitaker, R.C. et al., 1997. Predicting Obesity in Young Adulthood from Childhood and Parental Obesity. *New England Journal of Medicine*, 337(13), pp.869–873. Available at: http://www.nejm.org/doi/abs/10.1056/NEJM199709253371301.

WHO, 2009. Global Health Risks: Mortality and burden of disease attributable to selected major risks. *Bulletin of the World Health Organization*, 87, pp.646–646.

Willer, C.J. et al., 2009. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics*, 41(1), pp.25–34. Available at: http://www.nature.com/doifinder/10.1038/ng.287.

Williams, M., 2011a. *Inositol phosphate metabolism R-HSA-1483249*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-1483249.

Williams, M., 2011b. *Phospholipid metabolism R-HSA-1483257*, Reactome, release 68. Available at: https://reactome.org/content/detail/R-HSA-1483257.

Withrow, D. & Alter, D.A., 2011. The economic burden of obesity worldwide: a systematic review of the direct costs of obesity. *Obesity Reviews*, 12(2), pp.131–141. Available at: http://doi.wiley.com/10.1111/j.1467-789X.2009.00712.x.

Wittke-Thompson, J.K., Pluzhnikov, A. & Cox, N.J., 2005. Rational Inferences about Departures from Hardy-Weinberg Equilibrium. *The American Journal of Human Genetics*, 76(6), pp.967–986. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1196455/pdf/AJHGv76p967.pdf.

Wolkenhauer, O. et al., 2013. The road from systems biology to systems medicine. *Pediatric Research*, 73(4–2), pp.502–507. Available at: https://www.nature.com/articles/pr20134.pdf.

Wolkenhauer, O., 2013. The role of theory and modeling in medical research.

Frontiers in Physiology, 4, p.377. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24391594.

World Health Organization. Department for Prevention of Noncommunicable Diseases, 2017. *Noncommunicable Diseases Progress Monitor, 2017*, Geneva. Available at: http://apps.who.int/iris/bitstream/10665/258940/1/9789241513029-eng.pdf?ua=1.

World Health Organization, 2014. *Global Status Report On Noncommunicable Diseases 2014*, Geneva: WHO Library Cataloguing-in-Publication Data. Available at: http://apps.who.int/iris/bitstream/10665/148114/1/9789241564854_eng.pdf?ua=1.

World Health Organization, 2018. Obesity and overweight. *Fact sheets*. Available at: http://www.who.int/news-room/fact-sheets/detail/obesity-and-overweight [Accessed August 18, 2018].

Wu, C. et al., 2011. A Comparison of Association Methods Correcting for Population Stratification in Case-Control Studies. *Annals of Human Genetics*, 75(3), pp.418–427. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3215268/pdf/nihms333160.pdf.

Xi, B. et al., 2011. Influence of physical inactivity on associations between single nucleotide polymorphisms and genetic predisposition to childhood obesity. *American Journal of Epidemiology*, 173(11), pp.1256–1262.

Yang, W., Kelly, T. & He, J., 2007. Genetic epidemiology of obesity. *Epidemiologic Reviews*, 29(1), pp.49–61.

Yoon, S. et al., 2018. Efficient pathway enrichment and network analysis of GWAS summary data using GSA-SNP2. *Nucleic Acids Research*, 46(10), pp.e60–e60. Available at: https://academic.oup.com/nar/article-abstract/46/10/e60/4942469.

Zaitlen, N. et al., 2013. Using Extended Genealogy to Estimate Components of Heritability for 23 Quantitative and Dichotomous Traits P. M. Visscher, ed. *PLoS Genetics*, 9(5), p.e1003520. Available at: http://dx.plos.org/10.1371/journal.pgen.1003520.

Zaki, M.J., 2000. Generating non-redundant association rules. In *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '00*. New York, New York, USA: ACM Press, pp. 34–43. Available at: http://portal.acm.org/citation.cfm?doid=347090.347101.

Zeiler, M.D., 2012. ADADELTA: An Adaptive Learning Rate Method. *arXiv:1212.5701*. Available at: http://arxiv.org/abs/1212.5701.

Zeng, P. et al., 2015. Statistical analysis for genome-wide association study. *Journal of biomedical research*, 29(4), pp.285–97. Available at: http://www.ncbi.nlm.nih.gov/pubmed/26243515.

Zhang, J. et al., 2011. The impact of next-generation sequencing on genomics. *Journal of Genetics and Genomics*, 38(3), pp.95–109. Available at: http://www.ncbi.nlm.nih.gov/pubmed/21477781.

Zhang, K. et al., 2010. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Research*, 38(Web Server), pp.W90–W95. Available at: http://www.ncbi.nlm.nih.gov/pubmed/20435672.

Zhang, Q., Long, Q. & Ott, J., 2014. AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects A. Rzhetsky, ed. *PLoS Computational Biology*, 10(6), p.e1003627. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24901472.

Zhang, Y.-B. et al., 2016. Genome-wide association study identifies multiple susceptibility loci for craniofacial microsomia. *Nature Communications*,

7, p.10605. Available at: http://www.nature.com/doifinder/10.1038/ncomms10605.

Zhou, J. & Troyanskaya, O.G., 2015. Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, 12(10), pp.931–934. Available at: http://www.nature.com/reprints/index.html.

Zhou, T., Yao, J. & Liu, Z., 2017. Gene Ontology, Enrichment Analysis, and Pathway Analysis. In *Bioinformatics in Aquaculture*. Chichester, UK: John Wiley & Sons, Ltd, pp. 150–168. Available at: http://doi.wiley.com/10.1002/9781118782392.ch10.

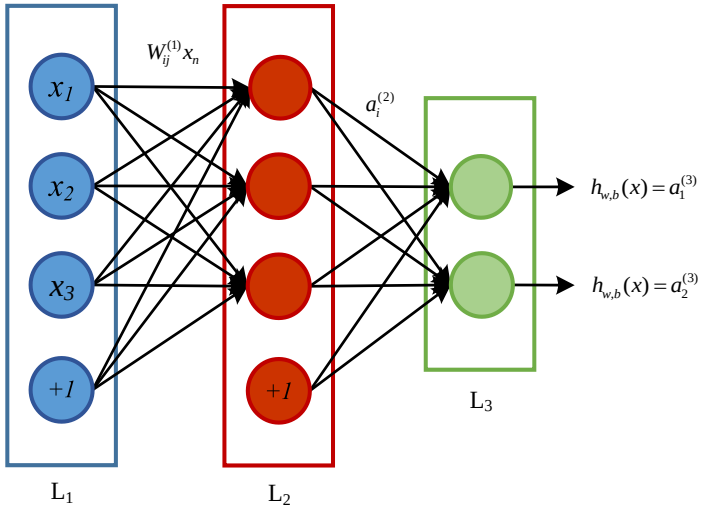
Zhou, W. et al., 2018. Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies. *Nature Genetics*, 50(9), pp.1335–1341. Available at: http://www.nature.com/articles/s41588-018-0184-y.

Zhu, J. et al., 2014. Associations of Genetic Risk Score with Obesity and Related Traits and the Modifying Effect of Physical Activity in a Chinese Han Population Y.-H. Hsu, ed. *PLoS ONE*, 9(3), p.e91442. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24626232.

Ziegler, A., König, I.R. & Thompson, J.R., 2008. Biostatistical Aspects of Genome-Wide Association Studies. *Biometrical Journal*, 50(1), pp.8–28. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18217698.

Appendix A: MLP activation calculation example.

The following example represents the computation of the activation of the neurons in a simple NN with three layers and two output nodes. This is a small version of the architecture adopted in this thesis although it serves for explanatory purposes only.



Example of MLP network with an input layer L_1, one hidden layers L_2 and an output layer L_3 with two output units.

Following the definition of parameters summarised in Table 3-2, and given a fixed setting of parameters *W,b*, the neural network displayed above calculates the outputs $h_{W,b}(x)$ as follow:

For each neuron:

$$h_{W,b}(x) = f(W^T x) = f\left(\sum_{i=1}^{n} W_i x_i + b\right)$$

The activation $a_i^{(l)}$ in layer 1 (L_1) is: $a_i^{(1)} = x_i$.

The activations of the units in the hidden layer (L_2) are calculated as follow:

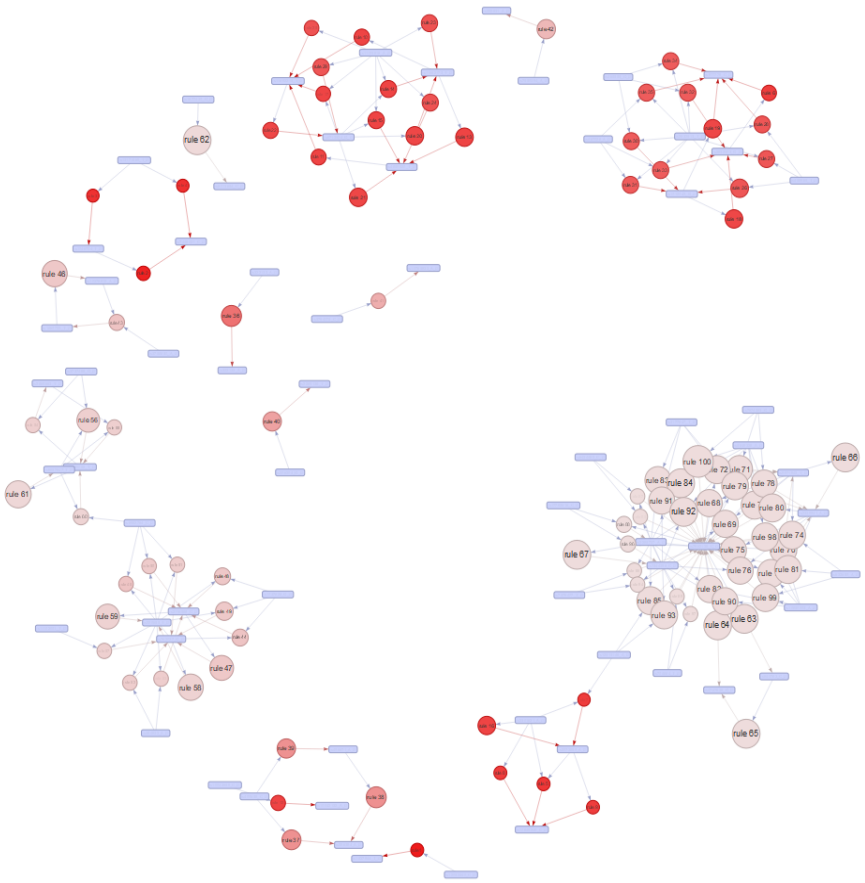
$$a_1^{(2)} = f(W_{11}^{(1)}x_1 + W_{12}^{(1)}x_2 + W_{13}^{(1)}x_3 + b_1^{(1)})$$
$$a_2^{(2)} = f(W_{21}^{(1)}x_1 + W_{22}^{(1)}x_2 + W_{23}^{(1)}x_3 + b_1^{(1)})$$
$$a_3^{(2)} = f(W_{31}^{(1)}x_1 + W_{32}^{(1)}x_2 + W_{33}^{(1)}x_3 + b_1^{(1)})$$

Once activations in the hidden layer are calculated, the outputs for above MLP can be computed:

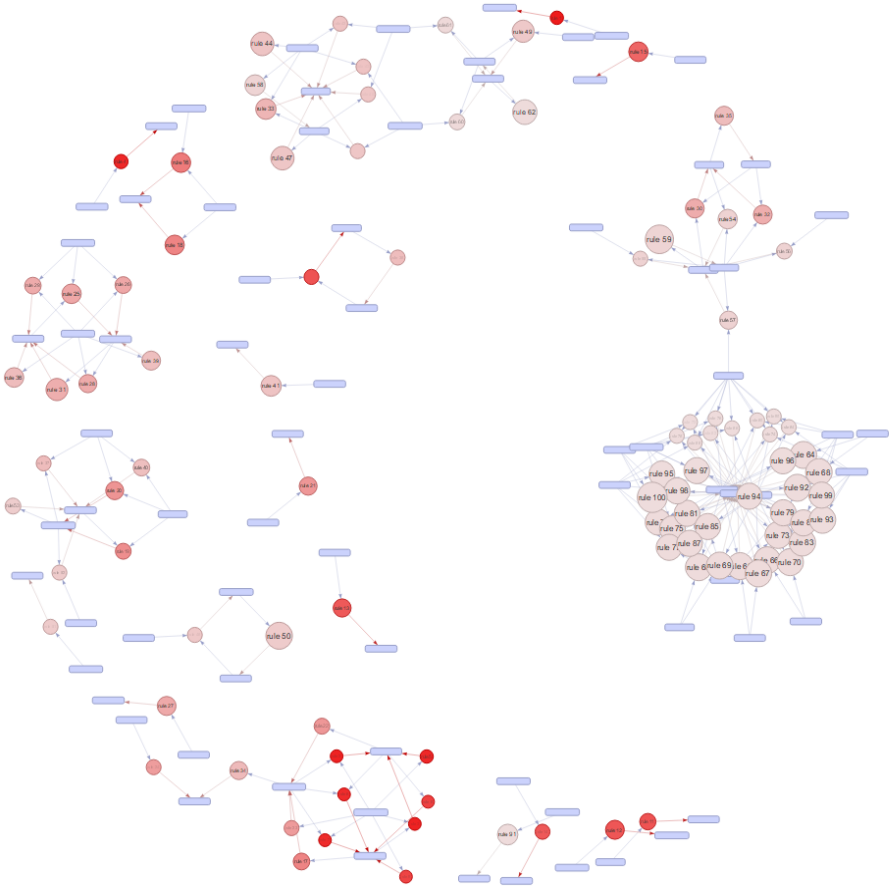
$$h_{w,b}(x) = a_1^{(3)} = f(W_{11}^{(2)}a_1 + W_{12}^{(2)}a_2 + W_{13}^{(2)}a_3 + b_1^{(2)})$$
$$h_{w,b}(x) = a_2^{(3)} = f(W_{21}^{(2)}a_1 + W_{22}^{(2)}a_2 + W_{23}^{(2)}a_3 + b_1^{(2)})$$

Appendix B: Rule network plot for 100 rules

In this Appendix, network plots for the top 100 rules are presented for cases and controls, which were derived from association analysis experiments conducted in this thesis. As the network grows, rules are more difficult to recognise from the clusters. Network plots were generated in R using the package *arulesViz*.



Network plot for 100 rules in case set.

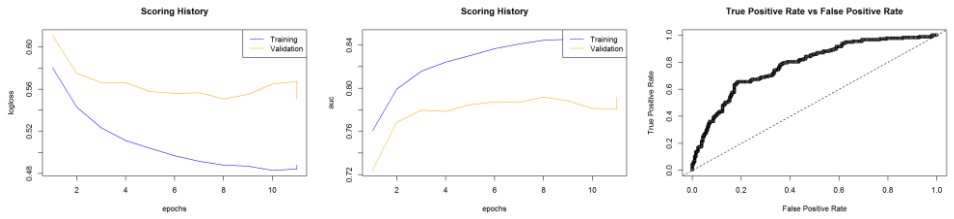


Network plot for 100 rules in control set.

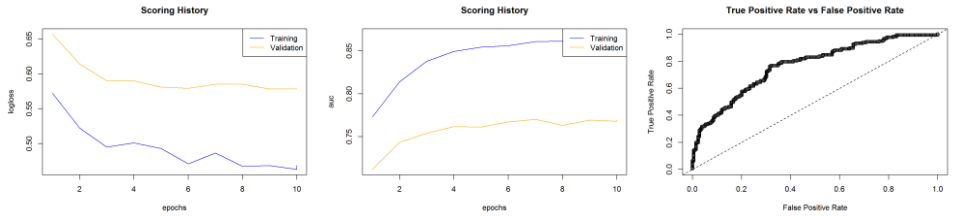
Appendix C: Performance plots SAERMA

Performance plots for final experiment using ARM and SAE for classification analysis using MLP.

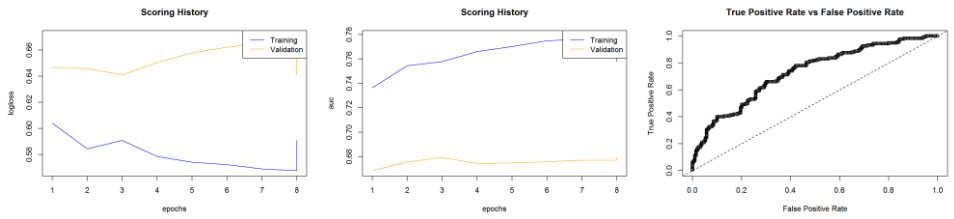
Plots corresponding to Table 4-20:



a) 150



b) 150-100



c) 150-100-50

Logloss, AUC and ROC curve plots for the top 300 rules. This included 204 SNPs which are compressed into: 150-100-50 units

Global parameters:

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	Yes
Stopping tolerance	0.01

Stopping rounds	4
Max model generated	200

Global tuning parameters for classification tasks

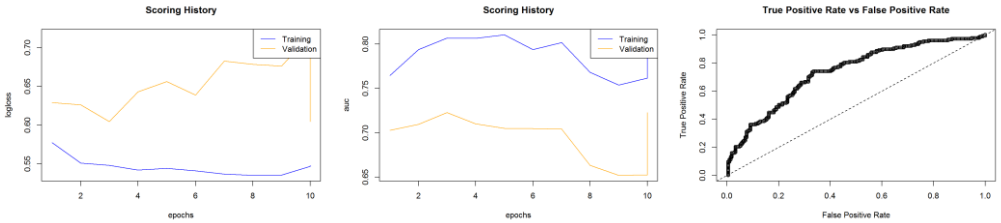
Next, specific parameters for each model:

Input	Parameter	Value
150	Activation	TanhWithDropout
	Hidden	2
	Neurons	10
	Epochs	50
	L1	$2.2x10^{-5}$
	L2	$3.5x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
150-100	Activation	MaxoutWithDropout
	Hidden	2
	Neurons	20
	Epochs	10
	L1	0.0
	L2	$1.0x10^{-4}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
150-100-50	Activation	TanhWithDropout
	Hidden	3
	Neurons	30
	Epochs	10
	L1	$3.2x10^{-5}$
	L2	$5.3x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5

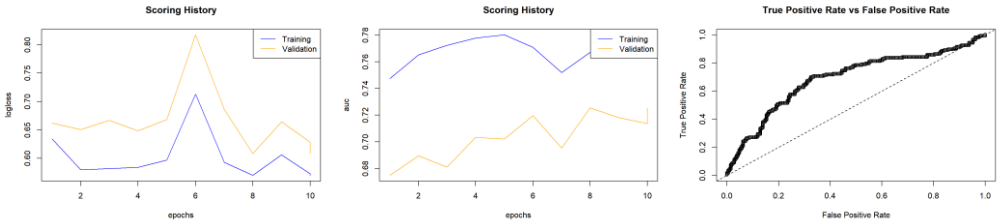
Model-specific tuning parameters

Based on empirical analysis, these configurations produced the best results.

