RaSaR: A Novel Methodology for the Detection of Epistasis

Jade JK Hind

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

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Complex diseases which affect a large proportion of our population today demand more strategic methods to produce significant association results. As it currently stands there are numerous disorders and diseases which are yet to be identified with a genetic causal variant despite evidence produced by research efforts which indicate the existence of high genetic concordance. Breast Cancer is one of the most prominent cancers in the female population with approximately 55K new cases each year in the UK and approximately 11K deaths.

The genetic component of Breast Cancer is a popular research area and has uncovered many genetic associations from high to low penetrance. The dataset used within this research is obtained from the DRIVE project, one of five introduced under the GAME-ON initiative. The general research use DRIVE dataset contains approximately 533K single-nucleotide polymorphisms (SNPs), with more than 280K sequenced with reference to the 5 most prominent cancers; colon, breast, ovarian, prostate and lung. SNP's are sequenced for approximately 28K subjects, of which approximately 14K were diagnosed with one of three stages of Breast Cancer; unknown, in-situ and invasive.

Epistasis is a progressive approach that complements the 'common disease, common variant' hypothesis that highlights the potential for connected networks of genetic variants collaborating to produce a phenotypic expression. Epistasis is commonly performed as a pairwise or limitless-arity capacity that considers variant networks as either variant vs variant or as high order interactions. This type of analysis extends the number of tests that were previously performed in a standard approach such as GWAS, in which FDR was already an issue, therefore by multiplying the number of tests up to a factorial rate also increases the issue of FDR.

Further to this, epistasis introduces its own limitations of computational complexity that are generated based on the analysis performed; to consider the most intense approach, a multivariate analysis introduces a time complexity of $O(n!)$. Throughout this thesis, approaches, methods and techniques for epistasis analysis and GWAS are discussed, as well as the limitations that exist and how to address these issues.

Proposed in this thesis is a novel methodology, methodology and methods for the detection of epistasis using interpretable methods and best practice to outline interactions through filtering processes. RaSaR refers to process of Random Sampling Regularisation which randomly splits and produces sample sets to conduct a voting system to regularise the significance and reliability of biological markers, SNPs. Parallel to this, the proposed methodology takes into consideration and adjusts for the common limitations of computational complexity and false discovery using filter selection and a novel method to association analysis.

Preliminary results are promising, outlining a concise detection of interactions using benchmarking standard approaches that consider the common approaches to multiple testing. Results for the detection of epistasis, in the classification of breast cancer patients, indicated nine outlined risk candidate interactions from five variants and a singular candidate variant with high protective association.

ACKNOWLEDGEMENTS

As a notably challenging life period producing one of the most ambitious pieces of academic work, it is with great sincerity that I give thanks to following people who have all participated in fulfilling basic, significant and crucial areas of my life:

To my Director of Studies, Prof. Paulo Lisboa, I extend my gratitude and thanks for the boundless knowledge that inspired, animated and progressed this research. As a distinguished and respected Professor, I can honestly express my awareness of the fortunate opportunity to study under such an intelligent and experienced supervisor.

To my second and third supervisors, Prof. Abir Hussain and Prof. Dhiya Al-Jumeily, whose continued support throughout my higher education academic life and career has led to considerable developments in my capabilities and confidence, for which I will always be appreciative. To further extend this acknowledgement, the respect, patience and faith that you have shown me throughout this process has been nothing short of inspiring.

To my colleagues and friends, Basma and Aday, the expression camaraderie comes to mind. To only mention the support, trust, and kindness that you have provided would be a crime. With encouragement, entertainment and compassion, my endurance and excitement in this work has been sustained, for which I am deeply grateful for. I was extremely lucky to have been provided with such enthusiastic, determined and modest colleagues. I give a special thanks to one of my most logical, intelligent and patient friends, Sam, whose conversations have always inspired critical thought and creativity, leaving me a little more knowledgeable every time. I further extend this list to include one of my oldest, most capable and engaging friends, Rose, thank you for the years gone and the years to go.

To my family. My loving, supportive, entertaining and inimitable family. My gratitude on this page will never been enough to convey how much you have and will motivate me. To my stepdad and stepbrother, I am lucky to have such practical and persevering influences in my life. To my auntie and uncle, you still and always will remain an aspiring guidance for my future, thank you for everything you have done for me. To my cousins, thank you for being the sisters I never asked for but am so grateful to have. To a mother, like no other mother, I wouldn’t be the person I am today without an eccentric, diligent and strong role model like you.

And finally, to my best-friend and partner. There are no words that can explain how much I appreciate everything you do for me. You are the solid foundation in my life that keeps me grounded, sane and assured in the person I am.

Thank you all for being exceptional.

GLOSSARY

Allele	A variant within a locus
Chromosome	The structure of DNA distributed across, most commonly, 23 pairs of chromosomes.
Computational Complexity	The amount of resources required to run a task.
DNA	Contains the instructions required to produce the functional elements of an organism.
False Discovery Rate	The rate of type 1 and type 2 errors.
Familywise Error Rate	The probability of at least one type 1 error occurring.
Gene	Commonly contains the instructions for the production of a segment or full protein.
Genotype	Genetic material that corresponds to a physical expression/ phenotype.
Haplotype	A group of alleles that are inherited in blocks
Heterozygous	A genotype expression that contains different nucleotide expressions e.g. Aa
Homozygous	A genotype expression that contains the same nucleotide expression e.g. AA or aa
Incidence	The percentage of the sample population that have a SNP/s state.
Linkage Disequilibrium	The non-random correlation between nearby alleles within the same chromosome
Locus	The location of a variant/nucleotide
Lymph nodes	A component of the body's immune system
Major Allele	The allele occurring most frequently in the sample population
Minor Allele	The allele occurring least frequently in the sample population
Nucleotide	A single building block of DNA; 1 molecule of sugar, 1 molecule of phosphoric acid and 1 pyrimidine/purine (Adenine, Guanine, Cytosine or Thymine).
OR	A measure of association between outcome and exposure.
Penetrance	The frequency of the population that are affected by the phenotype, who also carry the SNP/s state.
Phenotype	Physical expression of genotype
Population Stratification/ Structure	Systematic difference in allele frequencies that commonly occur between sub-populations due to effects such as ancestry.
Protein	Polypeptide chains of amino acid that are used for the structure function and regulation of the body's organs and tissue.

RR	The ratio of the probability of an outcome in an exposed group to the probability of an outcome in an unexposed group
Single Nucleotide Polymorphism	A common genetic variation in the human genome.
Somatic cells	A cell of an organism that is not a gamete (reproductive cell)
Time Complexity	The amount of time taken to run an algorithm
Type 1 error	False positive discoveries
Type 2 error	False negative discoveries

ACRONYMS

A, G,C,T	Adenine, Guanine, Cytosine, Thymine
BP	Base-Pair
CHAID	Chi-square Automatic Interaction Detector
CI	Confidence Interval
DNA	Deoxyribonucleic acid
ER	Oestrogen Receptor
FDR	False Discovery Rate
FWER	Familywise Error Rate
GC	Genomic Control
GENO	Genotype Call
GWAS	Genome-Wide Association Study
HWE	Hardy-Weinberg Equilibrium
LD	Linkage Disequilibrium
MAF	Minor Allele Frequency
MIND	Missingness in Individuals
OR	Odd’s Ratio
PCA	Principle Component Analysis
PPP	Petal Plot Policy Confidence Scale
QC	Quality Control
RaSaR	Random Sampling Regularisation
RNA	Ribonucleic acid
RR	Risk Ratio
SNP	Single Nucleotide Polymorphism
X^2	Chi-Squared metric
α	Threshold
μ	Mean
σ	Standard Deviation

Chapter 1: INTRODUCTION

The following chapter provides an overview of the thesis, considering the aims and influences that have inspired the work. Genomics and Bioinformatics are established fields that cater to exploring and solving some of the most prominent and timely research questions. Throughout this thesis, the focus is predominantly based on a bioinformatics approach for the analysis of genetic data. There are many approaches and methods that have been developed to analyse genetic data that consider the complex nature and representation of the data, incorporating techniques and adaptations that account for specific issues and bias that are exclusive to human genetics. The limitations of these approaches and methods open areas for improvement, while some issues are specific the methods, others are widespread. To outline one of the most modern approaches, epistasis responds to the phenomenon of systems or networks of genetic components interconnecting to produce a phenotypic response, this also coordinates with the hypothesis of 'common disease, common variant' [1], a term applicable to the theory that interactions of common variants cause common disease.

1.1 PROBLEM STATEMENT

Breast cancer is a complex disease; multifactorial effects represent the phenotypic response of the subject. While there is currently an abundance of techniques for the analysis of genetic data, there is still a limited contribution of reproducible genetic signals that provide evidence of association with sporadic breast cancer, which is estimated to encompass 66% of breast cancer cases [2]. This study will investigate the interactions that exist in subjects associated to breast neoplasms in invasive breast cancer. By considering a representative set of SNPs from the genome, further analysis can be conducted from genome-wide analysis to suggest potential SNPs for interactions.

Current efforts in breast cancer have led to early screening, with great successes in reducing the number of advanced cases [3][4]. Further to this, the introduction of genetic knowledge also outlines patients and their family for potential susceptibility to cancer through examples of the BRCA1 and BRCA2 gene [5][6]. These measures provide a preventative outlook for patient health, a leading direction to personalised medicine that will address patients on an individual basis taking into consideration factors such as genetic make-up[6]. The identification of these SNPs could lead to the classification of susceptible breast cancer patients and potentially the pharmacological or therapy treatments that are most suitable for these individuals[7].

The focus phenotype of this research is breast cancer due to the outlined genetic link in previous research that suggests a broad association to genetic components; the focus phenotype can be further defined to sporadic breast cancer which concerns the development of breast cancer outside of familial genetic causation [5][8][9]. Breast cancer is an internationally recognised issue that concludes one of the most prominent mortality rates for cancer in women for 2016 [10]. As such, research efforts for this disease have continually increased over the past decade, a focus of this on genetics, resulting in large datasets of information for large sample sizes. Particularly in this research, a dataset of ~28K (pre-qc) subjects and ~500K SNPs that incorporated genotyped data from not only a common backbone of SNP's (~280K) but additionally a collection of variants that are specific to the 5 most prominent cancers; breast, ovarian, prostate, colon and lung[11].

There are current issues in bioinformatics that can result in misinterpretation or misunderstanding due to 'black box' methods. Therefore, an additional consideration of this work will be influenced by methods that present interpretable options for processes throughout the methodology. Progressive approach, epistasis, invites new avenues of research[12]. The phenomenon that suggests combinations of biological material such as SNP variants are working as a system or network to produce the phenotypic outcome is becoming a favourable lead. Given the elusiveness of genetic causation in the face of high heritability, epistasis suggests an enigmatic genetic component is not being detected, as the signal for networks of SNPs are masking one another. Therefore, the research subject of epistasis invites a new problem area to explore. Current practices in epistasis detection range from pairwise to exhaustive search criteria; the limiting nature of pairwise detection could lead to loss of information by oversight however exhaustive search present their own problems with the demand for computational power [13].

Computational complexity encompasses the most problematic area of limitations for the limitless-arity technique as these analyses can be conducted using linear or parallel threading. These approaches require extensive hardware to accommodate the requirements of the analysis, however this will still be time intensive. This introduces the approach of feature selection to limit the number of input features to the analysis; but raises a new question of what feature selection technique to use [13]. Genome-wide Association Study (GWAS) is a common approach that considers the effect of the whole genome based on the phenotypic response variable but is subject to high false discovery rate [14]. There are currently established methods that adjust for the inflated values that are present in false positives by using the output p-values from the association analysis [15]. As these methods reduce the inflated value monotonically, the impact of a false positive that shows high significance will still be one of the most significant values when adjusted by a multiple testing method. Unless the researcher is prepared to disregard the results entirely, these false positives will still be outlined for significance.