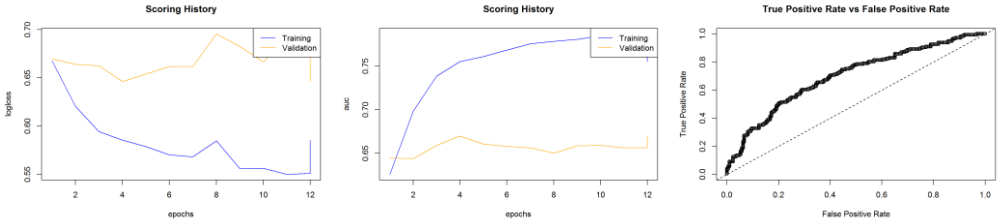
Plots corresponding to Table 4-21:



a) 125



b) 125-75



c) 125-75-50

Logloss, AUC and ROC curve plots for the top 200 rules. This included 161 SNPs which are compressed into: 125-75-50 units

Global parameters:

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	No
Stopping tolerance	-
Stopping rounds	-
Max model generated	200

Global tuning parameters for classification tasks

Next, specific parameters for each model:

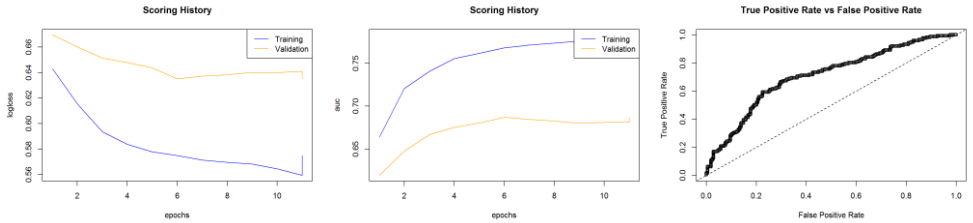
Input	Parameter	Value

125	Activation	TanhWithDropout
	Hidden	4
	Neurons	25
	Epochs	100
	L1	$6.4x10^{-5}$
	L2	$5.8x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5
125-75	Activation	TanhWithDropout
	Hidden	4
	Neurons	25
	Epochs	50
	L1	$4.3x10^{-5}$
	L2	$6.3x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
125-75-50	Activation	MaxoutWithDropout
	Hidden	2
	Neurons	10
	Epochs	100
	L1	0.0
	L2	$1.0x10^{-4}$
	Input_dropout_ratio	0.1
	Hidden_dropout_ratios	0.5

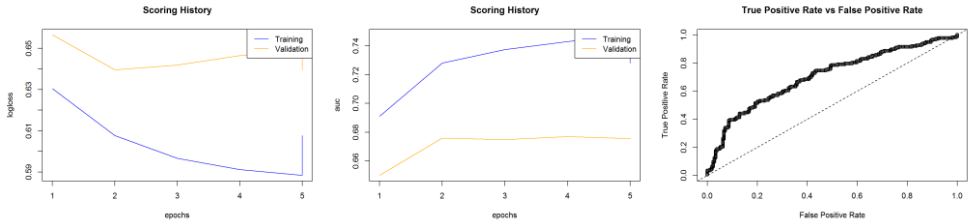
Model-specific tuning parameters

Based on empirical analysis, these configurations produced the best results.

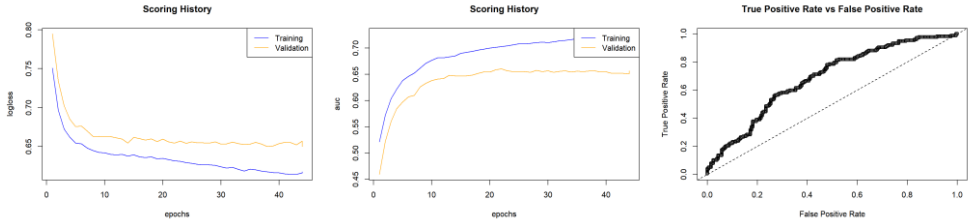
Plots corresponding to Table 4-22:



a) 90



b) 90-50



c) 90-50-25

Logloss, AUC and ROC curve plots for the top 100 rules. This included 124 SNPs which are compressed into: 90-50-25 units

Global parameters:

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	Yes
Stopping tolerance	0.01
Stopping rounds	4
Max model generated	200

Global tuning parameters for classification tasks

Next, specific parameters for each model:

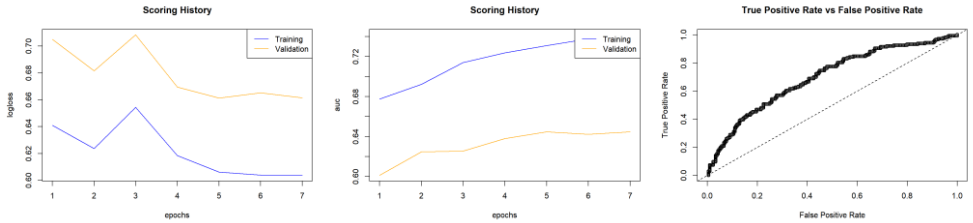
Input	Parameter	Value

90	Activation	MaxoutWithDropout
	Hidden	2
	Neurons	20
	Epochs	100
	L1	$5.2x10^{-5}$
	L2	$9.2x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
90-50	Activation	TanhWithDropout
	Hidden	2
	Neurons	20
	Epochs	10
	L1	$7.7x10^{-5}$
	L2	$4.1x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
90-50-25	Activation	RectifierWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$5.5x10^{-5}$
	L2	$6.7x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5

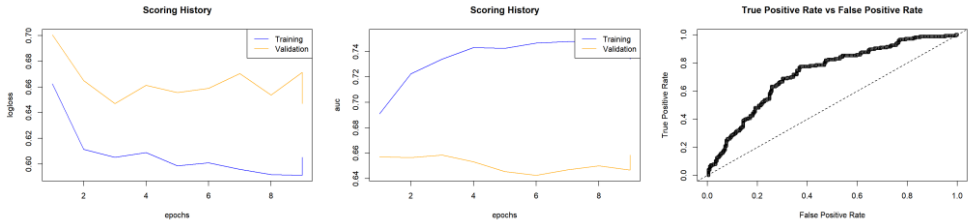
Model-specific tuning parameters

Based on empirical analysis, these configurations produced the best results.

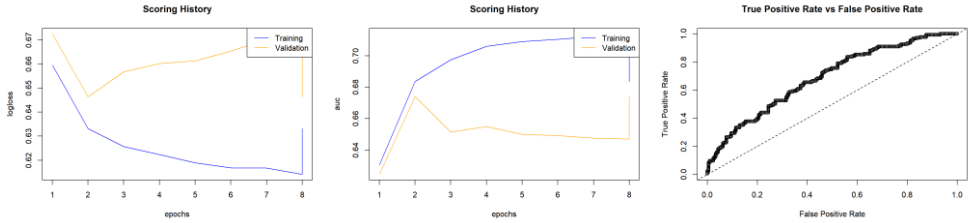
Plots corresponding to Table 4-23:



a) 75



b) 75-50



c) 75-50-25

Logloss, AUC and ROC curve plots for the top 50 rules. This included 92 SNPs which are compressed into: 75-50-25 units

Global parameters:

Global Parameters	
Parameter	**Value**
Adaptive learning	ADADELTA (rho = 0.99 and epsilon = $1x10^{-8}$)
Early stopping	Yes
Stopping tolerance	0.01
Stopping rounds	2
Max model generated	200

Global tuning parameters for classification tasks

Next, specific parameters for each model:

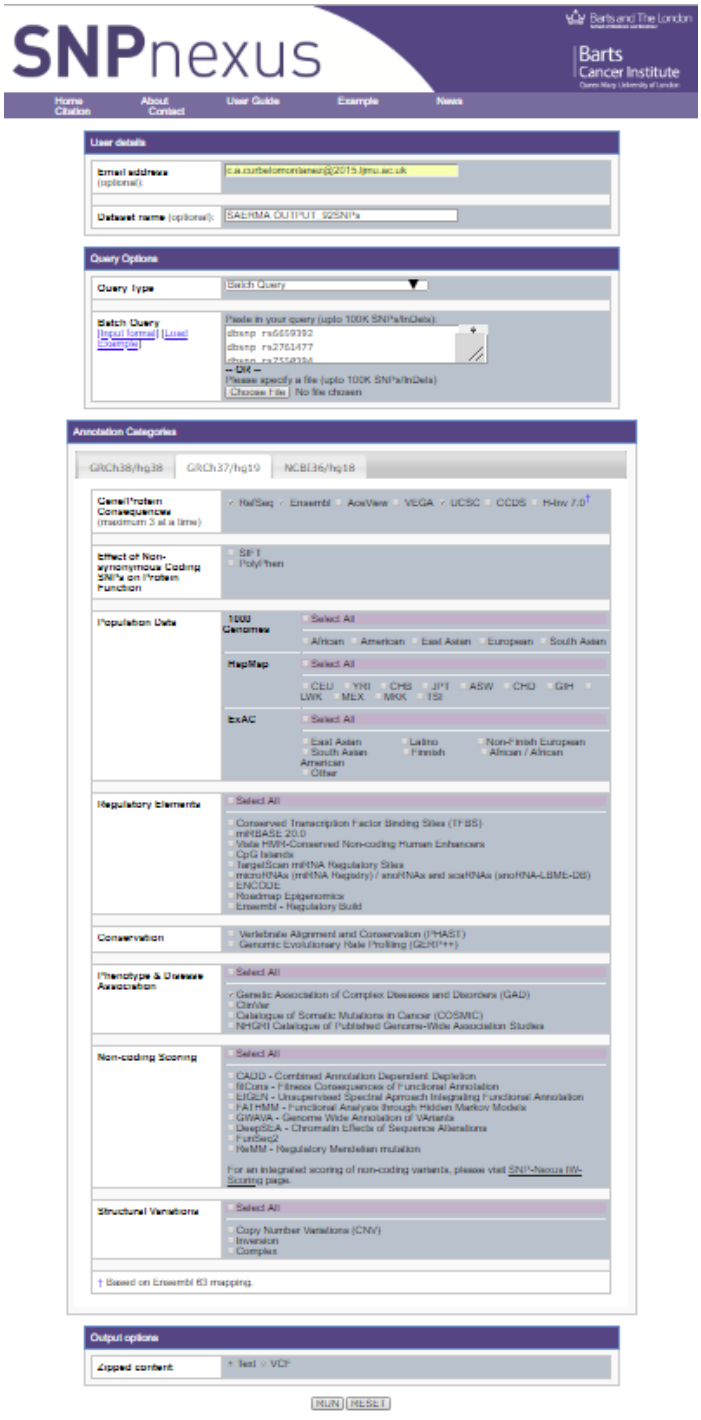
Input	Parameter	Value

90	Activation	MaxoutWithDropout
	Hidden	2
	Neurons	20
	Epochs	100
	L1	$5.2x10^{-5}$
	L2	$9.2x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
90-50	Activation	TanhWithDropout
	Hidden	2
	Neurons	20
	Epochs	10
	L1	$7.7x10^{-5}$
	L2	$4.1x10^{-5}$
	Input_dropout_ratio	0.05
	Hidden_dropout_ratios	0.5
90-50-25	Activation	RectifierWithDropout
	Hidden	2
	Neurons	20
	Epochs	50
	L1	$5.5x10^{-5}$
	L2	$6.7x10^{-5}$
	Input_dropout_ratio	0.0
	Hidden_dropout_ratios	0.5

Model-specific tuning parameters

Based on empirical analysis, these configurations produced the best results.

Appendix D: SNPnexus query output



SNPnexus main page screenshot

Source: http://www.snp-nexus.org/

The description for the output table provided below has been extracted from the SNPnexus web (Dayem Ullah et al. 2018):

- SNP: SNP ID as represented in genomic position format.
- Chromosome (C): Chromosome location for mapped variant.
- Position: Start position of the variant in the chromosome.
- Overlapped Gene (OG): Name of the gene to which the variant is overlapped.
- Type: Gene type (i.e. protein-coding, miRNA, non-coding, Pseudogene, snoRNA, lincRNA, etc).
- Annotation: Summary of whether the variant overlapped with the coding, intronic or untranslated regions of the various transcript isoforms of the gene, as annotated from Ensembl gene system.
- Nearest Upstream Gene (NUG). If a variant is not overlapped with any gene and, considering gene alignment on the positive strand as left to right, then NUG is the gene whose end position is nearest to the variant on the left.
- Type of NUG: Gene type (i.e. protein-coding, miRNA, non-coding, Pseudogene, snoRNA, lincRNA etc).
- Distance to NUG: distance from the end position of the nearest upstream gene.
- Nearest Downstream Gene (NDG): If a variant is not overlapped with any gene and considering gene alignment on the positive strand as left to right, then NDG is the gene whose start position is nearest to the variant on the right.
- Type of NDG: Gene type, e.g., protein-coding, miRNA, non-coding, Pseudogene, snoRNA, lincRNA etc.
- Distance to NDG: distance from the start position of the nearest downstream gene.

The table contains information on overlapped or nearest genes for the 204 SNPs comprising the top 300 rules identified applying the proposed algorithm SAERMA:

SNP	Chr	Pos	OG	Type	Annotation	NUG	Type of NUG	Distance to NUG	NDG	Type of NDG	Distance to NDG
rs11102001	1	110299691	GSTM5	protein_coding	non-coding intronic						
rs11102001	1	110299691	RP4-735C1.4	antisense	non-coding intronic						
rs11102001	1	110299691	EPS8L3	protein_coding	coding nonsyn,3downstream						
rs11802668	1	12000670	PLOD1	protein_coding	non-coding intronic,intronic						
rs6659392	1	178598873				RNA5SP69	rRNA	68709	RP5-1098D14.1	lincRNA	21968
rs2761477	1	178626526				RP5-1098D14.1	lincRNA	5109	MIR4424	miRNA	20358
rs7550394	1	199015253	RP11-16L9.4	lincRNA	non-coding intronic						
rs322931	1	199019855	RP11-16L9.4	lincRNA	non-coding intronic						
rs3900967	1	203583705				OPTC	protein_coding	105713	ATP2B4	protein_coding	11984
rs17015701	1	210337691				SYT14	protein_coding	55	SERTAD4-AS1	antisense	67110
rs11119426	1	210340969				SYT14	protein_coding	3333	SERTAD4-AS1	antisense	63832
rs3795308	1	23689083	ZNF436	protein_coding	coding syn						
rs9887921	1	23700811				C1orf213	protein_coding	2479	RP5-1057J7.7	lincRNA	4062
rs12074072	1	49054581	AGBL4	protein_coding	intronic						
rs11207744	1	61931461				NFIA	protein_coding	2996	AC099791.1	miRNA	163242
rs11692215	2	100746573	AFF3	protein_coding	intronic,non-coding intronic						
rs6713524	2	100747357	AFF3	protein_coding	intronic,non-coding intronic						
rs4850920	2	100757308	AFF3	protein_coding	intronic,non-coding intronic						
rs11682173	2	100759810				AFF3	protein_coding	609	LINC01104	lincRNA	64906
rs11123816	2	100783653				AFF3	protein_coding	24452	LINC01104	lincRNA	41063

rs6715763	2	118580670	DDX18	protein_coding	intronic,non-coding intronic,3downstream,5upstream						
rs13030497	2	118589029	DDX18	protein_coding	3utr,non-coding						
rs837870	2	130094203				AC079586.1	lincRNA	62782	snoU13	snoRNA	90792
rs6727787	2	131656585	ARHGEF4	protein_coding	non-coding intronic,intronic						
rs10204782	2	17114343				AC008069.2	pseudogene	77792	AC080094.1	lincRNA	36958
rs6434482	2	192423103				MYO1B	protein_coding	132988	NABP1	protein_coding	119691
rs12053340	2	201412611	SGOL2	protein_coding	intronic						
rs1527944	2	201456245	AOX1	protein_coding	intronic						
rs7579477	2	20486117	PUM2	protein_coding	intronic						
rs6717227	2	226013860				DOCK10	protein_coding	106698	AC067961.1	lincRNA	250566
rs728654	2	226030132				DOCK10	protein_coding	122970	AC067961.1	lincRNA	234294
rs11691744	2	235335472				AC097713.5	pseudogene	96990	AC097713.3	lincRNA	11500
rs11686781	2	235338531				AC097713.5	pseudogene	100049	AC097713.3	lincRNA	8441
rs12328617	2	238173270				AC112715.2	protein_coding	6951	COL6A3	protein_coding	59376
rs762027	2	32960091	TTC27	protein_coding	intronic,3downstream,5upstream						
rs2116588	2	32977741	TTC27	protein_coding	intronic						
rs220655	2	33013756	TTC27	protein_coding	intronic,3downstream						
rs1037626	2	33039690	TTC27	protein_coding	intronic,3downstream						
rs17293732	2	5602284				AC073143.1	pseudogene	148017	AC107057.1	lincRNA	87628
rs10180670	2	5607349				AC073143.1	pseudogene	153082	AC107057.1	lincRNA	82563
rs2080390	2	71058226	CD207	protein_coding	coding syn						
rs17662453	2	71061108	CD207	protein_coding	coding syn						

rs1452044	2	78945755				AC092660.1	lincRNA	119223	RNU6-827P	snRNA	163818
rs17015634	2	79001963				AC092660.1	lincRNA	175431	RNU6-827P	snRNA	107610
rs2154762	3	37663435	ITGA9	protein_coding	intronic						
rs4678980	3	37665385	ITGA9	protein_coding	intronic						
rs1479955	3	65228793				AC104331.1	miRNA	10135	MAGI1	protein_coding	110407
rs17349380	3	65231027				AC104331.1	miRNA	12369	MAGI1	protein_coding	108173
rs12486426	3	77197856	ROBO2	protein_coding	intronic						
rs425222	4	167232997				AC093874.1	miRNA	84013	Y_RNA	misc_RNA	1090
rs12506355	4	167265509				Y_RNA	misc_RNA	31323	RP11-217C7.1	lincRNA	44344
rs1080788	4	167300345				Y_RNA	misc_RNA	66159	RP11-217C7.1	lincRNA	9508
rs1492533	4	167310371	RP11-217C7.1	lincRNA	non-coding						
rs1565650	4	167321777	RP11-217C7.1	lincRNA	non-coding intronic						
rs1367555	4	188420909				RP11-91J3.1	pseudogene	55457	RP11-237D3.1	lincRNA	16398
rs1955311	4	30471257				RP11-174E22.2	pseudogene	461319	PCDH7	protein_coding	250780
rs6857847	4	89514572	HERC3	protein_coding	intronic,5utr						
rs17799056	4	89527323	HERC3	protein_coding	intronic						
rs2869663	4	89542281	HERC3	protein_coding	intronic						
rs3737488	4	89607905	HERC3	protein_coding	coding syn,coding *syn						
rs7448421	5	178303868	ZNF354B	protein_coding	intronic						
rs1046724	5	178315686				ZNF354B	protein_coding	563	ZFP2	protein_coding	7209
rs13171869	5	178335516	ZFP2	protein_coding	intronic						
rs2546440	5	180575250				OR2V1	protein_coding	22946	OR2V2	protein_coding	6693
rs16894413	5	65127219	NLN	protein_coding	intronic						
rs10043659	5	71781839	ZNF366	protein_coding	intronic						

rs10042132	5	71789021	ZNF366	protein_coding	intronic						
rs17335290	5	94200808	MCTP1	protein_coding	intronic,3downstream						
rs17792616	6	114309165	HDAC2	protein_coding	intronic						
rs17792616	6	114309165	RP3-399L15.3	antisense	non-coding intronic						
rs4897427	6	130810954				TMEM200A	protein_coding	46746	Y_RNA	misc_RNA	84187
rs2876086	6	130818014				TMEM200A	protein_coding	53806	Y_RNA	misc_RNA	77127
rs7765392	6	135005295				RP11-557H15.4	lincRNA	24321	RP1-287H17.1	lincRNA	21903
rs6569962	6	135011576				RP11-557H15.4	lincRNA	30602	RP1-287H17.1	lincRNA	15622
rs6914876	6	135014966				RP11-557H15.4	lincRNA	33992	RP1-287H17.1	lincRNA	12232
rs259404	6	147104187	ADGB	protein_coding	intronic,non-coding intronic						
rs1737319	6	163796510				RP3-495O10.1	pseudogene	12015	CAHM	lincRNA	37587
rs16894934	6	163805742				RP3-495O10.1	pseudogene	21247	CAHM	lincRNA	28355
rs4565296	6	16399647	ATXN1	protein_coding	intronic						
rs2206734	6	20694884	CDKAL1	protein_coding	intronic						
rs6935599	6	20717095	CDKAL1	protein_coding	intronic						
rs10947072	6	30372278				UBQLN1P1	pseudogene	40221	MICC	pseudogene	10214
rs6930977	6	30435288	TMPOP1	pseudogene	non-coding						
rs1627354	7	107677984	LAMB4	protein_coding	coding nonsyn,5upstream,non-coding						
rs7811376	7	111577642	DOCK4	protein_coding	intronic,non-coding intronic						
rs1860722	7	123037157				LYPLA1P1	pseudogene	166362	IQUB	protein_coding	55297
rs10278912	7	123037887				LYPLA1P1	pseudogene	167092	IQUB	protein_coding	54567
rs6958382	7	21498702	SP4	protein_coding	intronic						
rs2711098	7	24569427				NPY	protein_coding	237943	RNU6-1103P	snRNA	7524