Additionally, one of the main focuses of this work considers to role of False Discovery Rate (FDR) and the issue of replicability in genomic studies [16][17][18]. Replicability has plagued the field of genomics for decades and is often attributed to the metric values and study processes that are used resulting in the development of correction methods for metrics [14][19], and processes for genetic specific adjustments such as population stratification [20] and linkage disequilibrium [21]. Given these efforts, the issue of replicability has improve significantly in recent years [22] however the introduction of epistasis presents its' own challenges with established methods for univariate analysis unsuited for the processing of epistasis detection [23].

To summarise, in order to develop a novel methodology, the problem to be addressed was established using criteria to guide the decisions. To focus on interpretable methods, epistasis detection is a relatively untapped resource for sporadic breast cancer with limited publications. To conduct an epistasis approach, a multivariate method must be used which covers an approach of pairwise to limitless arity, with the latter exposing the most potential combinations. A limitless-arity approach analyses every combination that is available from the input data to expose high-order interactions and is commonly exposed to the time complexity $O(n!)$, this method introduces the issue of computational complexity and intensity.

Within this chapter, introduced are the aims & objectives, scope and contributions to knowledge. The last section provides an overview of the thesis chapters and a small description of the contents of each.

1.2 AIMS & OBJECTIVES

Epistasis is a progressive approach to genomic analysis that considers the evolving hypothesis of system/ networking components of genetic material. While epistasis is not a new concept, the approaches that are available fall to limitations such as computational complexity and false discovery. The aim of this study is to develop a novel methodology, Random Sampling Regularisation (**RaSaR**), for the detection of epistasis (See Chapter 5:), using interpretable methods and best practice that will cater for the pitfalls of computational complexity and false discovery. Additionally, the methodology will be evaluated for effectiveness, which in this case refers to the proposed methodology's false positive outcome, using a genetic dataset and to benchmark against standard methods (See Chapter 3: and 6.4) in genomic publications. The objectives of this thesis were identified as:

Objective	Description
RO1: *Best Practice Quality Control*	To adopt best practice when conducting quality control to remove bias and erroneous data, to ensure that quality of the data for the purpose of performing further analysis.
RO2: *False Positive Rate Reduction*	To overcome/ reduce the occurrence of false positive genetic signals.
RO3: *Feature Filtering*	To outline features for epistatic analysis, accommodating for the occurring weaker/masked signals that are hypothesised for epistasis while maintaining **objective RO2**.
RO4: *Hardware Limitations*	To explore and appoint a method, technique or software that can perform epistasis analysis without the requirement of High-Performance Computing (HPC).
RO5: *Outline Candidate Variants*	To identify and outline candidate variants and interactions using a genetic dataset.

Throughout the thesis, the key elements of genomics and bioinformatics in reference to Epistasis will be explored to develop the proposed methodology and evaluate its viability through measures of reliability and replicability (See section 5.6 and Chapter 7:).

1.3 Scope

The proposed methodology aims to increase robustness detection of significant SNPs and SNP interactions, by filtering candidate features using sampling and statistical methods. It should be considered as good practice in the first stage process of exploring SNP markers. The research is of relevance to complex diseases where interactions between SNPs may be important. The methodology will be focusing on several areas that concern the analysis of Single Nucleotide Polymorphisms (SNPs). These areas include quality control, association analysis, multiple testing for high-order interactions and statistical epistasis analysis. The research uses large sample sizes and variants sequenced in line with current research in the area of the focus phenotype.

1.4 Contributions to Knowledge

Condensing the pipeline of the methodology, the approach chosen as the area of focus in this thesis is Epistasis, of which two main methods are conducted to accomplish the tasks; pairwise and limitless arity. Of these two, the method that leads to the least information loss is limitless arity. Using prominent research from this approach, the two main techniques flourish in Multidimensional Reduction (MDR) and Itemset Mining. These techniques are commonly coupled with other approaches that aim to reduce the computational complexity of limitless arity by reducing the number of variants for analysis. This introduces the problem of how to reduce the number of variants without increasing the false positive rate, whether that be the commonly occurring or artificially incited rate. The purpose; to analyse the behaviour of variants in each subset to measure the deviating behaviour and to regularise the measure by using all outputs to produce one overall mean. As epistasis introduces the concept of 'masking' (variants present reduced signals due to the presence or regulation of another variant), there is expectation that variants will not show a strong association to the phenotype and as such relying on the mean of the value is not the main concern of this method, alternatively the standard deviation, σ, is also produced from the value which will be used conservatively while the mean threshold will be set leniently.

The contributions of this research are layered and each address an identified problem. The processes outlined in this thesis aimed to produce a methodology for epistasis detection, using interpretable methods and best practice that also catered for the common pitfalls, computational complexity and false discovery. The novelty of this methodology exists in 3 areas:

False Positive Reduction Analysis

A novel methodology for false positive reduction association analysis by cohort sampling ($\times k$) followed by cross-validation (J-fold) resulting in a dataset of 9 p-values for each individual SNP, to which feature selection can then be applied. This novelty refers to the objectives outlines in **RO2** and **RO3**. To refer to the results obtained from this research. While a larger feature set size was outlined for analysis with the RaSaR method, interactions output from LAMPlink analysis resulted in only 4 SNP interaction combinations, as outlined in Table 1.1. All combinations that are greyed out in Table 1.1 did not yield a statistical association in this study.

TABLE 1.1: RESULT OF COMBINATIONS OUTLINED FROM EACH METHOD

Combination	Optimum PPP Conf.	C1	C2	C3	C4
rs4602520-rs6910087-rs7246472	0.833	O	O		
9q21.13-9q21.13	0.25	O			
rs4602520-rs4144827	0.5	O			
1q41-rs3924215	0.75		O		O
rs6911024-rs12170250	0.75		O		
rs6852865-rs4602520	1		O		
rs6852865-rs4602520-rs6910087-rs7246472	0.375		O		
rs6852865-rs6910087-rs7246472	0.667		O		
rs3924215-rs6011609	0.5			O	
1p12-rs6011609	0.5			O	
rs4602520-rs6911024-rs7246472	0.5				O
rs6911024-rs7246472	0.5				O
rs4602520-rs7246472	0.5	O	O		O
rs4602520-rs6911024	0.5				O

Incorporating the confidence of the PPP, only 3 viable options would be considered in a normal process of the methodology as the PPP confidence of 9q21.13-9q21.13 was 0.25, which is <0.5 and would not be considered for further analysis. Therefore, all viable combinations outlined by RaSaR indicated a true association based on the sample population. Reduced false positive concludes an improved detection of epistasis, demonstrated by the results. Of the combinations outlined, the proposed method was able to detect all but one combination (the singular variant 1p12 was identified by every method as the top result).

REDUCTION IN HARDWARE REQUIREMENTS

Due to the novel methods occurring in association analysis and feature filtering, the methodology can be performed without high performance computing equipment. The use of the feature filtering method is a main component in the methodology for reducing the hardware requirements. By outlining candidate variants using statistical filtering, the dataset can be downsized to adhere to the computational limitations of the researcher. Additionally, the use of open-source software LAMPlink allows for computationally efficient analysis of a reduced number of SNPs. This novelty refers to the objective **RO4**. During the process computational expense was limited, analysis was performed without the requirement of extensive hardware with limitless-arity analysis in LAMPlink for all methods taking an average of 20.5s.

IDENTIFICATION OF PREDICTIVE FEATURES

The identification of predictive features comprising individual SNPs and interactions between SNPs using decision tree methods followed by rigorous statistical significance testing for classification of cases vs. controls. This novelty refers to the objective **RO5**. Table 6.17 provides the genomic characteristics of the outlined variants from the RaSaR method, with exception for SNP rs6852865 which was not identified by the RaSaR method. Already outlined in section 7.1 are the limitations of the method which state the compromise of this method in its susceptibility to False Negative results. Given this, the method was still able to detect the remainder of the interactions, outlining 4 novel candidate variants and 1 variant that has already had various publication surrounding it’s link to breast cancer (See section 6.7).

1.5 STRUCTURE OF THE THESIS

Continuing from the previous chapter, the remainder of the thesis illustrates the considerations of this project within the scope of the problem statement. **Chapter 2** provides an overview of the core information that builds the foundation of this thesis. Genetics is a complex area that requires explanation if unfamiliar, given the speciality terms and components that are used sparingly across this thesis. The study of genetics is based on one of the most impressive and complex systems that is currently known and therefore this chapter discusses a succinct section of this field that provides information solely on the context scope. To conclude the section, the focus disease is discussed, providing statistical information regarding incidence, mortality and associated factors contributing to the development of breast cancer. The chapter concludes with a discussion of the concepts that introduce the subject of this research and analyses the impact on the methodology going forward, while providing information about decisions that determine the direction.

Chapter 3 moves the focus towards bioinformatics, discussing the techniques and methods that are used in the analysis of genetics data. To conclude the section, a summary of the information is provided which outlines the main points and outcomes of the section. Concluding the chapter, a discussion combines the information and outlines the outcomes of the chapter.

Already discussed is the complexity and various structure forms of genetic data, **Chapter 4** provides a representative section that cannot be to guide understanding of the processes of the methodology. This section includes information of the various representations of data and further provides the binary transformations that are used throughout the methodology. Also explained is the data as represented by PLINK to introduce the outputs and inputs of information before analysis and transformation are conducted. Finally, this section introduces the breast cancer dataset that will be used to evaluate the methodology.

Within **Chapter 5**, the proposed novel methodology is described, stage by stage, introducing the methods and techniques that are used along with the purpose of use. Stage 1 introduces the important stage of quality control, removing erroneous data and provides threshold choices for each process. Stage 2 introduces one of the novel elements of the methodology compared with standard approaches. The purpose and inspiration of this choice is discussed and explained using a definition of how the process is conducted and how it effects the process thread. Stage 3 introduces the stage of association analysis, utilising the deviated step in stage 2 to conclude the novel filtering method. Stage 4 provides information about the process of feature selection, using the output of the association analysis to explore the data and apply thresholds to extract informative features. Stage 5 uses the extracted features from the output of feature selection to perform epistasis analysis using pre-defined limitless-arity approach. Stage 6 is the most informative and uses the output of the multivariate analysis to explore the relationships and associations. This section describes the approaches, techniques and methods used to gain insight into the associated interaction within the sample population. This section is concluded by a summary that outlines the purpose of the methodology and the intended outcomes of its use.

Chapter 6 outlines the results from the execution of the proposed novel methodology alongside comparable standard methods. Processing the stages as defined in Chapter 5, plots and information indicates the input, output and conclusion of the section. Stage 4 outlines threshold methods applicable to standard case-control approaches to benchmark and evaluate the effectiveness of the proposed methodology. Each method is described, and the section concludes by providing the features selected from all methods. The chapter progresses using the input outlined from stage 4 and concludes the results in stage 6, with a discussion about the interpretation provided in the final section of the chapter.

The concluding **Chapter 7** discusses the effectiveness of the method, the outcome of information and also outlined the limitations of this methodology. Future work is proposed which discusses the evaluation of the methodology and the adaptation for optimisation dependent on the given data structure and study design.

Chapter 2: Literature Review

The following chapter will discuss the areas of interest that motivate this work. As a cross-disciplinary approach to the research, the three fields of biology, computer science and statistics will be combined to address the defined problem as outlined in Introduction. The intent of the following chapter is to introduce the origins of this research by discussing the fundamental knowledge base as a primer to the remaining chapters.

Genomic studies pose a particularly difficult challenge in the navigation of genetic mutations that can either directly or indirectly effect a phenotypic expression in its host. While the most current studies use association analysis in a univariate capacity, it is becoming clearer, as the field develops, that singular mutations are unlikely to produce the severe expression that is common in many genetic disorders and disease [24].

With the introduction of GWAS, too came the added complexities that are now commonly associated with diseases. GWAS highlighted the cross chromosome and gene associations that were present in relation to disease and disorders [25]. Due to the discoveries uncovered by GWAS which suggests that genetic causation is far ‘messier’ than previously considered, therefore using SNP-SNP Interaction to uncover SNPs that work cooperatively to produce the phenotypic expression of these disorders and diseases is a considered approach of this work. SNP-SNP interaction considers the possibility that SNP mutations may occur in the population but are not effective unless one or a group of SNPs present a particular status which aggravates or “malfunctions”, resulting in the disease and disorders that are common in the population but are yet to be associated with a genetic component [26]. While SNP-SNP Interaction would result in an exhaustive search mechanism that may indicate potential associations and interactive groups; the computational complexity of full genome scan would render most computers useless.