rs2521766	7	24772686	DFNA5	protein_coding	intronic,non-coding intronic						
rs42695	7	28430860	CREB5	protein_coding	intronic						
rs2410518	8	17197888	MTMR7	protein_coding	intronic,non-coding intronic						
rs7386192	8	17209290	MTMR7	protein_coding	intronic						
rs7001002	8	21485754				AC022716.1	miRNA	78159	GFRA2	protein_coding	62161
rs2272641	8	23294761	ENTPD4	protein_coding	5upstream,coding nonsyn,non-coding						
rs13269315	8	4231556	CSMD1	protein_coding	intronic						
rs11786916	8	4232568	CSMD1	protein_coding	intronic						
rs6982764	8	85276246	RALYL	protein_coding	intronic						
rs484932	9	135241588				SETX	protein_coding	11216	TTF1	protein_coding	9420
rs11138445	9	82609762				RP11-403N16.3	lincRNA	102840	RP11-394O9.1	lincRNA	35732
rs10976907	9	8268343				RP11-29B9.2	lincRNA	307263	PTPRD	protein_coding	45903
rs10816047	9	9109843	PTPRD	protein_coding	intronic						
rs4750827	10	132547184				Y_RNA	misc_RNA	243719	MIR378C	miRNA	213667
rs12570210	10	132962633	TCERG1L	protein_coding	non-coding intronic,intronic						
rs3011642	10	22394843				DNAJC1	protein_coding	102145	ADIPOR1P1	pseudogene	57078
rs3011644	10	22397995				DNAJC1	protein_coding	105297	ADIPOR1P1	pseudogene	53926
rs10828296	10	22460830				ADIPOR1P1	pseudogene	6107	EBLN1	protein_coding	36913
rs1926690	10	22471741				ADIPOR1P1	pseudogene	17018	EBLN1	protein_coding	26002
rs11593316	10	22477498				ADIPOR1P1	pseudogene	22775	EBLN1	protein_coding	20245
rs6482203	10	22479936				ADIPOR1P1	pseudogene	25213	EBLN1	protein_coding	17807
rs10826675	10	29952872	SVIL	protein_coding	intronic,non-coding intronic						
rs9888055	10	29955649	SVIL	protein_coding	intronic,non-coding intronic						

rs7098565	10	29963887	SVIL	protein_coding	intronic,non-coding intronic						
rs10998726	10	71099222	HK1	protein_coding	intronic,non-coding intronic						
rs2042867	10	79722242				H2AFZP5	pseudogene	8608	POLR3A	protein_coding	12665
rs735638	10	79732721				H2AFZP5	pseudogene	19087	POLR3A	protein_coding	2186
rs4979935	10	79735473	POLR3A	protein_coding	3utr						
rs406671	10	90493534	LIPK	protein_coding	intronic						
rs396394	10	90500521	LIPK	protein_coding	intronic						
rs400659	10	90506027	LIPK	protein_coding	intronic						
rs10882256	10	95272834	CEP55	protein_coding	intronic						
rs3794075	11	12282908	MICAL2	protein_coding	intronic,3downstream,non-coding,non-coding intronic						
rs10765933	11	12283928	MICAL2	protein_coding	intronic,non-coding,non-coding intronic,3downstream						
rs10765933	11	12283928	RP11-265D17.2	antisense	non-coding intronic						
rs704664	11	44787201	TSPAN18	protein_coding	intronic						
rs3758650	11	616865	CDHR5	protein_coding	3utr,3downstream						
rs12280583	11	83325416	DLG2	protein_coding	intronic						
rs10501544	11	83334780	DLG2	protein_coding	intronic						
rs3809288	12	111652522	CUX2	protein_coding	intronic,non-coding intronic						
rs10083213	12	111654363	CUX2	protein_coding	intronic,non-coding intronic						
rs11065884	12	111818701	RP3-473L9.4	lincRNA	non-coding intronic						
rs10849949	12	111893537	ATXN2	protein_coding	intronic,5upstream,non-						

					coding,3utr,3downstream						
rs2073950	12	111894072	ATXN2	protein_coding	intronic,non-coding intronic,5upstream						
rs2301621	12	111895272	ATXN2	protein_coding	intronic,non-coding intronic,5upstream,non-coding						
rs1544396	12	112062875				RP11-686G8.2	antisense	24815	BRAP	protein_coding	17075
rs2285727	12	112093078	BRAP	protein_coding	intronic,non-coding intronic						
rs632650	12	112131698	ACAD10	protein_coding	3downstream,intronic,non-coding intronic						
rs640783	12	112168050	ACAD10	protein_coding	non-coding intronic,intronic,3downstream						
rs7962138	12	112180177	ACAD10	protein_coding	non-coding intronic,intronic						
rs4767939	12	112206895	RP11-162P23.2	protein_coding	intronic						
rs4767939	12	112206895	ALDH2	protein_coding	intronic						
rs4648328	12	112222788	RP11-162P23.2	protein_coding	intronic						
rs4648328	12	112222788	ALDH2	protein_coding	intronic						
rs9971942	12	112256042				RP3-462E2.3	lincRNA	4818	AC003029.1	pseudogene	21529
rs3177647	12	112277576	AC003029.1	pseudogene	non-coding						
rs3177647	12	112277576	MAPKAPK5-AS1	lincRNA	non-coding,3downstream						
rs12315146	12	112323016	MAPKAPK5	protein_coding	intronic,5upstream						
rs2339941	12	112490764	NAA25	protein_coding	intronic,non-coding intronic						
rs10850003	12	112571169	TRAFD1	protein_coding	intronic,non-coding intronic						

rs7974383	12	112794714	HECTD4	protein_coding	intronic						
rs11048320	12	26031743				RP11-443N24.3	lincRNA	44627	RP11-443N24.4	lincRNA	4145
rs10844474	12	33269608				AC026357.1	miRNA	74853	SNORD112	snoRNA	245910
rs10506104	12	33278996				AC026357.1	miRNA	84241	SNORD112	snoRNA	236522
rs6488130	12	33308022				AC026357.1	miRNA	113267	SNORD112	snoRNA	207496
rs7132180	12	49159610				LINC00935	protein_coding	41	ADCY6	protein_coding	365
rs17184182	12	67662924				GGTA2P	pseudogene	2178	CAND1	protein_coding	137
rs12426730	12	95002540	TMCC3	protein_coding	intronic,3downstream						
rs9300779	13	103682440				METTL21EP	pseudogene	134057	SLC10A2	protein_coding	13910
rs168518	13	106913481				RNA5SP38	rRNA	105642	LINC00460	lincRNA	115430
rs354463	13	106914872				RNA5SP38	rRNA	107033	LINC00460	lincRNA	114039
rs6492017	13	106928488				RNA5SP38	rRNA	120649	LINC00460	lincRNA	100423
rs7332922	13	52134425				MIR4703	miRNA	7622	RNU6-65P	snRNA	5123
rs3093872	14	20811332	RPPH1	antisense	non-coding						
rs10872856	14	21286675				RNASE1	protein_coding	15238	RP11-219E7.3	lincRNA	51762
rs10143202	14	61124940	SIX1	protein_coding	coding nonsyn						
rs7171993	15	35145065				RP11-83J16.3	pseudogene	2469	AQR	protein_coding	2667
rs3743121	15	35147345				RP11-83J16.3	pseudogene	4749	AQR	protein_coding	387
rs8029757	15	53688298				RP11-209E8.1	lincRNA	266403	WDR72	protein_coding	117640
rs2660825	15	73557903	NEO1	protein_coding	intronic						
rs306204	15	85388376	ALPK3	protein_coding	intronic						
rs2289138	15	85407564	ALPK3	protein_coding	intronic,non-coding intronic						
rs896364	15	85519657				SLC28A1	protein_coding	781	PDE8A	protein_coding	4014
rs12900078	15	85523969	PDE8A	protein_coding	intronic,5upstream						

rs11853328	15	93698096				RGMA	protein_coding	65663	AC112693.2	protein_coding	51199
rs8043910	16	78698238	WWOX	protein_coding	intronic						
rs17249591	17	31854060	ASIC2	protein_coding	intronic						
rs17249607	17	31855826	ASIC2	protein_coding	intronic						
rs17836850	17	31862176	ASIC2	protein_coding	intronic						
rs17836850	17	31862176	RP11-31I22.4	antisense	non-coding intronic						
rs12947624	17	71776939	LINC00469	lincRNA	non-coding intronic						
rs7208029	17	71786783	LINC00469	lincRNA	non-coding intronic						
rs7208029	17	71786783	AC125421.1	lincRNA	non-coding						
rs9897183	17	71811006	LINC00469	lincRNA	non-coding intronic						
rs2619980	17	71816888	LINC00469	lincRNA	non-coding intronic						
rs3813026	17	76123528	TMC6	protein_coding	intronic,5upstream,non-coding intronic						
rs10432142	18	7199531				SLC25A51P2	pseudogene	63713	LRRC30	protein_coding	31592
rs322132	19	11377939				DOCK6	protein_coding	4782	TSPAN16	protein_coding	28885
rs322133	19	11377975				DOCK6	protein_coding	4818	TSPAN16	protein_coding	28849
rs16970932	19	36673300	ZNF565	protein_coding	3utr,3downstream						
rs16989256	19	58313191	ZNF586	protein_coding	intronic						
rs7250447	19	58319123	ZNF586	protein_coding	intronic						
rs7250447	19	58319123	ZNF552	protein_coding	intronic,3utr,3downstream						
rs2303756	19	58331012	ZNF586	protein_coding	3utr						
rs34871518	19	58354265	CTD-2583A14.10	protein_coding	intronic						
rs34871518	19	58354265	ZNF587B	protein_coding	3downstream,intronic						

rs1047920	20	48605107	SNAI1	protein_coding	3utr						
rs6013654	20	51887837	TSHZ2	protein_coding	intronic						
rs6013654	20	51887837	RP4-678D15.1	antisense	non-coding intronic						
rs1877430	20	51888122	TSHZ2	protein_coding	intronic						
rs1877430	20	51888122	RP4-678D15.1	antisense	non-coding intronic						
rs467028	21	31292936	GRIK1	protein_coding	intronic,non-coding intronic						
rs465555	21	31293409	GRIK1	protein_coding	intronic,non-coding intronic						
rs466081	21	31296871	GRIK1	protein_coding	intronic,non-coding intronic						
rs458965	21	31311490	GRIK1	protein_coding	intronic,non-coding intronic						
rs1012042	21	31319562				GRIK1	protein_coding	7211	AF096876.1	lincRNA	142473
rs2832503	21	31324181				GRIK1	protein_coding	11830	AF096876.1	lincRNA	137854
rs977779	21	31325538				GRIK1	protein_coding	13187	AF096876.1	lincRNA	136497
rs2832513	21	31366120				GRIK1	protein_coding	53769	AF096876.1	lincRNA	95915
rs9977816	21	45653614	ICOSLG	protein_coding	intronic						
rs132323	22	29567431				KREMEN1	protein_coding	3110	RNU6-1219P	snRNA	20254
rs132366	22	29599984				RNU6-1219P	snRNA	12188	EMID1	protein_coding	1856
rs9622771	22	38675738				AL020993.1	miRNA	5308	CSNK1E	protein_coding	10959
rs17753394	22	38678387				AL020993.1	miRNA	7957	CSNK1E	protein_coding	8310
rs1884417	X	140316168				LDOC1	protein_coding	44858	SPANXC	protein_coding	19428
rs5953842	X	142554570	GS1-256O22.5	antisense	non-coding intronic						
rs12689477	X	54076358				PHF8	protein_coding	967	U3	snoRNA	14920
rs12688331	X	54208112	FAM120C	protein_coding	intronic,non-coding intronic						

Appendix E: SNPnexus phenotype & disease association

Result table with phenotype and disease association information from the Genetic Association Database (GAD), for the 204 SNPs identified with SAERMA. The following features were used to report the information for the SNPs indexed from GAD:

- SNP: SNP name
- Association: Confirmed association
- Phenotype: Phenotype description
- Disease_Class: Type of disease
- Gene: Gene name
- Pubmed: Pubmed ID of publication of the study

The output file obtained from SNPnexus was subject to a filter criterion so only disease/phenotypes within metabolic and cardiovascular disease class were retained. Other disease classes such as *vision, renal, aging, neurological, chemdependency, psych, haematological, developmental or immune* were not considered relevant. The disease/phenotypes within the metabolic and cardiovascular class were: cholesterol (HDL), waist circumference, triglycerides, lipids, metabolism, obesity, cholesterol, obesity|asthma, coronary disease, body mass index, diabetes mellitus type 2, type 2 diabetes, cardiovascular diseases, diabetes mellitus, hypertension, diastolic blood pressure, cholesterol (LDL), coronary artery disease, diabetes mellitus type 2|obesity, insulin, insulin resistance, blood pressure, body fat distribution, body weight, waist-hip ratio, body weight changes, body weight and measures, diabetes type 2 and triglycerides, diabetes type 2|diabetes type 1.

SNP	Association	Gene	Disease Class	Phenotype	Pubmed
rs10042132	Y	C7orf63	METABOLIC	Cholesterol, HDL	17903299
rs10043659	Y	C7orf63	METABOLIC	Cholesterol, HDL	17903299
rs10083213	Y	TPH2	METABOLIC	Waist Circumference	0
rs1012042	Y	MLL3	METABOLIC	Triglycerides	17903299
rs1012042	Y	NCAM2	METABOLIC	Lipids	17903299
rs1012042	Y	NCAM2	METABOLIC	Metabolism	0
rs1012042	Y	NCAM2	METABOLIC	Obesity	21552555
rs1012042	Y	NCAM2	METABOLIC	Triglycerides	0
rs1012042	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs10180670	Y	AFF3	METABOLIC	Cholesterol	17903299
rs10180670	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs10204782	Y	AFF3	METABOLIC	Cholesterol	17903299
rs10204782	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs10204782	Y	ROCK2	CARDIOVASCULAR	Death, Sudden, Cardiac	21658281
rs1037626	Y	AFF3	METABOLIC	Cholesterol	17903299
rs1037626	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs1046724	Y	PPP1R2P3	METABOLIC	Triglycerides	0
rs10816047		PTPRD	METABOLIC	obesity|asthma	20816195
rs10816047	Y	PTPRD	CARDIOVASCULAR	Coronary Disease	0
rs10816047	Y	PTPRD	METABOLIC	Body Mass Index	0
rs10816047	Y	PTPRD	METABOLIC	Cholesterol, HDL	17903299
rs10816047	Y	PTPRD	METABOLIC	Diabetes Mellitus, Type 2	20174558
rs10816047	Y	PTPRD	METABOLIC	type 2 diabetes	20174558
rs10826675	Y	MPP7	CARDIOVASCULAR	Cardiovascular Diseases	17903304
rs10826675	Y	MPP7	METABOLIC	Body Mass Index	17903300
rs10826675	Y	SVIL	METABOLIC	Diabetes Mellitus	0
rs10849949		ATXN2	CARDIOVASCULAR	Cardiovascular Diseases	21060863
rs10849949		ATXN2	CARDIOVASCULAR	hypertension	19430479
rs10849949		ATXN2	CARDIOVASCULAR	hypertension	20542020
rs10849949		ATXN2	METABOLIC	Obesity	20016785
rs10849949	Y	ATXN2	CARDIOVASCULAR	Diastolic blood pressure	19430483
rs10849949	Y	TPH2	METABOLIC	Waist Circumference	0
rs10850003	Y	TPH2	METABOLIC	Waist Circumference	0
rs10882256	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs10882256	Y	ANK3	METABOLIC	Triglycerides	17903299
rs10882256	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21378988
rs10882256	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21606135
rs10882256	Y	ZNF32	METABOLIC	Body Mass Index	0
rs10998726		HK1	METABOLIC	Diabetes Mellitus, Type 2	19096518
rs10998726	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs10998726	Y	ANK3	METABOLIC	Triglycerides	17903299
rs10998726	Y	HK1	METABOLIC	Diabetes Mellitus, Type 2|Obesity	19651813

rs10998726	Y	ZNF32	METABOLIC	Body Mass Index	0
rs11065884	Y	TPH2	METABOLIC	Waist Circumference	0
rs11102001	Y	EPHB2	METABOLIC	Insulin	0
rs11102001	Y	EPHB2	METABOLIC	Insulin Resistance	0
rs11102001	Y	NTNG1	METABOLIC	Body Mass Index	0
rs11102001	Y	Sep-10	CARDIOVASCULAR	Blood Pressure	53709906
rs11207744	Y	EPHB2	METABOLIC	Insulin	0
rs11207744	Y	EPHB2	METABOLIC	Insulin Resistance	0
rs11692215	Y	AFF3	METABOLIC	Body Fat Distribution	0
rs11692215	Y	AFF3	METABOLIC	Cholesterol	17903299
rs11692215	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs11786916		CSMD1	CARDIOVASCULAR	hypertension	19960030
rs11786916	Y	CSMD1	CARDIOVASCULAR	Blood Pressure	17903302
rs11786916	Y	CSMD1	CARDIOVASCULAR	Coronary Disease	21347282
rs11786916	Y	CSMD1	METABOLIC	Body Fat Distribution	0
rs11786916	Y	CSMD1	METABOLIC	Body Mass Index	0
rs11786916	Y	CSMD1	METABOLIC	Diabetes Mellitus	0
rs11786916	Y	CSMD1	METABOLIC	Insulin Resistance	21901158
rs11786916	Y	DEFA3	METABOLIC	Cholesterol, HDL	17903299
rs12074072	Y	EPHB2	METABOLIC	Insulin	0
rs12074072	Y	EPHB2	METABOLIC	Insulin Resistance	0
rs12315146	Y	TPH2	METABOLIC	Waist Circumference	0
rs12426730	Y	SLC26A5	METABOLIC	Triglycerides	0
rs12426730	Y	TPH2	METABOLIC	Waist Circumference	0
rs12486426	Y	LRRN1	CARDIOVASCULAR	Blood Pressure	0
rs12486426	Y	LRRN1	METABOLIC	Body Weight	17903300
rs12486426	Y	LRRN1	METABOLIC	Cholesterol, HDL	125323093
rs12486426	Y	LRRN1	METABOLIC	Triglycerides	17903299
rs12486426	Y	ROBO2	CARDIOVASCULAR	Blood Pressure	0
rs12486426	Y	ROBO2	METABOLIC	Cholesterol	17903299
rs12486426	Y	ROBO2	METABOLIC	Cholesterol, LDL	17903299
rs12486426	Y	ROBO2	METABOLIC	Diabetes Mellitus	17903298
rs12486426	Y	TKT	METABOLIC	Waist Circumference	0
rs12570210	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21378988
rs12570210	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21606135
rs12570210	Y	TCERG1L	METABOLIC	Diabetes Mellitus, Type 2	21490949
rs12947624	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs13171869	Y	PPP1R2P3	METABOLIC	Triglycerides	0
rs13269315		CSMD1	CARDIOVASCULAR	hypertension	19960030
rs13269315	Y	CSMD1	CARDIOVASCULAR	Blood Pressure	17903302
rs13269315	Y	CSMD1	CARDIOVASCULAR	Coronary Disease	21347282
rs13269315	Y	CSMD1	METABOLIC	Body Fat Distribution	0
rs13269315	Y	CSMD1	METABOLIC	Body Mass Index	0