While using a small number of SNP isn’t exactly computationally expensive, if we consider 500K SNPs using a bivariate analysis, a result of 125 billion possible SNP pairs are possible [27]; further to this, these calculations do not take into account the possible genotypic state. Considering that any regular genome wide analysis will include approx. 300,000 – 500,000 SNPs, we can now start to get a better understanding of why SNP-SNP Interaction is not an easy problem to approach [28].

One way of reducing the computational complexities of a SNP-SNP Interaction analysis is to reduce the number of SNPs. The decision as to which SNPs to include then becomes a problem as evidenced from previous studies [29][30], significantly associated SNPs may not be true positive associations. It is also noted that SNP mutations that cause phenotypic expression may not appear significant when a univariate analysis is used [31]. Therefore, we require a method that can pick out SNPs that are likely to show significance in an interaction analysis regardless of their significance in an initial univariate analysis.

Complex disease and disorders continue to be evasive in the underlying genetic component, this is likely due to effects other than genetic inheritance and/or mutation [28]. Given the prevalence across the world, the occurrence of these diseases displays different prevalence among countries, environments, diets and classes [32]. While it is difficult to determine an environmental-gene factor, given the number of environmental factors that could contribute to the phenotypic expression of a disease, we can consider the potential of SNP-SNP interactions that could indicate a potential environmental factor due to the pathophysiology of the gene it belongs to. Firstly, introduced in this chapter is the root subject that will guide and effect the decisions and approaches of this work. To introduce genetics to this thesis, discussions will focus on the central dogma process that guides the biological expression of the focus data type.

2.1 Genetics

Genetics as a field considers a whole range of biological structures, functions and physiology that are present in every living organism. For this research, we are concentrating on human genetics particularly concerning the variability that is present between humans. The following section provides information about the basics of genetics. We discuss the basic components that make up the human genome, the function, activity and the impact on the body regarding the reported research in this document.

2.1.1 Fundamentals of Genetics

Genetics are determinant of not only our hair and eye colour but our bodily functions, personality and desires [33]. Genetic make-up is referred to as genotypic information; genotypic information produces the traits in humans and the physical expression of that trait is referred to as a phenotype [33].

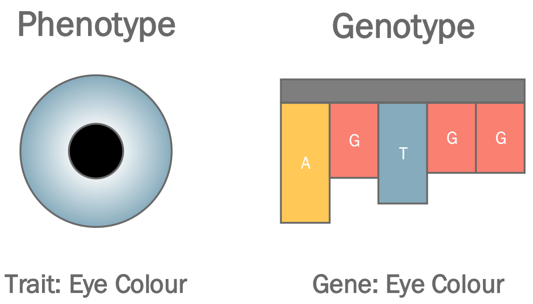


Figure 2-1: Demonstration of Phenotype and Genotype

Terms *genome*, *gene* and *base-pair* refer to size and area of the genetic make-up [34];

Genome	the total set of instructions for protein production of an organisms or the total amount of genetic material in a cell.
Gene	an area of the genome with instructions to produce a specific protein; (Note: not all of the genome is divided into genes)
Base-Pair	one pair of nucleotides.

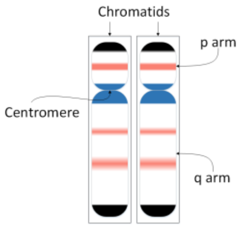


Figure 2-2: Chromosome

Each cell in the human body contains 46 chromosomes or 23 pairs of chromosomes. Chromosome pairs are passed to us in as haploid cells from both of our parents, each chromosome is then paired with its corresponding pair i.e. Chromosome 1.

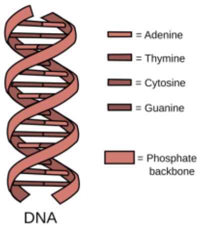


Figure 2-3: DNA

The structure of our genetic information, consisting of a sugar-phosphate backbone and nucleotide base-pairs; Adenine, Guanine, Thymine and Cytosine. DNA is coiled around histones which have been linked to environmental interaction [198].

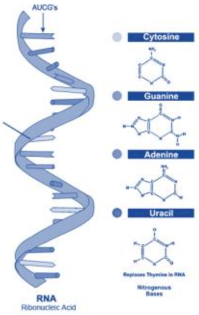


Figure 2-4: RNA

Produced during the transcription phase of the central dogma, RNA is a complementary copy of a specific DNA strand. It is used in the translation process to produce proteins; codons of 3 nucleotide bases are used to indicate the sequence of amino acids which form a polypeptide chain [199].

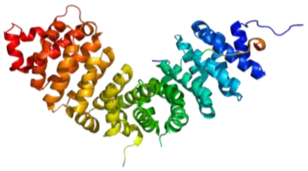


Figure 2-5: Protein

Proteins are created from the polypeptide chains that are formed during the translation process. Proteins are required for the function, structure and regulation of the bodies tissue and organs [200].

Genes provide the areas of the genome that specifically produce proteins. Areas of the genome that are used in the production of proteins are referred to as exons, or coding regions, while areas that are not used for protein production are referred to as introns, or non-coding regions. Intron regions are found outside of genes, but can also be found within [35].

The central dogma is the process that reads the DNA sequence of a gene and translates it into a protein which will form structure or function within the human body; this includes the phases transcription and translation [34]. Transcription refers to the process of unzipping the double-helix structure of DNA to create a copy of the sequence via complementary nucleotides. Translation uses the new single strand of DNA, referred to as RNA, to build a protein; using a sequence of 3 nucleotides called codons, a protein is built from the amino acids that correspond to the recipe of the 3 codons, once complete the protein is folded and transformed to its functional state.

The process leading up to the point of protein expression is complex and many issues may occur such as protein misfolding [36]; however the focus of this thesis concentrates on the building blocks of genetic information, nucleotides.