rs13269315	Y	CSMD1	METABOLIC	Diabetes Mellitus	0
rs13269315	Y	CSMD1	METABOLIC	Insulin Resistance	21901158
rs13269315	Y	DEFA3	METABOLIC	Cholesterol, HDL	17903299
rs1452044	Y	AFF3	METABOLIC	Cholesterol	17903299
rs1452044	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs1479955	Y	LRRN1	CARDIOVASCULAR	Blood Pressure	0
rs1479955	Y	LRRN1	METABOLIC	Body Weight	17903300
rs1479955	Y	LRRN1	METABOLIC	Cholesterol, HDL	125323093
rs1479955	Y	LRRN1	METABOLIC	Triglycerides	17903299
rs1479955	Y	TKT	METABOLIC	Waist Circumference	0
rs1544396	Y	TPH2	METABOLIC	Waist Circumference	0
rs16894413	Y	C7orf63	METABOLIC	Cholesterol, HDL	17903299
rs17015634	Y	AFF3	METABOLIC	Cholesterol	17903299
rs17015634	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs17184182	Y	GRIP1	CARDIOVASCULAR	Coronary Artery Disease	0
rs17184182	Y	GRIP1	METABOLIC	Waist-Hip Ratio	0
rs17184182	Y	SLC26A5	METABOLIC	Triglycerides	0
rs17249591	Y	ACCN1	METABOLIC	Body Mass Index	0
rs17249591	Y	ACCN1	METABOLIC	Body Weight	0
rs17249591	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs17249607	Y	ACCN1	METABOLIC	Body Mass Index	0
rs17249607	Y	ACCN1	METABOLIC	Body Weight	0
rs17249607	Y	ACCN1	METABOLIC	Body Weight	35806600
rs17249607	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs17293732	Y	AFF3	METABOLIC	Cholesterol	17903299
rs17293732	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs17349380	Y	LRRN1	CARDIOVASCULAR	Blood Pressure	0
rs17349380	Y	LRRN1	METABOLIC	Body Weight	17903300
rs17349380	Y	LRRN1	METABOLIC	Cholesterol, HDL	125323093
rs17349380	Y	LRRN1	METABOLIC	Triglycerides	17903299
rs17349380	Y	TKT	METABOLIC	Waist Circumference	0
rs17662453	Y	AFF3	METABOLIC	Cholesterol	17903299
rs17662453	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs17753394	Y	LARGE	CARDIOVASCULAR	Blood Pressure	17903302
rs17753394	Y	LARGE	METABOLIC	Cholesterol	0
rs17753394	Y	LARGE	METABOLIC	Cholesterol, HDL	0
rs17753394	Y	LARGE	METABOLIC	Cholesterol, LDL	0
rs17753394	Y	LARGE	METABOLIC	Cholesterol, LDL	17903299
rs17792616	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs17792616	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs17792616	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs17792616	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs17836850	Y	ACCN1	METABOLIC	Body Mass Index	0
rs17836850	Y	ACCN1	METABOLIC	Body Weight	0

rs17836850	Y	ACCN1	METABOLIC	Body Weight	71613200
rs17836850	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs1877430	Y	TSHZ2	METABOLIC	Cholesterol	0
rs1877430	Y	TSHZ2	METABOLIC	Cholesterol, LDL	0
rs1877430	Y	TSHZ2	METABOLIC	Triglycerides	35806598
rs1955311	Y	HTRA3	METABOLIC	Cholesterol, LDL	0
rs2042867	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs2042867	Y	ANK3	METABOLIC	Triglycerides	17903299
rs2042867	Y	ZNF32	METABOLIC	Body Mass Index	0
rs2073950		ATXN2	CARDIOVASCULAR	Cardiovascular Diseases	21060863
rs2073950		ATXN2	CARDIOVASCULAR	hypertension	19430479
rs2073950		ATXN2	CARDIOVASCULAR	hypertension	20542020
rs2073950		ATXN2	METABOLIC	Obesity	20016785
rs2073950	Y	ATXN2	CARDIOVASCULAR	Diastolic blood pressure	19430483
rs2073950	Y	TPH2	METABOLIC	Waist Circumference	0
rs2080390	Y	AFF3	METABOLIC	Cholesterol	17903299
rs2080390	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs2116588	Y	AFF3	METABOLIC	Cholesterol	17903299
rs2116588	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs2154762		ITGA9	CARDIOVASCULAR	hypertension	20414254
rs2154762		ITGA9	CARDIOVASCULAR	hypertension	20479155
rs2154762	Y	LRRN1	CARDIOVASCULAR	Blood Pressure	0
rs2154762	Y	LRRN1	METABOLIC	Body Weight	17903300
rs2154762	Y	LRRN1	METABOLIC	Cholesterol, HDL	125323093
rs2154762	Y	LRRN1	METABOLIC	Triglycerides	17903299
rs220655	Y	AFF3	METABOLIC	Cholesterol	17903299
rs220655	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs2206734		CDKAL1	METABOLIC	Diabetes Mellitus	18264689
rs2206734		CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19172244
rs2206734		CDKAL1	METABOLIC	Diabetes Mellitus|Diabetes Mellitus, Type 2|	19139842
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19502414
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18544707
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19258404
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19862325
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19324937
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19741467
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19401414
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19602701
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19794065
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18437351
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19020323
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18984664
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19592620
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19933996

rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19734900
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19892838
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19247372
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18469204
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19380854
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19720844
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18516622
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19225753
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19228808
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19008344
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	17463248
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18461161
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19082521
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19020324
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19002430
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	19033397
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18633108
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18991055
rs2206734		CDKAL1	METABOLIC	diabetes, type 2	18477659
rs2206734		CDKAL1	METABOLIC	diabetes, type 2 triglycerides	17463246
rs2206734		CDKAL1	METABOLIC	Obesity	20712903
rs2206734		CDKAL1	METABOLIC	Obesity	20816152
rs2206734		CDKAL1	UNKNOWN	diabetes, type 2 | diabetes, type 1	18426861
rs2206734	N	CDKAL1	UNKNOWN	diabetes, type 2 | diabetes, type 1	19455305
rs2206734	Y	CDKAL1	METABOLIC	Body Mass Index	22344219
rs2206734	Y	CDKAL1	METABOLIC	Body Mass Index	22344221
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	20581827
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17463249
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19401414
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	18711366
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19734900
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19056611
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17463246
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17460697
rs2206734	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	21490949
rs2206734	Y	CDKAL1	METABOLIC	diabetes, type 2	18591388
rs2206734	Y	CDKAL1	METABOLIC	diabetes, type 2	17460697
rs2206734	Y	CDKAL1	METABOLIC	diabetes, type 2	17463249
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	17463246
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	17463248
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	17460697
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	19401414
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	18711366

rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	17554300
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	17463249
rs2206734	Y	CDKAL1	METABOLIC	type 2 diabetes	18372903
rs2285727	Y	TPH2	METABOLIC	Waist Circumference	0
rs2301621		ATXN2	CARDIOVASCULAR	Cardiovascular Diseases	21060863
rs2301621		ATXN2	CARDIOVASCULAR	hypertension	19430479
rs2301621		ATXN2	CARDIOVASCULAR	hypertension	20542020
rs2301621		ATXN2	METABOLIC	Obesity	20016785
rs2301621	Y	ATXN2	CARDIOVASCULAR	Diastolic blood pressure	19430483
rs2301621	Y	TPH2	METABOLIC	Waist Circumference	0
rs2339941	Y	NAA25	METABOLIC	Cholesterol, LDL	17903299
rs2339941	Y	TPH2	METABOLIC	Waist Circumference	0
rs2410518	Y	MTMR7	METABOLIC	Body Mass Index	0
rs2410518	Y	MTMR7	METABOLIC	Body Mass Index	17903300
rs2410518	Y	MTMR7	METABOLIC	Body Weight Changes	17903300
rs2521766	Y	DFNA5	METABOLIC	Cholesterol, LDL	35806598
rs2521766	Y	EIF2AK1	METABOLIC	Triglycerides	17903299
rs2546440	Y	PPP1R2P3	METABOLIC	Triglycerides	0
rs2619980	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs2711098	Y	EIF2AK1	METABOLIC	Triglycerides	17903299
rs2832503	Y	MLL3	METABOLIC	Triglycerides	17903299
rs2832503	Y	NCAM2	METABOLIC	Lipids	17903299
rs2832503	Y	NCAM2	METABOLIC	Metabolism	0
rs2832503	Y	NCAM2	METABOLIC	Obesity	21552555
rs2832503	Y	NCAM2	METABOLIC	Triglycerides	0
rs2832503	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs2832513	Y	MLL3	METABOLIC	Triglycerides	17903299
rs2832513	Y	NCAM2	METABOLIC	Lipids	17903299
rs2832513	Y	NCAM2	METABOLIC	Metabolism	0
rs2832513	Y	NCAM2	METABOLIC	Obesity	21552555
rs2832513	Y	NCAM2	METABOLIC	Triglycerides	0
rs2832513	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs2876086	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs2876086	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs2876086	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs2876086	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs3177647	Y	TPH2	METABOLIC	Waist Circumference	0
rs3758650	Y	CDHR5	METABOLIC	Diabetes Mellitus	0
rs3795308	Y	EPHB2	METABOLIC	Insulin	0
rs3795308	Y	EPHB2	METABOLIC	Insulin Resistance	0
rs3809288	Y	TPH2	METABOLIC	Waist Circumference	0
rs3813026	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs396394	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392

rs396394	Y	ANK3	METABOLIC	Triglycerides	17903299
rs396394	Y	ZNF32	METABOLIC	Body Mass Index	0
rs400659	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs400659	Y	ANK3	METABOLIC	Triglycerides	17903299
rs400659	Y	ZNF32	METABOLIC	Body Mass Index	0
rs406671	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs406671	Y	ANK3	METABOLIC	Triglycerides	17903299
rs406671	Y	ZNF32	METABOLIC	Body Mass Index	0
rs42695	Y	EIF2AK1	METABOLIC	Triglycerides	17903299
rs458965	Y	GRIK1	METABOLIC	Body Mass Index	17903300
rs458965	Y	GRIK1	METABOLIC	Body Mass Index	22446040
rs458965	Y	GRIK1	METABOLIC	Body Weight	17903300
rs458965	Y	MLL3	METABOLIC	Triglycerides	17903299
rs458965	Y	NCAM2	METABOLIC	Lipids	17903299
rs458965	Y	NCAM2	METABOLIC	Metabolism	0
rs458965	Y	NCAM2	METABOLIC	Obesity	21552555
rs458965	Y	NCAM2	METABOLIC	Triglycerides	0
rs458965	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs4648328		ALDH2	CARDIOVASCULAR	Cardiovascular Diseases	20541757
rs4648328		ALDH2	CARDIOVASCULAR	Coronary Artery Disease	20417517
rs4648328		ALDH2	CARDIOVASCULAR	hypertension	20877124
rs4648328		ALDH2	CARDIOVASCULAR	hypertension	15167446
rs4648328		ALDH2	CARDIOVASCULAR	hypertension	12484509
rs4648328		ALDH2	CARDIOVASCULAR	hypertension	17785925
rs4648328		ALDH2	CARDIOVASCULAR	hypertension	11510748
rs4648328		ALDH2	METABOLIC	Diabetes Mellitus, Type 2	18216179
rs4648328		ALDH2	METABOLIC	diabetes, type 2	9752691
rs4648328	Y	TPH2	METABOLIC	Waist Circumference	0
rs465555	Y	GRIK1	METABOLIC	Body Mass Index	17903300
rs465555	Y	GRIK1	METABOLIC	Body Mass Index	22446040
rs465555	Y	GRIK1	METABOLIC	Body Weight	17903300
rs465555	Y	MLL3	METABOLIC	Triglycerides	17903299
rs465555	Y	NCAM2	METABOLIC	Lipids	17903299
rs465555	Y	NCAM2	METABOLIC	Metabolism	0
rs465555	Y	NCAM2	METABOLIC	Obesity	21552555
rs465555	Y	NCAM2	METABOLIC	Triglycerides	0
rs465555	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs466081	Y	GRIK1	METABOLIC	Body Mass Index	17903300
rs466081	Y	GRIK1	METABOLIC	Body Mass Index	22446040
rs466081	Y	GRIK1	METABOLIC	Body Weight	17903300
rs466081	Y	MLL3	METABOLIC	Triglycerides	17903299
rs466081	Y	NCAM2	METABOLIC	Lipids	17903299
rs466081	Y	NCAM2	METABOLIC	Metabolism	0
rs466081	Y	NCAM2	METABOLIC	Obesity	21552555
rs466081	Y	NCAM2	METABOLIC	Triglycerides	0

rs466081	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs467028	Y	GRIK1	METABOLIC	Body Mass Index	17903300
rs467028	Y	GRIK1	METABOLIC	Body Mass Index	22446040
rs467028	Y	GRIK1	METABOLIC	Body Weight	17903300
rs467028	Y	MLL3	METABOLIC	Triglycerides	17903299
rs467028	Y	NCAM2	METABOLIC	Lipids	17903299
rs467028	Y	NCAM2	METABOLIC	Metabolism	0
rs467028	Y	NCAM2	METABOLIC	Obesity	21552555
rs467028	Y	NCAM2	METABOLIC	Triglycerides	0
rs467028	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs4678980		ITGA9	CARDIOVASCULAR	hypertension	20414254
rs4678980		ITGA9	CARDIOVASCULAR	hypertension	20479155
rs4678980	Y	LRRN1	CARDIOVASCULAR	Blood Pressure	0
rs4678980	Y	LRRN1	METABOLIC	Body Weight	17903300
rs4678980	Y	LRRN1	METABOLIC	Cholesterol, HDL	125323093
rs4678980	Y	LRRN1	METABOLIC	Triglycerides	17903299
rs4750827	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21378988
rs4750827	Y	LIPA	CARDIOVASCULAR	Coronary Artery Disease	21606135
rs4767939		ALDH2	CARDIOVASCULAR	Cardiovascular Diseases	20541757
rs4767939		ALDH2	CARDIOVASCULAR	Coronary Artery Disease	20417517
rs4767939		ALDH2	CARDIOVASCULAR	hypertension	20877124
rs4767939		ALDH2	CARDIOVASCULAR	hypertension	15167446
rs4767939		ALDH2	CARDIOVASCULAR	hypertension	12484509
rs4767939		ALDH2	CARDIOVASCULAR	hypertension	17785925
rs4767939		ALDH2	CARDIOVASCULAR	hypertension	11510748
rs4767939		ALDH2	METABOLIC	Diabetes Mellitus, Type 2	18216179
rs4767939		ALDH2	METABOLIC	diabetes, type 2	9752691
rs4767939	Y	TPH2	METABOLIC	Waist Circumference	0
rs4850920	Y	AFF3	METABOLIC	Body Fat Distribution	0
rs4850920	Y	AFF3	METABOLIC	Cholesterol	17903299
rs4850920	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs4897427	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs4897427	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs4897427	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs4897427	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs4979935	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs4979935	Y	ANK3	METABOLIC	Triglycerides	17903299
rs4979935	Y	ZNF32	METABOLIC	Body Mass Index	0
rs6013654	Y	TSHZ2	METABOLIC	Cholesterol	0
rs6013654	Y	TSHZ2	METABOLIC	Cholesterol, LDL	0
rs6013654	Y	TSHZ2	METABOLIC	Triglycerides	35806598
rs632650	Y	TPH2	METABOLIC	Waist Circumference	0
rs640783	Y	TPH2	METABOLIC	Waist Circumference	0

rs6569962	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs6569962	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs6569962	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs6569962	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs6713524	Y	AFF3	METABOLIC	Body Fat Distribution	0
rs6713524	Y	AFF3	METABOLIC	Cholesterol	17903299
rs6713524	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs6914876	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs6914876	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs6914876	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs6914876	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs6935599		CDKAL1	METABOLIC	Diabetes Mellitus	18264689
rs6935599		CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19172244
rs6935599		CDKAL1	METABOLIC	Diabetes Mellitus|Diabetes Mellitus, Type 2|	19139842
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19502414
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18544707
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19258404
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19862325
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19324937
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19741467
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19401414
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19602701
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19794065
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18437351
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19020323
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18984664
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19592620
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19933996
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19734900
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19892838
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19247372
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18469204
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19380854
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19720844
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18516622
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19225753
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19228808
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19008344
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	17463248
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18461161
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19082521
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19020324
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19002430
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	19033397

rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18633108
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18991055
rs6935599		CDKAL1	METABOLIC	diabetes, type 2	18477659
rs6935599		CDKAL1	METABOLIC	diabetes, type 2 triglycerides	17463246
rs6935599		CDKAL1	METABOLIC	Obesity	20712903
rs6935599		CDKAL1	METABOLIC	Obesity	20816152
rs6935599		CDKAL1	UNKNOWN	diabetes, type 2 | diabetes, type 1	18426861
rs6935599	N	CDKAL1	UNKNOWN	diabetes, type 2 | diabetes, type 1	19455305
rs6935599	Y	CDKAL1	METABOLIC	Body Mass Index	22344219
rs6935599	Y	CDKAL1	METABOLIC	Body Mass Index	22344221
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	20581827
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	18372903
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17463249
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19401414
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	18711366
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19734900
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	19056611
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17463246
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	17460697
rs6935599	Y	CDKAL1	METABOLIC	Diabetes Mellitus, Type 2	21490949
rs6935599	Y	CDKAL1	METABOLIC	diabetes, type 2	18591388
rs6935599	Y	CDKAL1	METABOLIC	diabetes, type 2	17460697
rs6935599	Y	CDKAL1	METABOLIC	diabetes, type 2	17463249
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	17463246
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	17463248
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	17460697
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	19401414
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	18711366
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	17554300
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	17463249
rs6935599	Y	CDKAL1	METABOLIC	type 2 diabetes	18372903
rs6958382	Y	EIF2AK1	METABOLIC	Triglycerides	17903299
rs7098565	Y	MPP7	CARDIOVASCULAR	Cardiovascular Diseases	17903304
rs7098565	Y	MPP7	METABOLIC	Body Mass Index	17903300
rs7098565	Y	SVIL	METABOLIC	Diabetes Mellitus	0
rs7208029	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs735638	Y	ANK3	METABOLIC	Cholesterol, LDL	143226392
rs735638	Y	ANK3	METABOLIC	Triglycerides	17903299
rs735638	Y	ZNF32	METABOLIC	Body Mass Index	0
rs7386192	Y	MTMR7	METABOLIC	Body Mass Index	0
rs7386192	Y	MTMR7	METABOLIC	Body Mass Index	17903300
rs7386192	Y	MTMR7	METABOLIC	Body Weight Changes	17903300

rs7448421	Y	PPP1R2P3	METABOLIC	Triglycerides	0
rs7579477	Y	AFF3	METABOLIC	Cholesterol	17903299
rs7579477	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs762027	Y	AFF3	METABOLIC	Cholesterol	17903299
rs762027	Y	AFF3	METABOLIC	Cholesterol, HDL	17903299
rs7765392	Y	RBMXP1	METABOLIC	Body Mass Index	17903300
rs7765392	Y	RBMXP1	METABOLIC	Body Weight Changes	17903300
rs7765392	Y	RBMXP1	METABOLIC	Body Weights and Measures	17903300
rs7765392	Y	RBMXP1	METABOLIC	Cholesterol, HDL	17903299
rs7962138	Y	TPH2	METABOLIC	Waist Circumference	0
rs7974383	Y	C12orf51	CARDIOVASCULAR	Blood Pressure	17903302
rs7974383	Y	C12orf51	METABOLIC	Cholesterol	53709897
rs7974383	Y	C12orf51	METABOLIC	Cholesterol, HDL	21909109
rs7974383	Y	C12orf51	METABOLIC	Cholesterol, LDL	89516495
rs7974383	Y	C12orf51	METABOLIC	Waist-Hip Ratio	19396169
rs7974383	Y	TPH2	METABOLIC	Waist Circumference	0
rs8043910		WWOX	CARDIOVASCULAR	Cardiovascular Diseases	20942981
rs8043910	Y	CDH3	METABOLIC	Diabetes Mellitus	17903298
rs8043910	Y	WWOX	CARDIOVASCULAR	Cardiovascular Diseases	17903304
rs9622771	Y	LARGE	CARDIOVASCULAR	Blood Pressure	17903302
rs9622771	Y	LARGE	METABOLIC	Cholesterol	0
rs9622771	Y	LARGE	METABOLIC	Cholesterol, HDL	0
rs9622771	Y	LARGE	METABOLIC	Cholesterol, LDL	0
rs9622771	Y	LARGE	METABOLIC	Cholesterol, LDL	17903299
rs977779	Y	MLL3	METABOLIC	Triglycerides	17903299
rs977779	Y	NCAM2	METABOLIC	Lipids	17903299
rs977779	Y	NCAM2	METABOLIC	Metabolism	0
rs977779	Y	NCAM2	METABOLIC	Obesity	21552555
rs977779	Y	NCAM2	METABOLIC	Triglycerides	0
rs977779	Y	NCAM2	METABOLIC	Triglycerides	17903299
rs9887921	Y	EPHB2	METABOLIC	Insulin	0
rs9887921	Y	EPHB2	METABOLIC	Insulin Resistance	0
rs9888055	Y	MPP7	CARDIOVASCULAR	Cardiovascular Diseases	17903304
rs9888055	Y	MPP7	METABOLIC	Body Mass Index	17903300
rs9888055	Y	SVIL	METABOLIC	Diabetes Mellitus	0
rs9897183	Y	PEMT	CARDIOVASCULAR	Coronary Artery Disease	21378990
rs9971942	Y	TPH2	METABOLIC	Waist Circumference	0
rs9977816	Y	MLL3	METABOLIC	Triglycerides	17903299
rs9977816	Y	NCAM2	METABOLIC	Lipids	17903299
rs9977816	Y	NCAM2	METABOLIC	Metabolism	0
rs9977816	Y	NCAM2	METABOLIC	Obesity	21552555
rs9977816	Y	NCAM2	METABOLIC	Triglycerides	0
rs9977816	Y	NCAM2	METABOLIC	Triglycerides	17903299

Appendix F: *i*-GSEA4GWAS results for GO terms

The following table contains a list with the different GO terms correlated to obesity using the GWAS data, via gene set enrichment analysis.

Gene set name	GO term	Gene set P-value	Gene set FDR	Significant genes
REGULATION_OF_SECRETION	GO:0051046	0.002	0.05716	16
SPLICEOSOME	GO:0005681	0.006	0.059166666	10
DNA_CATABOLIC_PROCESS	GO:0006308	0.002	0.061	9
HELICASE_ACTIVITY	GO:0004386	0.004	0.061	18
NEGATIVE_REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	GO:0051241	0.002	0.06385714	13
PROTEIN_RNA_COMPLEX_ASSEMBLY	GO:0022618	0.002	0.06523077	15
NUCLEOPLASM_PART	GO:0044451	0.002	0.066296294	52
CYCLIC_NUCLEOTIDE_MEDIATED_SIGNALING	GO:0019935	0.001	0.0663	46
CHROMOSOME__PERICENTRIC_REGION	GO:0000775	0.001	0.069	14
PROTEIN_C_TERMINUS_BINDING	GO:0008022	< 0.001	0.06983333	24
DNA_HELICASE_ACTIVITY	GO:0003678	0.001	0.0714	9
STRUCTURE_SPECIFIC_DNA_BINDING	GO:0043566	0.004	0.07272223	17
ANGIOGENESIS	GO:0001525	0.004	0.0745	20
DNA_DIRECTED_RNA_POLYMERASE_II__HOLOENZYME	GO:0016591	0.003	0.07482759	17
VIRAL_REPRODUCTIVE_PROCESS	GO:0022415	0.004	0.07541176	12
POSITIVE_REGULATION_OF_CYTOKINE_BIOSYNTHETIC_PROCESS	GO:0042108	0.002	0.076	9
PROTEIN_DIMERIZATION_ACTIVITY	GO:0046983	0.002	0.0766	61
NEUROTRANSMITTER_BINDING	GO:0042165	0.004	0.077125	20
G_PROTEIN_SIGNALING__COUPLED_TO_CAMP_NUCLEOTIDE_SECOND_MESSENGER	GO:0007188	0.005	0.0776129	31
RECEPTOR_COMPLEX	GO:0043235	0.001	0.0779	25
SECOND_MESSENGER_MEDIATED_SIGNALING	GO:0019932	< 0.001	0.080133334	63
NEUROTRANSMITTER_RECEPTOR_ACTIVITY	GO:0030594	0.006	0.080484845	19
SEQUENCE_SPECIFIC_DNA_BINDING	GO:0043565	< 0.001	0.0805	24
RNA_PROCESSING	GO:0006396	0.002	0.08214285	36

POSITIVE_REGULATION_OF_TRANSLATION	GO:0045727	< 0.001	0.08469231	13
G_PROTEIN_SIGNALING__COUPLED_TO_CYCLIC_NUCLEOTIDE_SECOND_MESSENGER	GO:0007187	< 0.001	0.0855	46
SYNAPSE	GO:0045202	0.001	0.08645454	14
CYTOKINE_METABOLIC_PROCESS	GO:0042107	< 0.001	0.0865	15
CYTOKINE_PRODUCTION	GO:0001816	< 0.001	0.086555555	28
CAMP_MEDIATED_SIGNALING	GO:0019933	0.006	0.09071053	31
RIBONUCLEOPROTEIN_COMPLEX	GO:0030529	0.002	0.090972975	26
REGULATION_OF_NEUROTRANSMITTER_LEVELS	GO:0001505	0.001	0.09163889	13
REGULATION_OF_MITOSI	GO:0007088	0.006	0.092685714	12
VIRAL_REPRODUCTION	GO:0016032	0.003	0.0935	13
SYNAPTIC_TRANSMISSION	GO:0007268	0.001	0.09375	69
PROTEIN_HOMODIMERIZATION_ACTIVITY	GO:0042803	0.001	0.09385294	41
M_PHASE_OF_MITOTIC_CELL_CYCLE	GO:0000087	0.006	0.09446154	20
POSITIVE_REGULATION_OF_CELL_DIFFERENTIATION	GO:0045597	0.001	0.09678572	15
MITOSIS	GO:0007067	0.008	0.09795121	19
NITROGEN_COMPOUND_CATABOLIC_PROCESS	GO:0044270	0.009	0.09863637	11
DEPHOSPHORYLATION	GO:0016311	0.003	0.09890697	27
REGULATION_OF_CYTOKINE_BIOSYNTHETIC_PROCESS	GO:0042035	< 0.001	0.102	14
DOUBLE_STRANDED_DNA_BINDING	GO:0003690	0.013	0.10208333	10
PHOSPHOPROTEIN_PHOSPHATASE_ACTIVITY	GO:0004721	0.004	0.10266667	32
G_PROTEIN_SIGNALING__ADENYLATE_CYCLASE_ACTIVATING_PATHWAY	GO:0007189	0.004	0.10293999	11
VASCULATURE_DEVELOPMENT	GO:0001944	0.006	0.10334783	21
ENDONUCLEASE_ACTIVITY	GO:0004519	0.012	0.10357777	8
REGULATION_OF_MITOTIC_CELL_CYCLE	GO:0007346	0.003	0.103893615	9
CELL_SUBSTRATE_ADHESION	GO:0031589	0.006	0.10389796	16
CYTOKINE_BIOSYNTHETIC_PROCESS	GO:0042089	0.001	0.10585714	14

PROTEIN_COMPLEX_BINDING	GO:0032403	0.007	0.106037036	21
MITOTIC_CELL_CYCLE	GO:0000278	0.009	0.106625006	39
NEGATIVE_REGULATION_OF_CELL_DIFFERENTIATION	GO:0045596	0.005	0.10696226	16
POSITIVE_REGULATION_OF_PROTEIN_METABOLIC_PROCESS	GO:0051247	0.013	0.10732692	24
IRAL_INFECTIOUS_CYCLE	GO:0019058	0.011	0.107836366	10
GENERATION_OF_A_SIGNAL_INVOLVED_IN_CELL_CELL_SIGNALING	GO:0003001	< 0.001	0.108	16
REGULATION_OF_MULTICELLULAR_ORGANISMAL_PROCESS	GO:0051239	0.004	0.1272456	54
POSITIVE_REGULATION_OF_CELLULAR_PROTEIN_METABOLIC_PROCESS	GO:0032270	0.015	0.1275	23
TRANSMISSION_OF_NERVE_IMPULSE	GO:0019226	0.006	0.12795652	73
RESPONSE_TO_RADIATION	GO:0009314	0.014	0.12827693	15
REGULATION_OF_CELL_CYCLE	GO:0051726	0.008	0.1285625	52
PROTEIN_HETERODIMERIZATION_ACTIVITY	GO:0046982	0.013	0.12877941	28
INTERMEDIATE_FILAMENT	GO:0005882	0.013	0.1297	8
INTERMEDIATE_FILAMENT_CYTOSKELETON	GO:0045111	0.013	0.1297	8
ANATOMICAL_STRUCTURE_FORMATION	GO:0048646	0.01	0.12980281	24
REGULATION_OF_G_PROTEIN_COUPLED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	GO:0008277	0.005	0.12987931	10
REGULATION_OF_CELL_DIFFERENTIATION	GO:0045595	0.001	0.13015872	29
MRNA_PROCESSING_GO_0006397	GO:0006397	0.014	0.13040909	16
CELL_CYCLE_PROCESS	GO:0022402	0.016	0.13048612	42
DNA_REPAIR	GO:0006281	0.01	0.1307015	30
CELL_MATRIX_ADHESION	GO:0007160	0.014	0.13109589	15
UBIQUITIN_PROTEIN_LIGASE_ACTIVITY	GO:0004842	0.014	0.13147542	14
MYELOID_CELL_DIFFERENTIATION	GO:0030099	0.008	0.132	15
CELL_CYCLE_PHASE	GO:0022403	0.012	0.13202667	39
REGULATION_OF_PROTEIN_METABOLIC_PROCESS	GO:0051246	0.014	0.13293242	47
SMALL_CONJUGATING_PROTEIN_LIGASE_ACTIVITY	GO:0019787	0.017	0.13485526	14

ENDOSOME	GO:0005768	0.019	0.13974118	19
PROTEIN_AMINO_ACID_DEPHOSPHORYLATION	GO:0006470	0.013	0.14083117	25
REGULATION_OF_MAPKKK_CASCADE	GO:0043408	0.025	0.141	7
G_PROTEIN_COUPLED_RECEPTOR_ACTIVITY	GO:0004930	0.016	0.14111827	71
REGULATION_OF_TRANSFERASE_ACTIVITY	GO:0051338	0.017	0.14196809	49
M_PHASE	GO:0000279	0.019	0.14215	24
ATPASE_ACTIVITY	GO:0016887	0.015	0.14228916	34
POSITIVE_REGULATION_OF_TRANSFERASE_ACTIVITY	GO:0051347	0.017	0.14235869	31
RESPONSE_TO_LIGHT_STIMULUS	GO:0009416	0.022	0.1424557	13
RESPONSE_TO_HYPOXIA	GO:0001666	0.012	0.14247435	11
ADHERENS_JUNCTION	GO:0005912	0.015	0.14265853	10
DOUBLE_STRAND_BREAK_REPAIR	GO:0006302	0.028	0.1431978	8
ACID_AMINO_ACID_LIGASE_ACTIVITY	GO:0016881	0.022	0.14374074	15
REGULATION_OF_TRANSLATION	GO:0006417	0.02	0.1446	24
NA_SPLICING	GO:0008380	0.023	0.1446628	18
POSITIVE_REGULATION_OF_CATALYTIC_ACTIVITY	GO:0043085	0.022	0.14594382	51
HOMEOSTATIC_PROCESS	GO:0042592	0.01	0.14613636	65
ATPASE_ACTIVITY__COUPLED_TO_TRANSMEMBRANE_MOVEMENT_OF_IONS	GO:0042625	0.018	0.14767815	9
SMALL_PROTEIN_CONJUGATING_ENZYME_ACTIVITY	GO:0008639	0.021	0.14805263	15
AMINE_CATABOLIC_PROCESS	GO:0009310	0.027	0.15058586	10
TRANSCRIPTION_FACTOR_COMPLEX	GO:0005667	0.02	0.15113266	22
RESPONSE_TO_DNA_DAMAGE_STIMULUS	GO:0006974	0.018	0.15113401	38
RHODOPSIN_LIKE_RECEPTOR_ACTIVITY	GO:0001584	0.015	0.1514375	47
POSITIVE_REGULATION_OF_CELL_PROLIFERATION	GO:0008284	0.016	0.15253	49
HEMOPOIESIS	GO:0030097	0.018	0.15408911	28
GLUTAMATE_RECEPTOR_ACTIVITY	GO:0008066	0.009	0.15566666	15

IMMUNE_EFFECTOR_PROCESS	GO:0002252	0.021	0.15740776	14
NEGATIVE_REGULATION_OF_NUCLEOBASE__NUCLEOSIDE__NUCLEOTIDE_AND_NUCLEIC_ACID_METABOLIC_PROCESS	GO:0045934	0.019	0.15747747	64
REGULATION_OF_CELLULAR_PROTEIN_METABOLIC_PROCESS	GO:0032268	0.022	0.15839048	44
LIGASE_ACTIVITY	GO:0016874	0.023	0.15842374	26
CHROMOSOME_ORGANIZATION_AND_BIOGENESIS	GO:0051276	0.031	0.15851402	29
POSITIVE_REGULATION_OF_TRANSCRIPTION	GO:0045941	0.02	0.15860909	51
RESPONSE_TO_ENDOGENOUS_STIMULUS	GO:0009719	0.024	0.15879807	47
NEGATIVE_REGULATION_OF_TRANSCRIPTION	GO:0016481	0.021	0.15912296	59
POSITIVE_REGULATION_OF_NUCLEOBASE__NUCLEOSIDE__NUCLEOTIDE_AND_NUCLEIC_ACID_METABOLIC_PROCESS	GO:0045935	0.021	0.15914151	54
TRANSMEMBRANE_RECEPTOR_PROTEIN_KINASE_ACTIVITY	GO:0019199	0.017	0.15919445	27
NEURON_DEVELOPMENT	GO:0048666	0.018	0.15935898	27
LEUKOCYTE_DIFFERENTIATION	GO:0002521	0.029	0.15955372	15
ACTIN_BINDING	GO:0003779	0.021	0.15972413	31
LIGASE_ACTIVITY__FORMING_CARBON_NITROGEN_BONDS	GO:0016879	0.025	0.15997247	18
TRANSLATION	GO:0006412	0.023	0.16000892	34
NUCLEAR_ORGANIZATION_AND_BIOGENESIS	GO:0006997	0.036	0.16052501	8
HEMOPOIETIC_OR_LYMPHOID_ORGAN_DEVELOPMENT	GO:0048534	0.026	0.16086957	28
INTRACELLULAR_RECEPTOR_MEDIATED_SIGNALING_PATHWAY	GO:0030522	0.03	0.1611842	7
CELL_MIGRATION	GO:0016477	0.031	0.1612353	39
MACROMOLECULE_CATABOLIC_PROCESS	GO:0009057	0.026	0.16134399	29
AXON_GUIDANCE	GO:0007411	0.022	0.1619646	13
HUMORAL_IMMUNE_RESPONSE	GO:0006959	0.024	0.16202419	13
PROTEIN_TYROSINE_PHOSPHATASE_ACTIVITY	GO:0004725	0.027	0.16292684	24
CYTOSKELETAL_PROTEIN_BINDING	GO:0008092	0.023	0.1671938	53
RIBONUCLEOPROTEIN_COMPLEX_BIOGENESIS_AND_ASSEMBLY	GO:0022613	0.033	0.1679297	15

RNA_POLYMERASE_II_TRANSCRIPTION_FACTOR_ACTIVITY	GO:0003702	0.027	0.16915748	52
MICROTUBULE_ORGANIZING_CENTER	GO:0005815	0.032	0.16974615	17
MRNA_METABOLIC_PROCESS	GO:0016071	0.038	0.1699127	17
SKELETAL_DEVELOPMENT	GO:0001501	0.029	0.17011194	43
NEURON_DIFFERENTIATION	GO:0030182	0.024	0.170203	35
POSITIVE_REGULATION_OF_RNA_METABOLIC_PROCESS	GO:0051254	0.027	0.17060307	43
PROTEOLYSIS	GO:0006508	0.031	0.17099242	48
HEMATOPOIETIN_INTERFERON_CLASS__D200_DOMAIN__CYTOKINE_RECEPTOR_BINDING	GO:0005126	0.04	0.17252593	7
ATPASE_ACTIVITY__COUPLED	GO:0042623	0.029	0.18455148	30
NEGATIVE_REGULATION_OF_DEVELOPMENTAL_PROCESS	GO:0051093	0.025	0.18542336	59
ENZYME_LINKED_RECEPTOR_PROTEIN_SIGNALING_PATHWAY	GO:0007167	0.028	0.18664029	52
REGULATION_OF_PROTEIN_KINASE_ACTIVITY	GO:0045859	0.029	0.18673187	47
GENERATION_OF_NEURONS	GO:0048699	0.027	0.19535461	36
AMINO_ACID_METABOLIC_PROCESS	GO:0006520	0.04	0.19601429	25
CELL_CELL_ADHESION	GO:0016337	0.037	0.19835915	34
CARBOXYLIC_ACID_TRANSPORT	GO:0046942	0.043	0.20010489	14
REGULATION_OF_KINASE_ACTIVITY	GO:0043549	0.034	0.2018264	47
REGULATION_OF_CELLULAR_COMPONENT_ORGANIZATION_AND_BIOGENESIS	GO:0051128	0.039	0.20316552	36
CELLULAR_HOMEOSTASIS	GO:0019725	0.035	0.20458218	44
AMINE_RECEPTOR_ACTIVITY	GO:0008227	0.033	0.20658107	17
CHROMOSOMAL_PART	GO:0044427	0.044	0.20923841	29
VOLTAGE_GATED_POTASSIUM_CHANNEL_ACTIVITY	GO:0005249	0.043	0.21057333	18
ORGANIC_ACID_TRANSPORT	GO:0015849	0.044	0.21136242	14
PHOSPHORIC_ESTER_HYDROLASE_ACTIVITY	GO:0042578	0.04	0.21392259	51
REGULATION_OF_IMMUNE_SYSTEM_PROCESS	GO:0002682	0.043	0.21415131	22
POSITIVE_REGULATION_OF_TRANSCRIPTION__DNA_DEPENDENT	GO:0045893	0.039	0.21422727	42

IMMUNE_SYSTEM_DEVELOPMENT	GO:0002520	0.044	0.21430065	28
PROTEIN_DNA_COMPLEX_ASSEMBLY	GO:0065004	0.045	0.21537974	13
POSITIVE_REGULATION_OF_TRANSCRIPTION_FROM_RNA_POLYMERASE_II_PROMOTER	GO:0045944	0.04	0.21547772	23
MICROTUBULE_CYTOSKELETON_ORGANIZATION_AND_BIOGENESIS	GO:0000226	0.049	0.21563749	8
BIOPOLYMER_CATABOLIC_PROCESS	GO:0043285	0.048	0.21601887	26
LOCOMOTORY_BEHAVIOR	GO:0007626	0.047	0.21682693	26
NEGATIVE_REGULATION_OF_CATALYTIC_ACTIVITY	GO:0043086	0.043	0.22042684	23
AMINO_ACID_CATABOLIC_PROCESS	GO:0009063	0.043	0.22118181	9
TRANSMEMBRANE_RECEPTOR_PROTEIN_TYROSINE_KINASE_ACTIVITY	GO:0004714	0.043	0.22431764	23
TRANSCRIPTION_REPRESSOR_ACTIVITY	GO:0016564	0.049	0.22598214	47
AMINO_ACID_AND_DERIVATIVE_METABOLIC_PROCESS	GO:0006519	0.047	0.2265988	30

Appendix G: Association rules identified in canonical pathways

In this appendix, those pathways with a considerably large number of rules (> 20) identified in Section 5.1.2.2 are listed. Particularly, the rules for ECM receptor interaction pathway, prostate cancer pathway and the union of all canonical pathways are provided.