2.1.2 Genetic Mutation

Genetic mutations are changes in the base structure of DNA; this could be a single nucleotide base or could include a set of genes. These changes can be either hereditary or somatic; hereditary changes are passed down to offspring from their parents. However, in some cases, genetic mutations can occur in the egg or sperm cell which are not present in the parent's DNA, these types of mutations could explain genetic disorders that affect people whose parents do not possess the mutation in their genes [37].

Somatic mutations either occur during cell division when copying the DNA or by environmental factors such as exposure to UV rays. Mutations occurring in somatic cells cannot be passed down to offspring. Genetic Mutation occurs during meiosis (Sexual haploid division) and mitosis (Cell division used for growth, repair and asexual reproduction).

There are three types of mutation that occur in small regions; insertion, deletion and variation [38]. Insertion and deletion can have serious consequences is they are located in coding regions of the genome. This type of mutation can shift the protein reading sequence, this can lead to an incorrect subsequent protein. Mutations normally occur in non-coding regions, but even mutations that occur in coding regions will have little to no effect [39]. Mutations that do affect the body can influence the protein produced, the amount produced and/or when and where it is produced; this can cause serious health problem in the affected individual.

Normally mutations that cause disease in individuals are uncommon in the general population, however there are mutations that are common throughout the population. Common mutations normally control the differences in the population e.g. eye colour, hair colour, height etc. There are however common mutations that are associated with disease and disorders; this type of variation is one that this research is interested in investigating. Base-pair nucleotides in which a variation exists in more than 1% of the population are referred to as Single Nucleotides Polymorphisms (SNPs); the focus data type of this thesis. The next section will discuss SNPs and their biological impact.

2.1.3 SINGLE NUCLEOTIDE POLYMORPHISMS (SNPS)

SNPs are single base-pair nucleotides that span DNA to provide the instructions for protein production. Only nucleotide base-pairs that show variation in >1% of the population are considered SNPs. SNPs are denoted by the letters A, C, G and T which correspond to their representative amino acid, Adenine, Cytosine, Guanine and Thymine. Across the span of approximately 3 billion nucleotides, SNPs occur around every 300 nucleotides which is approximately a total of 10 million SNPs [40].

SNPs are made up of 2 alleles that can take one of 3 forms (Explained further in Chapter 4:). Dominant alleles present the phenotypic traits that are present in humans i.e. Brown eyes are a dominant gene. Recessive alleles are 'overpowered' by the presence of the dominant allele but may still present phenotypic traits due to the genotype present.

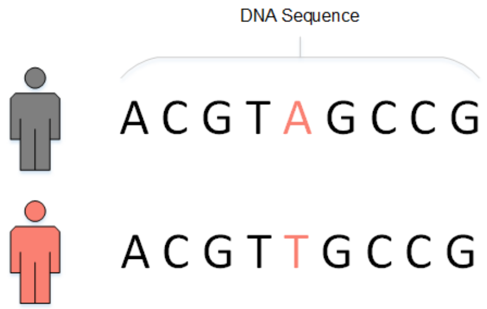


FIGURE 2-6: SINGLE NUCLEOTIDE POLYMORPHISM; A VARIATION IN A SINGLE LOCUS WHICH IS PRESENT IN MORE THAN 1% OF THE POPULATION.

There are 3 variations of genotype, if we represent the dominant allele as 'A' and the recessive allele as 'a'. Homozygous dominant refers to a genotype which presents the same allele on either chromatin, with both alleles being the dominant variant, and is represented by 'AA'. Homozygous Recessive refers to a genotype which presents the same allele on either chromatin, with both alleles being the recessive variants, and is represented by 'aa'. The final genotype is Heterozygous which refers to a genotype that has both a dominant and recessive allele and can be represented by, 'Aa' or 'aA'. If the genotype is known, variants of both mother and father can be used to generate a probability on the traits that any offspring would have, such as eye colour as demonstrated in Figure 2-7.

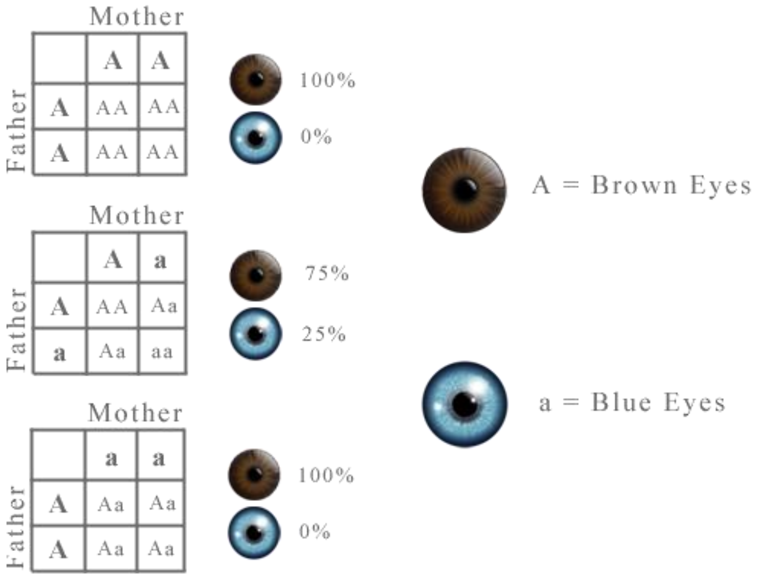


FIGURE 2-7: EXAMPLE OF TRAIT INHERITANCE

As the dominant allele, A, is brown eyes, whenever the dominant allele is present in the genotype combination, the phenotype expression of that trait will be present. So even in the case of a heterozygous genotype, the dominant allele is present and therefore the phenotypic expression will be based on the instructions of this allele.

There are two types of SNPs: Linked and Causative. Linked SNPs are located outside of genes but can still affect the function of the body such as drug response and disease risk. Causative SNPs are affective mutations that are located within genes; causative SNPs further divide into coding and non-coding SNPs which correspond to their location within the gene. Mutations occurring in the non-coding region of the gene do not affect the protein production of the gene but can still effect the time, locations and level of gene expression [41]. Mutation occurring in the coding region of the gene can affect the amino acid sequence, directly effecting the protein produced. Having discussed the fundamentals of genetic in the context of the research in the current section, the next section discusses standard and progressive study approaches to genetic analysis.