Rules for ECM Receptor Interaction Pathway: Cases

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs3212578_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.6542	0.9055	1.0231	9.3746
2	{rs17057233_C_D,rs7537288_G_D} => {rs3212578_A_D}	0.6712	0.8846	1.0204	7.4777
3	{rs7182678_G_D} => {rs12328617_A_D}	0.6132	0.8894	1.0180	4.3695
4	{rs12328617_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.6667	0.8772	1.0173	5.2058
5	{rs12328617_A_D,rs17057233_C_D} => {rs3212578_A_D}	0.6780	0.8817	1.0170	5.5284
6	{rs747546_T_D} => {rs4865534_G_D}	0.6439	0.8144	1.0168	3.7827
7	{rs17479287_A_D,rs7537288_G_D} => {rs3212578_A_D}	0.6223	0.8808	1.0161	3.5647
8	{rs17479287_A_D,rs7537288_G_D} => {rs1627354_A_D}	0.6223	0.8808	1.0161	3.5647
9	{rs9975613_C_D} => {rs1627354_A_D}	0.6177	0.8801	1.0152	3.1119
10	{rs12328617_A_D,rs1627354_A_D} => {rs17479287_A_D}	0.6394	0.8426	1.0146	2.8712
11	{rs12328617_A_D,rs362777_C_D} => {rs1627354_A_D}	0.6052	0.8793	1.0144	2.6049
12	{rs1627354_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.6496	0.8978	1.0144	3.6498
13	{rs1627354_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.6724	0.8968	1.0132	3.5541
14	{rs17057233_C_D,rs17479287_A_D} => {rs3212578_A_D}	0.6394	0.8781	1.0130	2.5729
15	{rs17057233_C_D,rs17479287_A_D} => {rs1627354_A_D}	0.6394	0.8781	1.0130	2.5729
16	{rs1627354_A_D,rs3212578_A_D} => {rs17479287_A_D}	0.6303	0.8407	1.0123	1.9377
17	{rs2839086_T_D} => {rs7537288_G_D}	0.6075	0.8725	1.0118	1.7677
18	{rs3212578_A_D,rs4865534_G_D} => {rs17057233_C_D}	0.6132	0.8953	1.0116	1.9743
19	{rs17057233_C_D} => {rs8041354_T_D}	0.7486	0.8458	1.0115	4.5372
20	{rs17479287_A_D} => {rs1627354_A_D}	0.7281	0.8767	1.0113	3.5977
21	{rs3212578_A_D} => {rs17057233_C_D}	0.7759	0.8950	1.0112	5.5353
22	{rs12328617_A_D,rs4865534_G_D} => {rs1627354_A_D}	0.6121	0.8762	1.0108	1.5355
23	{rs17057233_C_D,rs4865534_G_D} => {rs1627354_A_D}	0.6200	0.8762	1.0107	1.5986
24	{rs6983702_T_D} => {rs17057233_C_D}	0.6155	0.8942	1.0103	1.5867
25	{rs2275843_T_D} => {rs17057233_C_D}	0.6212	0.8936	1.0096	1.4305
26	{rs9890077_C_D} => {rs1627354_A_D}	0.6382	0.8752	1.0096	1.4137
27	{rs362777_C_D} => {rs8041354_T_D}	0.6644	0.8439	1.0093	1.4274
28	{rs7537288_G_D,rs8041354_T_D} => {rs3212578_A_D}	0.6280	0.8748	1.0091	1.2120
29	{rs3212578_A_D} => {rs7537288_G_D}	0.7543	0.8701	1.0090	2.8855

30	{rs1627354_A_D,rs7537288_G_D} => {rs12328617_A_D}	0.6587	0.8813	1.0087	1.3470
31	{rs17057233_C_D,rs747546_T_D} => {rs1627354_A_D}	0.6098	0.8744	1.0086	0.9859
32	{rs362777_C_D} => {rs1627354_A_D}	0.6883	0.8743	1.0085	1.5367
33	{rs17479287_A_D,rs7537288_G_D} => {rs12328617_A_D}	0.6223	0.8808	1.0081	0.9713
34	{rs4865534_G_D} => {rs1627354_A_D}	0.6997	0.8736	1.0077	1.3696
35	{rs12328617_A_D,rs17479287_A_D} => {rs3212578_A_D}	0.6348	0.8732	1.0073	0.8166
36	{rs4865534_G_D,rs7537288_G_D} => {rs12328617_A_D}	0.6041	0.8791	1.0062	0.5138
37	{rs7537288_G_D} => {rs12328617_A_D}	0.7577	0.8786	1.0056	1.2021
38	{rs9890077_C_D} => {rs8041354_T_D}	0.6132	0.8409	1.0056	0.3811
39	{rs17479287_A_D} => {rs3212578_A_D}	0.7235	0.8712	1.0050	0.7026
40	{rs747546_T_D} => {rs1627354_A_D}	0.6883	0.8705	1.0042	0.3749
41	{rs1627354_A_D,rs3212578_A_D} => {rs12328617_A_D}	0.6576	0.8771	1.0039	0.2704
42	{rs12328617_A_D,rs17057233_C_D} => {rs1627354_A_D}	0.6689	0.8698	1.0034	0.2175
43	{rs3212578_A_D} => {rs12328617_A_D}	0.7600	0.8766	1.0033	0.4425
44	{rs747546_T_D} => {rs17479287_A_D}	0.6587	0.8331	1.0031	0.1600
45	{rs17057233_C_D,rs7537288_G_D} => {rs1627354_A_D}	0.6598	0.8696	1.0031	0.1710
46	{rs747546_T_D} => {rs7537288_G_D}	0.6837	0.8647	1.0028	0.1617
47	{rs362777_C_D} => {rs17479287_A_D}	0.6553	0.8324	1.0023	0.0817
48	{rs4865534_G_D} => {rs8041354_T_D}	0.6712	0.8381	1.0023	0.0923
49	{rs9890077_C_D} => {rs7537288_G_D}	0.6303	0.8643	1.0022	0.0744
50	{rs2275843_T_D} => {rs12328617_A_D}	0.6086	0.8756	1.0022	0.0651
51	{rs12328617_A_D,rs8041354_T_D} => {rs1627354_A_D}	0.6325	0.8688	1.0021	0.0703
52	{rs362777_C_D} => {rs7537288_G_D}	0.6803	0.8642	1.0021	0.0906
53	{rs17479287_A_D} => {rs12328617_A_D}	0.7270	0.8753	1.0019	0.1027
54	{rs1627354_A_D} => {rs12328617_A_D}	0.7588	0.8753	1.0018	0.1342
55	{rs9890077_C_D} => {rs17057233_C_D}	0.6462	0.8861	1.0012	0.0242
56	{rs17057233_C_D} => {rs1627354_A_D}	0.7679	0.8676	1.0008	0.0300
57	{rs362777_C_D} => {rs12328617_A_D}	0.6883	0.8743	1.0006	0.0092
58	{rs2275843_T_D} => {rs1627354_A_D}	0.6030	0.8674	1.0006	0.0050

Rules for ECM Receptor Interaction Pathway: Controls

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs1627354_A_D,rs7537288_G_D} => {rs362777_C_D}	0.631	0.848	1.014	3.440
2	{rs1627354_A_D,rs17057233_C_D} => {rs8041354_T_D}	0.610	0.802	1.014	2.489
3	{rs12328617_A_D,rs17057233_C_D} => {rs8041354_T_D}	0.611	0.801	1.012	1.811
4	{rs6983702_T_D} => {rs362777_C_D}	0.618	0.846	1.011	1.985
5	{rs7537288_G_D} => {rs8041354_T_D}	0.655	0.800	1.011	2.159
6	{rs1627354_A_D,rs8041354_T_D} => {rs362777_C_D}	0.610	0.845	1.011	1.638
7	{rs17564993_T_D} => {rs12328617_A_D}	0.627	0.920	1.010	2.592
8	{rs3212578_A_D} => {rs7537288_G_D}	0.679	0.826	1.009	1.940
9	{rs8041354_T_D} => {rs17057233_C_D}	0.670	0.846	1.009	1.689
10	{rs9890077_C_D} => {rs17057233_C_D}	0.645	0.846	1.009	1.392
11	{rs6983702_T_D} => {rs17479287_A_D}	0.646	0.884	1.008	1.436
12	{rs12328617_A_D,rs9890077_C_D} => {rs17479287_A_D}	0.614	0.883	1.007	0.940
13	{rs9975613_C_D} => {rs12328617_A_D}	0.682	0.917	1.007	1.635

14	{rs12328617_A_D,rs7537288_G_D} => {rs362777_C_D}	0.625	0.842	1.007	0.807
15	{rs362777_C_D,rs747546_T_D} => {rs1627354_A_D}	0.629	0.914	1.007	1.170
16	{rs9890077_C_D} => {rs17479287_A_D}	0.673	0.883	1.007	1.217
17	{rs8041354_T_D} => {rs362777_C_D}	0.666	0.842	1.007	0.936
18	{rs6983702_T_D} => {rs4865534_G_D}	0.622	0.851	1.006	0.680
19	{rs7537288_G_D} => {rs362777_C_D}	0.689	0.842	1.006	1.001
20	{rs12328617_A_D,rs3212578_A_D} => {rs17479287_A_D}	0.654	0.882	1.006	0.808
21	{rs3212578_A_D} => {rs17479287_A_D}	0.725	0.881	1.006	1.112
22	{rs362777_C_D,rs4865534_G_D} => {rs1627354_A_D}	0.645	0.913	1.005	0.738
23	{rs362777_C_D} => {rs1627354_A_D}	0.763	0.912	1.005	1.339
24	{rs8041354_T_D} => {rs1627354_A_D}	0.722	0.912	1.004	0.810
25	{rs4865534_G_D,rs7537288_G_D} => {rs1627354_A_D}	0.623	0.911	1.004	0.304
26	{rs9975613_C_D} => {rs17479287_A_D}	0.654	0.880	1.004	0.287
27	{rs8041354_T_D} => {rs3212578_A_D}	0.653	0.825	1.003	0.237
28	{rs9975613_C_D} => {rs1627354_A_D}	0.677	0.911	1.003	0.367
29	{rs6983702_T_D} => {rs747546_T_D}	0.604	0.826	1.003	0.121
30	{rs9975613_C_D} => {rs3212578_A_D}	0.613	0.824	1.003	0.117
31	{rs9890077_C_D} => {rs4865534_G_D}	0.646	0.847	1.003	0.128
32	{rs1627354_A_D,rs17057233_C_D} => {rs4865534_G_D}	0.644	0.847	1.002	0.088
33	{rs2275843_T_D} => {rs1627354_A_D}	0.658	0.910	1.002	0.126
34	{rs1627354_A_D,rs17479287_A_D} => {rs4865534_G_D}	0.672	0.847	1.002	0.066
35	{rs8041354_T_D} => {rs17479287_A_D}	0.695	0.878	1.002	0.077
36	{rs17479287_A_D,rs362777_C_D} => {rs12328617_A_D}	0.667	0.912	1.002	0.076
37	{rs9890077_C_D} => {rs12328617_A_D}	0.695	0.912	1.002	0.088
38	{rs17057233_C_D} => {rs4865534_G_D}	0.710	0.846	1.001	0.067
39	{rs8041354_T_D} => {rs12328617_A_D}	0.722	0.912	1.001	0.089
40	{rs4865534_G_D} => {rs1627354_A_D}	0.768	0.909	1.001	0.092
41	{rs9975613_C_D} => {rs17057233_C_D}	0.624	0.840	1.001	0.022
42	{rs9890077_C_D} => {rs3212578_A_D}	0.627	0.823	1.001	0.014
43	{rs3212578_A_D} => {rs1627354_A_D}	0.747	0.909	1.001	0.032
44	{rs4865534_G_D,rs747546_T_D} => {rs17479287_A_D}	0.607	0.877	1.001	0.011
45	{rs4865534_G_D} => {rs17479287_A_D}	0.742	0.877	1.001	0.026
46	{rs9975613_C_D} => {rs747546_T_D}	0.613	0.824	1.001	0.006
47	{rs7537288_G_D} => {rs1627354_A_D}	0.743	0.908	1.000	0.006

Rules for Prostate Cancer Pathway: Cases

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs10510097_A_D} => {rs42695_A_D}	0.6314	0.8384	1.0137	2.3909
2	{rs11672342_T_D} => {rs16897333_G_D}	0.6257	0.8160	1.0117	1.6458
3	{rs2849379_T_D} => {rs11466212_G_D}	0.6621	0.8622	1.0092	1.4377
4	{rs16897333_G_D,rs4739561_T_D} => {rs11190421_A_D}	0.6098	0.8535	1.0084	0.8497
5	{rs6857523_A_D} => {rs4739561_T_D}	0.6382	0.8976	1.0077	1.0326
6	{rs11466212_G_D,rs4739561_T_D} => {rs17041230_A_D}	0.6667	0.8825	1.0062	0.7337
7	{rs42695_A_D,rs4739561_T_D} => {rs11190421_A_D}	0.6257	0.8514	1.0059	0.4643
8	{rs2849379_T_D} => {rs13116385_T_D}	0.6177	0.8044	1.0058	0.3964

9	{rs17041230_A_D,rs4739561_T_D} => {rs13116385_T_D}	0.6280	0.8035	1.0047	0.2719
10	{rs17041230_A_D,rs2849379_T_D} => {rs4739561_T_D}	0.6007	0.8949	1.0046	0.3147
11	{rs11190421_A_D,rs11466212_G_D} => {rs17041230_A_D}	0.6325	0.8811	1.0046	0.3334
12	{rs17041230_A_D,rs4739561_T_D} => {rs11190421_A_D}	0.6644	0.8501	1.0043	0.3235
13	{rs17041230_A_D,rs42695_A_D} => {rs11190421_A_D}	0.6121	0.8499	1.0041	0.2137
14	{rs16897333_G_D,rs4739561_T_D} => {rs17041230_A_D}	0.6291	0.8806	1.0039	0.2415
15	{rs11466212_G_D,rs16897333_G_D} => {rs17041230_A_D}	0.6018	0.8802	1.0035	0.1658
16	{rs11466212_G_D} => {rs17041230_A_D}	0.7520	0.8802	1.0035	0.4384
17	{rs2849379_T_D} => {rs16897333_G_D}	0.6212	0.8089	1.0028	0.0978
18	{rs11190421_A_D} => {rs42695_A_D}	0.7019	0.8293	1.0027	0.1677
19	{rs11190421_A_D,rs17041230_A_D} => {rs16897333_G_D}	0.6007	0.8086	1.0025	0.0637
20	{rs13116385_T_D} => {rs17041230_A_D}	0.7031	0.8791	1.0022	0.1247
21	{rs11190421_A_D} => {rs4739561_T_D}	0.7554	0.8925	1.0019	0.1419
22	{rs16897333_G_D} => {rs17041230_A_D}	0.7088	0.8787	1.0018	0.0838
23	{rs1291490_T_D} => {rs4739561_T_D}	0.6303	0.8921	1.0015	0.0382
24	{rs16897333_G_D} => {rs11190421_A_D}	0.6837	0.8477	1.0015	0.0445
25	{rs4739561_T_D} => {rs13116385_T_D}	0.7133	0.8008	1.0012	0.0442
26	{rs2849379_T_D} => {rs4739561_T_D}	0.6849	0.8919	1.0012	0.0340
27	{rs1291490_T_D} => {rs11466212_G_D}	0.6041	0.8551	1.0008	0.0082
28	{rs11190421_A_D} => {rs17041230_A_D}	0.7429	0.8777	1.0006	0.0138
29	{rs10510097_A_D} => {rs4739561_T_D}	0.6712	0.8912	1.0005	0.0057
30	{rs17041230_A_D} => {rs4739561_T_D}	0.7816	0.8911	1.0003	0.0046
31	{rs6857523_A_D} => {rs11466212_G_D}	0.6075	0.8544	1.0000	0.0000

Rules for Prostate Cancer Pathway: Controls

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs587230_A_D} => {rs17041230_A_D}	0.6047	0.8471	1.0140	2.7745
2	{rs17041230_A_D,rs42695_A_D} => {rs13116385_T_D}	0.6208	0.8494	1.0135	2.8788
3	{rs13116385_T_D,rs4739561_T_D} => {rs17041230_A_D}	0.6029	0.8457	1.0123	2.1215
4	{rs11190421_A_D,rs11466212_G_D} => {rs17041230_A_D}	0.6020	0.8455	1.0120	2.0331
5	{rs10510097_A_D,rs13116385_T_D} => {rs42695_A_D}	0.6055	0.8861	1.0119	2.4224
6	{rs11672342_T_D} => {rs2849379_T_D}	0.6655	0.8185	1.0111	2.5542
7	{rs11190421_A_D,rs17041230_A_D} => {rs11672342_T_D}	0.6091	0.8215	1.0103	1.4933
8	{rs17041230_A_D} => {rs13116385_T_D}	0.7066	0.8458	1.0092	2.4931
9	{rs13116385_T_D,rs4739561_T_D} => {rs42695_A_D}	0.6297	0.8833	1.0087	1.4888
10	{rs16897333_G_D} => {rs11466212_G_D}	0.6172	0.8156	1.0087	1.1059
11	{rs11466212_G_D,rs4739561_T_D} => {rs42695_A_D}	0.6091	0.8833	1.0087	1.3171
12	{rs2849379_T_D} => {rs10510097_A_D}	0.6646	0.8210	1.0087	1.5551
13	{rs11466212_G_D} => {rs11672342_T_D}	0.6628	0.8197	1.0082	1.3663
14	{rs11466212_G_D} => {rs42695_A_D}	0.7138	0.8827	1.0081	2.1699
15	{rs11466212_G_D} => {rs17041230_A_D}	0.6798	0.8407	1.0063	0.9603
16	{rs17041230_A_D,rs42695_A_D} => {rs4739561_T_D}	0.6288	0.8605	1.0063	0.7054
17	{rs587230_A_D} => {rs11190421_A_D}	0.6360	0.8910	1.0062	0.8221
18	{rs16897333_G_D} => {rs2849379_T_D}	0.6163	0.8144	1.0061	0.5501
19	{rs10510097_A_D,rs4739561_T_D} => {rs42695_A_D}	0.6100	0.8800	1.0049	0.4349