2.2 GENOMIC STUDY APPROACHES

The following section introduces the approaches used when investigating genetic data. These approaches consider the size of the dataset, the selection of features and the cohort design. The aim of these approaches is to aid in discovering potential candidate genes that can be utilised in a clinical setting for the purposes of improved pharmacological solutions, diagnostic criteria, susceptibility and finally the potential to identify predisposition to a disorder.

The analysis of genomic data is providing individuals with information and knowledge to take control of their health [6]; while still in its early stages, genomic analysis and its resulting outcomes can aid the healthcare sector, approaching the much debated subject of personalised medicine [6]. Personalised medicine caters for the needs of patients by considering their biological and epidemiological make-up. This will in future replace the current "one-size-fits-all" approach that is common in prescription medication; an individual's biological make-up could provide information for the most appropriate treatment response, i.e. indicating the amount, variety and response to particular drugs [7]. This is important in complex diseases as treatment is often based on a necessary 'trail-and-error' period which may or may not provide relief from the symptoms [42]. Further to this, even treatment options that aid in reducing or eliminating the problematic symptoms of a disease or disorder, can also cause a variety of side effects that can still effect patients Quality of Life [43][44].

There are several approaches to genomic study which are commonly used to identify risk variants in common, complex diseases such Breast Cancer; GWAS (Genome-wide Association study), Candidate Gene and Familial studies. One of the most popular genetic feature inputted for study analysis are SNPs (Single Nucleotide Polymorphisms), these are variants in base pairs within the DNA sequence [45].

While a majority of these SNPs will have little to no impact on the biological systems, the consequential causal sequence can lead to imbalances in chemicals, misfolds in protein polypeptide chains and instability in mRNA transcripts [36]. The involvement of these SNPs in the genetic analysis for the purpose of finding risk variants is due to the abundance of variation throughout the genome; proving promising and successful in many determined diseases so far [46][47]. Although there are many other genetic and biological studies that are successfully undertaken, the following identified approaches utilise the SNP feature input for the analysis of correlation and susceptibility in subjects. As such, the following sections encompass the approaches to data analysis that consider the variability that exists in genomics due to the structure of genetic data and the subject specification.

2.2.1 GWAS (GENOME WIDE ASSOCIATION STUDY)

GWAS provides a way in which the whole genome (genotyped SNPs) can be scanned to identify SNPs that confer risk for the identified and analysed phenotype. Presenting a hypothesis-free approach that has introduced an option for researchers to visualise whole genome effects for diseases.

The most common approach in GWAS utilises a case-control set-up [48]; Cases refer to a cohort affected by the disease subject of the study and Control refers to a cohort who are unaffected by the disease. The proceedings of a GWAS aims to find the correlation results between the cohorts and the disease. In an ordinary case-control GWAS, the odds ratio is the first considered statistics in which an OR > 1 suggests the association of an allele is a risk for disease, the greater the difference from 1, the more indicative of an association and an OR < 1 suggests a protective association against a disease [48]. Performing a chi-squared test from the results will provide significance of the alleles association; that is, how likely it is that the result is truly associated with the disease.

While GWAS presents a unique approach for analysis of genetic material, its requirements introduce both advantages and disadvantages. GWAS have also previously been acknowledged for their expense; however, this criticism is becoming obsolete as advances in technology are reducing the costly price [49]. This approach also outlines some disadvantages that effect the reliability of the study such including high false discovery rate and the overlooking of rare alleles which could potentially be important to the discovery of biomarkers [50]. As such, an important feature of GWAS are the requirements for a large sample size for reliability of result outcomes [51]. Unfortunately, this accommodation does not rectify the issues that are present with false discovery in GWAS and given the parameters that define the size of these studies, transfer learning is commonly adopted from methods that aim to reduce, rectify and eliminate the effects of bias and false discovery in 'Big Data'. A common approach from big data techniques is to use multiple testing adjustments, as discussed later on in Methodology. Successes in GWAS have previously outlined viable SNPs in complex diseases such as Crohn's Disease [52], Rheumatoid Arthritis [53] and Celiac Disease [54]. It has also previously been proposed that GWAS studies should be a first step in the genetic identification process [55].

Within the next section, the focus moves to an approach that contrasts with the whole-genome approach of GWAS to introduce an approach that focuses its efforts in areas of significance based on prior knowledge.

2.2.2 Candidate Gene Approach

In contrast to GWAS, the candidate gene approach focuses on a small selection of genes, SNPs and/or alleles that are chosen based on relevance in the role of the disease/phenotype in question [56]. While the most common approach is to use relevant markers to test association, the relevance of the gene can be based on prior knowledge of the biological, functional and physiological mechanisms that have an identified association to the disease/phenotype [57]. However, markers that have identified in previous studies for the focus phenotype can also be used[58].

The candidate gene approach presents a unique advantage in that it is quick and easy to determine the association between the disease and selected genes to determine the effect of genetic variants [56]. Issues that arise with this approach lie in the conservative choice of genetic information, by reducing the dataset and therefore the number of genetic association possibilities, variants that could arise in a genome scan are overlooked. This couples with the limitations of the research present to determine a candidate gene; with the speed at which genomics is evolving; the information to supply to this field can only indicate a limited amount of candidate genes [59].

In contrast to GWAS, the candidate gene approach focuses on a small selection of genes, SNPs and/or alleles that are selected based on known disease pathology. Within the next section, another approach is introduced that focuses its effort on the subject specification.

2.2.3 Familial Study Approach

Familial studies approach contains numerous types of study design built upon the unique advantages that are present in the focus research of related subjects. The study types can generally be summarised into 3 categories: Twin, Linkage Analysis and Other. The following sections provide a summarised outline of the study approaches in family studies and the unique value that they provide.