20	{rs7327621_A_D} => {rs42695_A_D}	0.6154	0.8798	1.0047	0.4066
21	{rs587230_A_D} => {rs42695_A_D}	0.6279	0.8797	1.0046	0.4156
22	{rs10510097_A_D,rs11190421_A_D} => {rs42695_A_D}	0.6270	0.8795	1.0044	0.3833
23	{rs2849379_T_D,rs42695_A_D} => {rs4739561_T_D}	0.6038	0.8588	1.0043	0.2893
24	{rs11672342_T_D,rs42695_A_D} => {rs11190421_A_D}	0.6324	0.8893	1.0043	0.3916
25	{rs13116385_T_D} => {rs42695_A_D}	0.7370	0.8794	1.0043	0.7402
26	{rs10510097_A_D} => {rs42695_A_D}	0.7156	0.8791	1.0039	0.5348
27	{rs11190421_A_D,rs4739561_T_D} => {rs17041230_A_D}	0.6324	0.8387	1.0039	0.2634
28	{rs11190421_A_D} => {rs11672342_T_D}	0.7227	0.8162	1.0038	0.5476
29	{rs10510097_A_D} => {rs17041230_A_D}	0.6825	0.8385	1.0036	0.3290
30	{rs11672342_T_D} => {rs17041230_A_D}	0.6816	0.8383	1.0034	0.2900
31	{rs11672342_T_D,rs42695_A_D} => {rs4739561_T_D}	0.6100	0.8579	1.0032	0.1696
32	{rs17041230_A_D} => {rs4739561_T_D}	0.7165	0.8576	1.0029	0.2870
33	{rs16897333_G_D} => {rs13116385_T_D}	0.6360	0.8404	1.0028	0.1381
34	{rs11190421_A_D,rs4739561_T_D} => {rs42695_A_D}	0.6619	0.8778	1.0025	0.1450
35	{rs17041230_A_D} => {rs11190421_A_D}	0.7415	0.8876	1.0023	0.2400
36	{rs42695_A_D} => {rs4739561_T_D}	0.7504	0.8570	1.0022	0.2291
37	{rs10510097_A_D} => {rs13116385_T_D}	0.6834	0.8396	1.0017	0.0765
38	{rs16897333_G_D} => {rs11672342_T_D}	0.6163	0.8144	1.0017	0.0424
39	{rs4739561_T_D} => {rs2849379_T_D}	0.6932	0.8107	1.0015	0.0604
40	{rs2849379_T_D} => {rs17041230_A_D}	0.6771	0.8365	1.0012	0.0376
41	{rs587230_A_D} => {rs4739561_T_D}	0.6109	0.8559	1.0009	0.0141
42	{rs11466212_G_D} => {rs13116385_T_D}	0.6780	0.8385	1.0005	0.0053
43	{rs11466212_G_D} => {rs10510097_A_D}	0.6583	0.8142	1.0003	0.0013
44	{rs16897333_G_D} => {rs42695_A_D}	0.6628	0.8759	1.0002	0.0015

All Canonical Pathways: Cases

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs10510097_A_D,rs17041230_A_D} => {rs17057233_C_D}	0.601	0.913	1.032	13.385
2	{rs3212578_A_D,rs4739561_T_D} => {rs6990501_G_D}	0.647	0.839	1.030	11.947
3	{rs10510097_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.605	0.911	1.029	11.444
4	{rs11628551_T_D,rs16897333_G_D} => {rs2722279_C_D}	0.645	0.872	1.029	11.832
5	{rs10510097_A_D,rs11628551_T_D} => {rs17057233_C_D}	0.626	0.911	1.029	12.343
6	{rs11628551_T_D,rs12328617_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.615	0.887	1.028	10.113
7	{rs17057233_C_D,rs2722279_C_D,rs7537288_G_D} => {rs11628551_T_D}	0.604	0.945	1.028	13.769
8	{rs362777_C_D,rs593241_T_D} => {rs17041230_A_D}	0.612	0.901	1.027	9.978
9	{rs11190421_A_D,rs11628551_T_D} => {rs747546_T_D}	0.629	0.812	1.027	8.342
10	{rs17041230_A_D,rs3212578_A_D,rs4739561_T_D} => {rs17057233_C_D}	0.611	0.909	1.027	9.822
11	{rs17057233_C_D,rs4865534_G_D} => {rs42695_A_D}	0.601	0.849	1.026	7.068
12	{rs11628551_T_D,rs17041230_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.613	0.885	1.026	8.617
13	{rs42695_A_D,rs4739561_T_D} => {rs4865534_G_D}	0.604	0.822	1.026	6.786
14	{rs2722279_C_D,rs7307707_T_D} => {rs11628551_T_D}	0.605	0.943	1.026	12.246
15	{rs4739561_T_D,rs7537288_G_D} => {rs6990501_G_D}	0.637	0.836	1.026	8.431
16	{rs2722279_C_D,rs3798315_A_D} => {rs11628551_T_D}	0.602	0.943	1.026	11.763

17	{rs6469232_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.617	0.908	1.026	9.492
18	{rs17041230_A_D,rs2722279_C_D,rs7537288_G_D} => {rs11628551_T_D}	0.601	0.943	1.026	11.605
19	{rs12450493_A_D,rs4739561_T_D} => {rs4865534_G_D}	0.618	0.821	1.026	7.074
20	{rs11628551_T_D,rs17057233_C_D,rs4739561_T_D} => {rs8041354_T_D}	0.622	0.857	1.025	7.627
21	{rs1627354_A_D,rs3136667_C_D} => {rs12450493_A_D}	0.653	0.874	1.025	9.598
22	{rs17057233_C_D,rs6990501_G_D} => {rs3212578_A_D}	0.644	0.889	1.025	9.395
23	{rs11190421_A_D,rs17041230_A_D} => {rs747546_T_D}	0.602	0.810	1.025	5.797
24	{rs2722279_C_D,rs3136667_C_D} => {rs16897333_G_D}	0.601	0.826	1.024	5.818
25	{rs2569538_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.620	0.883	1.024	7.663
26	{rs11628551_T_D,rs3212578_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.605	0.906	1.024	7.814
27	{rs2722279_C_D,rs3212578_A_D,rs7537288_G_D} => {rs11628551_T_D}	0.601	0.941	1.024	10.062
28	{rs13116385_T_D,rs2722279_C_D} => {rs11628551_T_D}	0.635	0.941	1.024	11.614
29	{rs17041230_A_D,rs7537288_G_D} => {rs2569538_A_D}	0.629	0.829	1.024	6.538
30	{rs12883673_C_D,rs2722279_C_D} => {rs11628551_T_D}	0.614	0.941	1.023	10.337
31	{rs11628551_T_D,rs17057233_C_D,rs7537288_G_D} => {rs3212578_A_D}	0.626	0.887	1.023	7.443
32	{rs3212578_A_D,rs6469232_A_D} => {rs17057233_C_D}	0.634	0.906	1.023	8.540
33	{rs2722279_C_D,rs7537288_G_D} => {rs11628551_T_D}	0.684	0.941	1.023	14.308
34	{rs3212578_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.654	0.906	1.023	9.375
35	{rs2722279_C_D,rs3212578_A_D} => {rs6990501_G_D}	0.614	0.833	1.023	5.752
36	{rs1627354_A_D,rs6990501_G_D} => {rs12450493_A_D}	0.618	0.872	1.023	6.444
37	{rs1627354_A_D,rs4865534_G_D} => {rs12450493_A_D}	0.610	0.872	1.023	6.141
38	{rs17057233_C_D,rs17479287_A_D} => {rs6469232_A_D}	0.601	0.825	1.023	5.109
39	{rs11466212_G_D,rs17041230_A_D} => {rs12450493_A_D}	0.655	0.871	1.023	7.880
40	{rs12450493_A_D,rs17479287_A_D} => {rs1627354_A_D}	0.621	0.886	1.022	6.763
41	{rs11628551_T_D,rs6990501_G_D} => {rs2722279_C_D}	0.643	0.867	1.022	7.062
42	{rs10510097_A_D} => {rs17057233_C_D}	0.681	0.905	1.022	10.271
43	{rs11628551_T_D,rs3212578_A_D,rs4739561_T_D} => {rs17057233_C_D}	0.637	0.905	1.022	7.895
44	{rs6857523_A_D} => {rs12328617_A_D}	0.635	0.893	1.022	7.136
45	{rs4865534_G_D,rs8041354_T_D} => {rs4739561_T_D}	0.611	0.910	1.022	6.931
46	{rs1627354_A_D,rs3212578_A_D,rs4739561_T_D} => {rs17057233_C_D}	0.602	0.904	1.022	6.325
47	{rs11628551_T_D,rs747546_T_D} => {rs2722279_C_D}	0.631	0.866	1.022	6.123
48	{rs17057233_C_D,rs2722279_C_D} => {rs6469232_A_D}	0.618	0.824	1.022	5.097
49	{rs11466212_G_D,rs17479287_A_D} => {rs1627354_A_D}	0.633	0.885	1.021	6.492
50	{rs17041230_A_D,rs17057233_C_D} => {rs6469232_A_D}	0.643	0.824	1.021	5.800
51	{rs12328617_A_D,rs17041230_A_D} => {rs362777_C_D}	0.615	0.804	1.021	4.728
52	{rs12328617_A_D,rs6990501_G_D} => {rs3212578_A_D}	0.631	0.885	1.021	6.329
53	{rs11466212_G_D,rs11628551_T_D} => {rs2722279_C_D}	0.680	0.865	1.021	7.975
54	{rs11190421_A_D,rs4739561_T_D} => {rs4865534_G_D}	0.618	0.818	1.021	4.840
55	{rs1627354_A_D,rs2722279_C_D} => {rs11466212_G_D}	0.637	0.872	1.021	6.128
56	{rs12450493_A_D,rs362777_C_D} => {rs17041230_A_D}	0.604	0.895	1.021	5.672
57	{rs17057233_C_D,rs362777_C_D} => {rs17041230_A_D}	0.623	0.895	1.021	6.255
58	{rs7537288_G_D,rs8080832_C_D} => {rs11628551_T_D}	0.606	0.938	1.021	7.932
59	{rs17057233_C_D,rs2722279_C_D,rs593241_T_D} => {rs11628551_T_D}	0.606	0.938	1.021	7.932
60	{rs13116385_T_D,rs17057233_C_D} => {rs8041354_T_D}	0.603	0.853	1.021	4.614
61	{rs17041230_A_D,rs4739561_T_D} => {rs362777_C_D}	0.628	0.803	1.021	4.950

62	{rs17041230_A_D,rs2722279_C_D,rs593241_T_D} => {rs11628551_T_D}	0.604	0.938	1.021	7.676
63	{rs11628551_T_D,rs3212578_A_D} => {rs7537288_G_D}	0.701	0.880	1.020	9.027
64	{rs2722279_C_D,rs42695_A_D} => {rs11628551_T_D}	0.654	0.938	1.020	9.626
65	{rs17041230_A_D,rs4865534_G_D} => {rs12450493_A_D}	0.606	0.869	1.020	4.861
66	{rs17057233_C_D,rs7537288_G_D} => {rs3212578_A_D}	0.671	0.885	1.020	7.478
67	{rs11628551_T_D,rs17041230_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.626	0.903	1.020	6.333
68	{rs3136667_C_D,rs7537288_G_D} => {rs42695_A_D}	0.621	0.844	1.020	4.848
69	{rs3212578_A_D,rs7537288_G_D} => {rs6990501_G_D}	0.627	0.831	1.020	4.868
70	{rs11672342_T_D,rs17057233_C_D} => {rs3212578_A_D}	0.601	0.884	1.020	4.955
71	{rs8080832_C_D} => {rs16897333_G_D}	0.613	0.823	1.020	4.379
72	{rs11466212_G_D,rs1627354_A_D} => {rs12450493_A_D}	0.651	0.869	1.020	6.141
73	{rs11466212_G_D,rs17057233_C_D} => {rs8041354_T_D}	0.641	0.853	1.020	5.492
74	{rs11628551_T_D,rs2569538_A_D} => {rs7537288_G_D}	0.656	0.880	1.020	6.466
75	{rs362777_C_D,rs7537288_G_D} => {rs17041230_A_D}	0.609	0.895	1.020	5.325
76	{rs747546_T_D} => {rs11190421_A_D}	0.683	0.863	1.020	7.289
77	{rs12883673_C_D} => {rs362777_C_D}	0.621	0.803	1.020	4.411
78	{rs10510097_A_D,rs11628551_T_D} => {rs3212578_A_D}	0.608	0.884	1.020	4.957
79	{rs17041230_A_D,rs17057233_C_D,rs2722279_C_D} => {rs11628551_T_D}	0.613	0.937	1.020	7.389
80	{rs2255146_A_D} => {rs4739561_T_D}	0.643	0.908	1.020	6.754
81	{rs12450493_A_D,rs4739561_T_D} => {rs6990501_G_D}	0.625	0.831	1.020	4.516
82	{rs11190421_A_D,rs4739561_T_D} => {rs362777_C_D}	0.606	0.803	1.020	3.870
83	{rs17041230_A_D,rs17479287_A_D} => {rs1627354_A_D}	0.641	0.884	1.020	5.752
84	{rs11628551_T_D,rs12328617_A_D,rs17057233_C_D} => {rs2722279_C_D}	0.608	0.864	1.019	4.398
85	{rs11628551_T_D,rs7307707_T_D} => {rs7537288_G_D}	0.612	0.879	1.019	4.757
86	{rs11628551_T_D,rs6990501_G_D} => {rs7537288_G_D}	0.652	0.879	1.019	5.784
87	{rs11628551_T_D,rs17041230_A_D} => {rs362777_C_D}	0.646	0.802	1.019	4.890
88	{rs11190421_A_D,rs17041230_A_D} => {rs2569538_A_D}	0.613	0.825	1.019	3.919
89	{rs6990501_G_D,rs8041354_T_D} => {rs4739561_T_D}	0.615	0.908	1.019	5.456
90	{rs11628551_T_D,rs362777_C_D} => {rs2722279_C_D}	0.627	0.864	1.019	4.658
91	{rs16897333_G_D,rs3212578_A_D} => {rs2722279_C_D}	0.605	0.864	1.019	4.121
92	{rs17057233_C_D,rs4739561_T_D} => {rs8041354_T_D}	0.675	0.852	1.019	6.119
93	{rs593241_T_D,rs8080832_C_D} => {rs11628551_T_D}	0.605	0.937	1.019	6.541
94	{rs16897333_G_D,rs17057233_C_D} => {rs2722279_C_D}	0.612	0.864	1.019	4.243
95	{rs1627354_A_D,rs17057233_C_D} => {rs12450493_A_D}	0.667	0.868	1.019	5.940
96	{rs1627354_A_D,rs4739561_T_D} => {rs6990501_G_D}	0.638	0.830	1.019	4.548
97	{rs11190421_A_D,rs1627354_A_D} => {rs12450493_A_D}	0.636	0.868	1.019	4.837
98	{rs2839086_T_D} => {rs593241_T_D}	0.613	0.881	1.019	4.462
99	{rs2722279_C_D,rs593241_T_D} => {rs11628551_T_D}	0.686	0.936	1.019	9.497
100	{rs17057233_C_D,rs2722279_C_D} => {rs11628551_T_D}	0.702	0.936	1.019	10.298
101	{rs12328617_A_D,rs2569538_A_D} => {rs7537288_G_D}	0.623	0.878	1.018	4.559
102	{rs16897333_G_D,rs17057233_C_D} => {rs3212578_A_D}	0.626	0.883	1.018	4.705
103	{rs2722279_C_D,rs8041354_T_D} => {rs11628551_T_D}	0.667	0.936	1.018	8.342
104	{rs11628551_T_D,rs17057233_C_D,rs3212578_A_D} => {rs2722279_C_D}	0.617	0.863	1.018	4.091
105	{rs11628551_T_D,rs12328617_A_D,rs17041230_A_D} => {rs7537288_G_D}	0.614	0.878	1.018	4.255
106	{rs3212578_A_D,rs4739561_T_D} => {rs17057233_C_D}	0.695	0.901	1.018	7.541
107	{rs11628551_T_D,rs12450493_A_D,rs17041230_A_D} => {rs1627354_A_D}	0.608	0.883	1.018	4.173

108	{rs11628551_T_D,rs12450493_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.612	0.901	1.018	4.729
109	{rs11628551_T_D,rs17057233_C_D} => {rs6469232_A_D}	0.669	0.821	1.018	5.298
110	{rs17057233_C_D,rs3136667_C_D} => {rs6469232_A_D}	0.627	0.821	1.018	3.856
111	{rs7182678_G_D} => {rs12328617_A_D}	0.613	0.889	1.018	4.369
112	{rs3212578_A_D,rs593241_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.613	0.936	1.018	6.155
113	{rs17041230_A_D,rs6990501_G_D} => {rs12450493_A_D}	0.618	0.867	1.018	4.044
114	{rs11628551_T_D,rs17057233_C_D,rs593241_T_D} => {rs3212578_A_D}	0.623	0.882	1.018	4.436
115	{rs11628551_T_D,rs17041230_A_D,rs3212578_A_D} => {rs593241_T_D}	0.610	0.880	1.018	4.076
116	{rs12883673_C_D,rs1627354_A_D} => {rs11628551_T_D}	0.629	0.936	1.018	6.595
117	{rs1627354_A_D,rs17041230_A_D} => {rs12450493_A_D}	0.662	0.867	1.018	5.238
118	{rs4739561_T_D,rs593241_T_D} => {rs6990501_G_D}	0.635	0.829	1.018	4.031
119	{rs11628551_T_D,rs11672342_T_D} => {rs593241_T_D}	0.618	0.880	1.018	4.219
120	{rs17057233_C_D,rs4739561_T_D} => {rs6990501_G_D}	0.656	0.829	1.018	4.628
121	{rs17057233_C_D,rs593241_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.611	0.936	1.018	5.929
122	{rs593241_T_D,rs8041354_T_D} => {rs11628551_T_D}	0.677	0.936	1.018	8.241
123	{rs12328617_A_D,rs17057233_C_D,rs4739561_T_D} => {rs3212578_A_D}	0.605	0.882	1.018	3.927
124	{rs1627354_A_D,rs8041354_T_D} => {rs11466212_G_D}	0.629	0.869	1.018	4.226
125	{rs10510097_A_D} => {rs3212578_A_D}	0.664	0.882	1.018	5.427
126	{rs11466212_G_D,rs4739561_T_D} => {rs8041354_T_D}	0.643	0.851	1.018	4.298
127	{rs12883673_C_D,rs7537288_G_D} => {rs11628551_T_D}	0.626	0.935	1.018	6.237
128	{rs17057233_C_D,rs593241_T_D} => {rs3212578_A_D}	0.672	0.882	1.018	5.639
129	{rs2722279_C_D,rs6469232_A_D} => {rs11628551_T_D}	0.642	0.935	1.018	6.702
130	{rs12883673_C_D,rs3136667_C_D} => {rs11628551_T_D}	0.625	0.935	1.017	6.121
131	{rs11628551_T_D,rs17041230_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.608	0.901	1.017	4.255
132	{rs11466212_G_D,rs6990501_G_D} => {rs1627354_A_D}	0.612	0.882	1.017	3.925
133	{rs17041230_A_D,rs3212578_A_D} => {rs17057233_C_D}	0.679	0.900	1.017	6.256
134	{rs11628551_T_D,rs12328617_A_D,rs17057233_C_D} => {rs3212578_A_D}	0.620	0.882	1.017	4.049
135	{rs12328617_A_D,rs3212578_A_D} => {rs7537288_G_D}	0.667	0.877	1.017	5.206
136	{rs12883673_C_D,rs17041230_A_D} => {rs11628551_T_D}	0.639	0.935	1.017	6.455
137	{rs2255146_A_D} => {rs17041230_A_D}	0.631	0.892	1.017	4.531
138	{rs12883673_C_D,rs593241_T_D} => {rs11628551_T_D}	0.622	0.935	1.017	5.893
139	{rs11628551_T_D,rs17057233_C_D} => {rs8041354_T_D}	0.693	0.851	1.017	5.830
140	{rs12883673_C_D,rs17057233_C_D} => {rs11628551_T_D}	0.638	0.935	1.017	6.334
141	{rs17041230_A_D,rs4739561_T_D} => {rs2569538_A_D}	0.644	0.824	1.017	3.926
142	{rs12328617_A_D,rs17057233_C_D} => {rs3212578_A_D}	0.678	0.882	1.017	5.528
143	{rs11628551_T_D,rs3136667_C_D,rs3212578_A_D} => {rs17057233_C_D}	0.615	0.900	1.017	4.243
144	{rs17041230_A_D,rs6990501_G_D} => {rs7537288_G_D}	0.625	0.877	1.017	3.934
145	{rs4739561_T_D,rs6469232_A_D} => {rs17057233_C_D}	0.645	0.900	1.017	4.857
146	{rs11628551_T_D,rs12328617_A_D} => {rs2722279_C_D}	0.688	0.862	1.017	5.494
147	{rs16897333_G_D} => {rs2722279_C_D}	0.695	0.862	1.017	5.740
148	{rs1627354_A_D,rs2722279_C_D} => {rs11628551_T_D}	0.683	0.935	1.017	7.559
149	{rs17041230_A_D,rs593241_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.617	0.934	1.017	5.345
150	{rs17041230_A_D} => {rs362777_C_D}	0.702	0.800	1.017	6.333
151	{rs42695_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.622	0.900	1.016	4.119
152	{rs1627354_A_D,rs7537288_G_D} => {rs12450493_A_D}	0.647	0.866	1.016	4.019

153	{rs11628551_T_D,rs3212578_A_D} => {rs2722279_C_D}	0.686	0.861	1.016	5.121
154	{rs593241_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.695	0.934	1.016	7.768
155	{rs13116385_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.646	0.934	1.016	5.963
156	{rs11466212_G_D,rs12883673_C_D} => {rs11628551_T_D}	0.613	0.934	1.016	5.032
157	{rs11628551_T_D,rs12450493_A_D} => {rs1627354_A_D}	0.689	0.881	1.016	5.317
158	{rs11466212_G_D,rs7537288_G_D} => {rs1627354_A_D}	0.647	0.881	1.016	4.087
159	{rs593241_T_D,rs7537288_G_D} => {rs3212578_A_D}	0.655	0.881	1.016	4.241
160	{rs11466212_G_D,rs11628551_T_D} => {rs8041354_T_D}	0.668	0.849	1.016	4.182
161	{rs17041230_A_D,rs2722279_C_D} => {rs11628551_T_D}	0.691	0.934	1.016	7.182
162	{rs1627354_A_D,rs8080832_C_D} => {rs11628551_T_D}	0.609	0.934	1.016	4.633
163	{rs11466212_G_D,rs17041230_A_D} => {rs1627354_A_D}	0.662	0.880	1.016	4.266
164	{rs17041230_A_D,rs8041354_T_D} => {rs17057233_C_D}	0.658	0.899	1.016	4.494
165	{rs3798315_A_D,rs7537288_G_D} => {rs11628551_T_D}	0.608	0.934	1.016	4.536
166	{rs1627354_A_D,rs3212578_A_D,rs7537288_G_D} => {rs11628551_T_D}	0.608	0.934	1.016	4.536
167	{rs12883673_C_D,rs3212578_A_D} => {rs11628551_T_D}	0.623	0.934	1.016	4.890
168	{rs11628551_T_D,rs6990501_G_D} => {rs3212578_A_D}	0.653	0.880	1.016	3.972
169	{rs11628551_T_D,rs4739561_T_D} => {rs8041354_T_D}	0.692	0.849	1.016	4.752
170	{rs2722279_C_D,rs2849379_T_D} => {rs11628551_T_D}	0.606	0.933	1.015	4.440
171	{rs3212578_A_D,rs42695_A_D} => {rs17057233_C_D}	0.646	0.899	1.015	4.113
172	{rs11190421_A_D,rs11628551_T_D} => {rs2722279_C_D}	0.667	0.860	1.015	3.921
173	{rs17057233_C_D,rs2722279_C_D} => {rs3212578_A_D}	0.660	0.880	1.015	3.992
174	{rs12450493_A_D} => {rs1627354_A_D}	0.750	0.880	1.015	7.355
175	{rs2722279_C_D} => {rs11628551_T_D}	0.791	0.933	1.015	12.281
176	{rs13116385_T_D,rs593241_T_D} => {rs11628551_T_D}	0.647	0.933	1.015	4.937
177	{rs8080832_C_D} => {rs4739561_T_D}	0.673	0.904	1.015	4.487
178	{rs12450493_A_D,rs12883673_C_D} => {rs11628551_T_D}	0.613	0.933	1.014	4.021
179	{rs11628551_T_D,rs17041230_A_D} => {rs593241_T_D}	0.706	0.877	1.014	4.858
180	{rs12883673_C_D} => {rs11628551_T_D}	0.721	0.932	1.014	6.970
181	{rs6469232_A_D,rs7537288_G_D} => {rs11628551_T_D}	0.643	0.932	1.014	4.521
182	{rs11628551_T_D,rs12328617_A_D} => {rs7537288_G_D}	0.699	0.875	1.014	4.443
183	{rs3136667_C_D,rs3212578_A_D} => {rs17057233_C_D}	0.669	0.898	1.014	4.021
184	{rs17057233_C_D,rs3798315_A_D} => {rs11628551_T_D}	0.611	0.932	1.014	3.842
185	{rs11628551_T_D,rs17041230_A_D} => {rs7537288_G_D}	0.704	0.874	1.014	4.378
186	{rs593241_T_D,rs747546_T_D} => {rs11628551_T_D}	0.638	0.932	1.014	4.128
187	{rs7537288_G_D,rs8041354_T_D} => {rs11628551_T_D}	0.669	0.932	1.014	4.803
188	{rs747546_T_D,rs7537288_G_D} => {rs11628551_T_D}	0.637	0.932	1.014	4.034
189	{rs11628551_T_D,rs3212578_A_D} => {rs17057233_C_D}	0.714	0.897	1.014	4.905
190	{rs6469232_A_D} => {rs17057233_C_D}	0.724	0.897	1.013	5.140
191	{rs3212578_A_D} => {rs6990501_G_D}	0.716	0.825	1.013	4.501

All Canonical Pathways: Controls

Rank	Rule	Supp.	Conf.	Lift	χ^2
1	{rs12450493_A_D,rs7537288_G_D} => {rs4739561_T_D}	0.634	0.875	1.024	9.692
2	{rs11628551_T_D,rs8041354_T_D} => {rs2569538_A_D}	0.602	0.863	1.023	7.317
3	{rs11190421_A_D,rs11628551_T_D,rs4739561_T_D} => {rs12450493_A_D}	0.602	0.905	1.023	8.676

4	{rs12450493_A_D,rs4739561_T_D} => {rs8041354_T_D}	0.615	0.809	1.023	6.825
5	{rs17479287_A_D,rs362777_C_D} => {rs10510097_A_D}	0.608	0.831	1.021	6.054
6	{rs6469232_A_D} => {rs593241_T_D}	0.635	0.829	1.021	7.145
7	{rs11190421_A_D,rs17041230_A_D} => {rs3798315_A_D}	0.606	0.817	1.021	5.788
8	{rs10510097_A_D,rs11628551_T_D} => {rs593241_T_D}	0.600	0.829	1.021	5.712
9	{rs4733616_T_D} => {rs2569538_A_D}	0.622	0.861	1.021	6.918
10	{rs12328617_A_D,rs7537288_G_D} => {rs2849379_T_D}	0.614	0.827	1.021	6.055
11	{rs11190421_A_D,rs11628551_T_D,rs42695_A_D} => {rs12450493_A_D}	0.623	0.903	1.021	8.126
12	{rs12328617_A_D,rs2569538_A_D} => {rs2849379_T_D}	0.633	0.826	1.021	6.604
13	{rs2275843_T_D} => {rs10510097_A_D}	0.601	0.831	1.021	5.392
14	{rs11190421_A_D,rs1627354_A_D,rs42695_A_D} => {rs12450493_A_D}	0.639	0.903	1.020	8.622
15	{rs4949184_A_D} => {rs42695_A_D}	0.600	0.893	1.020	6.661
16	{rs11190421_A_D,rs2569538_A_D} => {rs3798315_A_D}	0.610	0.816	1.020	5.394
17	{rs12450493_A_D,rs1627354_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.614	0.904	1.020	7.254
18	{rs362777_C_D,rs42695_A_D} => {rs10510097_A_D}	0.606	0.830	1.020	5.074
19	{rs12450493_A_D,rs1627354_A_D,rs17041230_A_D} => {rs11190421_A_D}	0.606	0.903	1.020	6.743
20	{rs12450493_A_D,rs9890077_C_D} => {rs11628551_T_D}	0.612	0.904	1.019	6.858
21	{rs11672342_T_D,rs12328617_A_D} => {rs2849379_T_D}	0.611	0.825	1.019	4.907
22	{rs3136667_C_D,rs42695_A_D} => {rs10510097_A_D}	0.600	0.829	1.019	4.623
23	{rs11190421_A_D,rs12450493_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.601	0.903	1.019	6.250
24	{rs11190421_A_D,rs17057233_C_D} => {rs3798315_A_D}	0.602	0.815	1.019	4.517
25	{rs11628551_T_D,rs12328617_A_D,rs4739561_T_D} => {rs12450493_A_D}	0.613	0.901	1.019	6.485
26	{rs17041230_A_D,rs17479287_A_D} => {rs13116385_T_D}	0.627	0.854	1.019	5.638
27	{rs1627354_A_D,rs3798315_A_D} => {rs11190421_A_D}	0.659	0.902	1.019	8.219
28	{rs1937834_C_D} => {rs362777_C_D}	0.607	0.852	1.019	4.954
29	{rs10510097_A_D,rs11190421_A_D} => {rs362777_C_D}	0.607	0.852	1.019	4.954
30	{rs1627354_A_D,rs2569538_A_D} => {rs2849379_T_D}	0.631	0.825	1.019	5.359
31	{rs11628551_T_D,rs12450493_A_D,rs1627354_A_D} => {rs11190421_A_D}	0.650	0.902	1.019	7.733
32	{rs4513489_G_D} => {rs4865534_G_D}	0.631	0.861	1.019	5.810
33	{rs11190421_A_D,rs11628551_T_D,rs12328617_A_D} => {rs12450493_A_D}	0.643	0.901	1.019	7.335
34	{rs11190421_A_D,rs1627354_A_D,rs4739561_T_D} => {rs12450493_A_D}	0.618	0.901	1.018	6.357
35	{rs11628551_T_D,rs362777_C_D} => {rs10510097_A_D}	0.615	0.829	1.018	4.763
36	{rs11190421_A_D,rs12450493_A_D,rs362777_C_D} => {rs1627354_A_D}	0.603	0.925	1.018	6.972
37	{rs12883673_C_D,rs4865534_G_D} => {rs4739561_T_D}	0.608	0.871	1.018	5.076
38	{rs1627354_A_D,rs2569538_A_D} => {rs3798315_A_D}	0.623	0.814	1.018	4.699
39	{rs12450493_A_D,rs2569538_A_D} => {rs8041354_T_D}	0.601	0.806	1.018	3.992
40	{rs12450493_A_D,rs1627354_A_D,rs4865534_G_D} => {rs11190421_A_D}	0.613	0.901	1.018	5.849
41	{rs17057233_C_D,rs4865534_G_D} => {rs4739561_T_D}	0.618	0.870	1.018	5.094
42	{rs11466212_G_D,rs12450493_A_D} => {rs11628551_T_D}	0.643	0.902	1.018	6.820
43	{rs11628551_T_D,rs12328617_A_D,rs42695_A_D} => {rs12450493_A_D}	0.638	0.900	1.018	6.507
44	{rs13116385_T_D,rs4739561_T_D} => {rs12883673_C_D}	0.600	0.842	1.018	4.107
45	{rs1627354_A_D,rs17041230_A_D} => {rs3798315_A_D}	0.617	0.814	1.018	4.318
46	{rs3212578_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.625	0.902	1.018	6.052
47	{rs3136667_C_D,rs4739561_T_D} => {rs17057233_C_D}	0.600	0.854	1.018	4.229

48	{rs12450493_A_D,rs3798315_A_D} => {rs11190421_A_D}	0.635	0.901	1.018	6.331
49	{rs11190421_A_D,rs4739561_T_D} => {rs7537288_G_D}	0.628	0.833	1.017	4.726
50	{rs12450493_A_D,rs9890077_C_D} => {rs11190421_A_D}	0.610	0.901	1.017	5.495
51	{rs12450493_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.674	0.902	1.017	7.761
52	{rs12450493_A_D,rs12883673_C_D} => {rs4739561_T_D}	0.634	0.870	1.017	5.345
53	{rs1627354_A_D,rs17041230_A_D} => {rs3136667_C_D}	0.638	0.841	1.017	5.025
54	{rs11628551_T_D,rs1627354_A_D,rs2722279_C_D} => {rs4865534_G_D}	0.609	0.860	1.017	4.420
55	{rs1627354_A_D,rs4739561_T_D} => {rs8041354_T_D}	0.624	0.805	1.017	4.255
56	{rs12450493_A_D,rs3212578_A_D} => {rs11628551_T_D}	0.653	0.901	1.017	6.422
57	{rs1627354_A_D,rs6469232_A_D} => {rs4865534_G_D}	0.601	0.859	1.017	3.941
58	{rs6469232_A_D} => {rs12883673_C_D}	0.644	0.841	1.017	4.836
59	{rs12450493_A_D,rs1627354_A_D} => {rs11190421_A_D}	0.725	0.900	1.016	9.762
60	{rs2722279_C_D,rs4739561_T_D} => {rs4865534_G_D}	0.649	0.859	1.016	5.121
61	{rs12450493_A_D,rs593241_T_D} => {rs11628551_T_D}	0.651	0.901	1.016	6.157
62	{rs10510097_A_D,rs1627354_A_D} => {rs362777_C_D}	0.629	0.850	1.016	4.385
63	{rs11190421_A_D,rs1627354_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.618	0.901	1.016	5.107
64	{rs11628551_T_D,rs6990501_G_D} => {rs12450493_A_D}	0.614	0.899	1.016	4.927
65	{rs4865534_G_D,rs593241_T_D} => {rs2722279_C_D}	0.614	0.895	1.016	4.814
66	{rs12450493_A_D,rs17479287_A_D} => {rs593241_T_D}	0.639	0.825	1.016	4.412
67	{rs17479287_A_D,rs3136667_C_D} => {rs17041230_A_D}	0.614	0.849	1.016	3.914
68	{rs3798315_A_D,rs4865534_G_D} => {rs2722279_C_D}	0.606	0.894	1.016	4.466
69	{rs11190421_A_D,rs1627354_A_D,rs3212578_A_D} => {rs11628551_T_D}	0.600	0.901	1.016	4.514
70	{rs11628551_T_D,rs8041354_T_D} => {rs12450493_A_D}	0.627	0.899	1.016	5.027
71	{rs11672342_T_D,rs4865534_G_D} => {rs11628551_T_D}	0.615	0.901	1.016	4.777
72	{rs6469232_A_D} => {rs4865534_G_D}	0.657	0.859	1.016	5.004
73	{rs11190421_A_D,rs12328617_A_D,rs42695_A_D} => {rs12450493_A_D}	0.634	0.899	1.016	5.141
74	{rs11466212_G_D,rs4865534_G_D} => {rs11628551_T_D}	0.614	0.900	1.016	4.670
75	{rs10510097_A_D} => {rs362777_C_D}	0.691	0.849	1.016	6.165
76	{rs1627354_A_D,rs17479287_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.605	0.900	1.015	4.282
77	{rs11190421_A_D,rs11628551_T_D,rs17479287_A_D} => {rs12450493_A_D}	0.616	0.898	1.015	4.486
78	{rs2722279_C_D,rs7537288_G_D} => {rs4739561_T_D}	0.625	0.868	1.015	4.059
79	{rs11628551_T_D,rs1627354_A_D,rs4739561_T_D} => {rs12450493_A_D}	0.615	0.898	1.015	4.381
80	{rs4865534_G_D,rs593241_T_D} => {rs11628551_T_D}	0.618	0.900	1.015	4.333
81	{rs11628551_T_D,rs4739561_T_D} => {rs12450493_A_D}	0.676	0.898	1.015	5.864
82	{rs12883673_C_D,rs42695_A_D} => {rs4739561_T_D}	0.629	0.868	1.015	3.890
83	{rs1627354_A_D,rs9890077_C_D} => {rs11190421_A_D}	0.619	0.899	1.015	4.246
84	{rs7327621_A_D} => {rs12450493_A_D}	0.628	0.898	1.015	4.363
85	{rs11190421_A_D,rs12328617_A_D,rs1627354_A_D} => {rs12450493_A_D}	0.659	0.898	1.015	5.173
86	{rs11190421_A_D,rs11628551_T_D} => {rs12450493_A_D}	0.706	0.898	1.015	6.803
87	{rs11466212_G_D} => {rs3212578_A_D}	0.674	0.834	1.015	4.700
88	{rs11628551_T_D,rs2722279_C_D} => {rs4865534_G_D}	0.668	0.858	1.015	4.616
89	{rs11190421_A_D,rs6469232_A_D} => {rs1627354_A_D}	0.627	0.921	1.015	5.029
90	{rs11190421_A_D,rs16897333_G_D} => {rs1627354_A_D}	0.616	0.921	1.015	4.745
91	{rs11190421_A_D,rs6990501_G_D} => {rs12450493_A_D}	0.611	0.898	1.015	3.878
92	{rs11628551_T_D,rs12328617_A_D,rs17479287_A_D} => {rs12450493_A_D}	0.633	0.897	1.014	4.252

93	{rs11628551_T_D,rs12328617_A_D,rs1627354_A_D} => {rs12450493_A_D}	0.657	0.897	1.014	4.813
94	{rs12328617_A_D,rs1627354_A_D} => {rs2849379_T_D}	0.677	0.821	1.014	4.557
95	{rs11190421_A_D,rs11628551_T_D,rs17041230_A_D} => {rs1627354_A_D}	0.603	0.921	1.014	4.214
96	{rs11190421_A_D,rs12450493_A_D,rs17057233_C_D} => {rs1627354_A_D}	0.602	0.921	1.014	4.127
97	{rs2569538_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.644	0.899	1.014	4.364
98	{rs11628551_T_D,rs12328617_A_D} => {rs12450493_A_D}	0.722	0.897	1.014	6.567
99	{rs7537288_G_D} => {rs4739561_T_D}	0.709	0.867	1.014	5.443
100	{rs12450493_A_D,rs17057233_C_D} => {rs11628551_T_D}	0.663	0.898	1.013	4.337
101	{rs12450493_A_D,rs2569538_A_D} => {rs11628551_T_D}	0.670	0.898	1.013	4.446
102	{rs3136667_C_D} => {rs17041230_A_D}	0.699	0.846	1.013	4.601
103	{rs2849379_T_D,rs42695_A_D} => {rs12328617_A_D}	0.648	0.922	1.013	4.554
104	{rs11190421_A_D,rs42695_A_D} => {rs12450493_A_D}	0.694	0.896	1.013	4.942
105	{rs6469232_A_D} => {rs2722279_C_D}	0.682	0.891	1.013	4.353
106	{rs1627354_A_D,rs4865534_G_D} => {rs11628551_T_D}	0.690	0.898	1.013	4.579
107	{rs11190421_A_D,rs4739561_T_D} => {rs12450493_A_D}	0.675	0.896	1.012	4.059
108	{rs11628551_T_D,rs42695_A_D} => {rs12450493_A_D}	0.698	0.896	1.012	4.595
109	{rs11628551_T_D,rs3798315_A_D} => {rs1627354_A_D}	0.649	0.919	1.012	3.978
110	{rs11190421_A_D,rs7537288_G_D} => {rs1627354_A_D}	0.665	0.919	1.012	3.985
111	{rs11190421_A_D,rs11628551_T_D} => {rs1627354_A_D}	0.722	0.918	1.011	5.132
112	{rs11190421_A_D,rs362777_C_D} => {rs1627354_A_D}	0.681	0.918	1.011	3.913