2.2.3.1. Twin Studies

Twin studies have made some of the most ground-breaking discoveries in risk variants for disease [60]; this is due to the extreme similarities in their genomes particularly when concentrating on identical twins. There is a dissection in the study types which are conducted as follows: Monozygous twins (Identical), Dizygous twins (Non-identical/ Fraternal), twins who are nurtured apart and twins who are adopted and nurtured by unrelated foster parents [27].

Monozygous twins provide an almost identical genome from the point of birth that implies that in the majority of cases any phenotypic, epidemiological or genotypic variation is caused by environmental intrusions and influences [27]. Dizygous twins are used within studies to determine the concordance in both Dizygous and Monozygous twins; if concordance is higher in Monozygous twins then genetic susceptibility is a risk [27]. The use of both, reared apart and reared apart with unrelated foster parents, is used to determine the environmental effects on the subject when faced with different nurturing techniques and environment e.g. Urban vs. Rural living [43]. This can help us to identify, or at least give us some indication as to the environmental factors that can affect our genetic architecture; these implications can be studied in non-related individuals but the advantage of twins studies lies in the genome similarities given the genetic implications are unlikely to be due to variants in the genome. While the advantages of twin studies are clear, a few issues hinder the progress and preservation of this approach. The attainment of twin data is relatively restricted given the limited cohort, focus on genotype can be skewed if subjects are exposed to different environments, gene expression on monozygous twins can differ and the twin unaffected by the disease is less willing to participate than the twin affected [44].

2.2.3.2. Linkage Analysis

Linkage Analysis focuses on the probability of transmission of alleles at closely located positions on the genome as an intact force, commonly referred to as a haplotype [44]. Linkage Analysis is an approach not confined to the family studies branch but can be adapted to many different approaches e.g. GWAS, Candidate. However, as a common approach in family studies. it uses related individual's data to map the genetic sites of a disease or disorder trait. In doing so, it provides a unique opportunity to explore the potential of susceptibility association to multiple loci [44]. This uses the recombination factor to produce values of variance that either indicate tightly linked (No Recombination) or unlinked; unlinked recombination factor will indicate a shift in loci over generations of families. Linkage Analysis is primary approach that identifies an initial cohort of variants and loci of potential genes for further analysis; this therefore provides our basis becoming a less commonly used approach once the initial studies are produced.

2.2.3.3. Other

Other family studies can include different approaches that can include the identification of risk in 1st- 2nd- or 3rd degree relative, the risk in female relatives, the risk of inheritance between offspring and parentage [27]. The main basis, and therefore the origination of family studies stems from the research conducted by the famous Mendel [45], in which he used 'sweet peas' to track the trait inheritance based on the manifested phenotype of the parentage. The most common approach in family studies is that of the case-control study type [45]; this uses affected subjects compared to unaffected subjects, therefore aiming to provide insight into the genetic differences that are present between the two cohorts essentially the identification of causal SNPs for the focus disease or disorder. Again, the issue with family studies lies in both the attainment of data and the limits in application to society, as the results cannot indicate and are not generalised to the public it only provides heritability and susceptibility risk for the cohort included in the study e.g. risk variant for offspring of 2 parents affected by Schizophrenia [45]. The next section introduces the final study approach of this chapter, a progressive approach that responds to continuing research.

2.2.4 EPISTASIS

Epistasis refers to the interaction between genes, but more commonly encompasses the interactions between genetic components. Epistasis covers three major categories; functional, compositional and statistical epistasis [26]. Functional epistasis addresses the interactions that occur between proteins, this is less adopted use of the term epistasis but covers the functional consequences within genetic pathways. Compositional epistasis describes the phenomenon of the blocking of one allelic effect by an allele at another locus [26]. This category considers the composition of the genotype, and as such, the discovery of such interactions relies on a substitution process to realise the effects of the ‘masking’ or ‘aggravating’ loci. Statistical epistasis is the analysis of the effect of combinations of alleles at different loci over all present genotypes within a population. This approach presents the most flexible but consuming option that measure the average deviation given multiple states, combinations and locations.

Epistasis benefits from an exhaustive technique that not only considers genetic components as singular entities but in combinatorial components. Genetic pathways already indicate a level of interaction as evidenced by the interactions that occur for regulation of gene expression, signal transduction and biochemical pathways [61].

While epistasis extends a potentially untapped source of information in genetic pathways and disease penetrance, its current successes are limited. This may be due to the limitation of the approach; computational complexity refers to the amount of resource required to perform the algorithm. Commonly in GWAS, the number of SNPs that are tested extends past 300,000, as demonstrated in Figure 2-8, the computational complexity, 3^n, of epistasis increases exponentially with every additional feature that is included. At 10 features, more than 50000 interactions are possible.

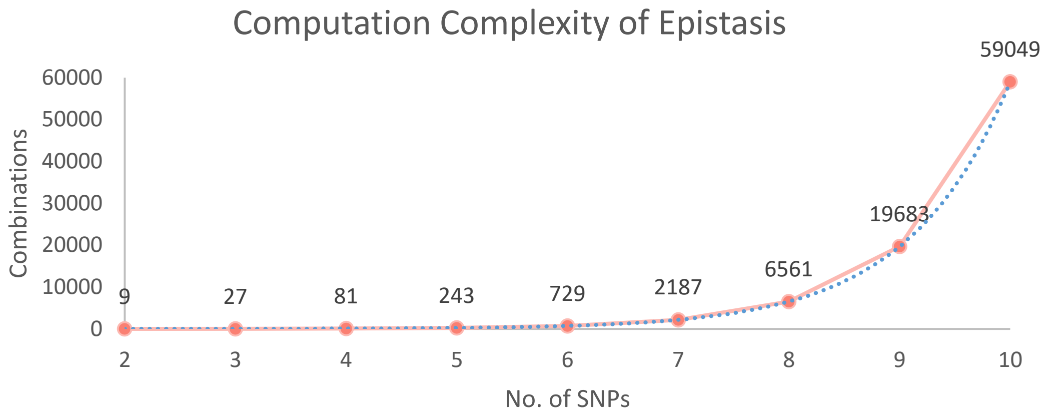


FIGURE 2-8: DEMONSTRATION OF THE INCREASE OF COMPUTATION COMPLEXITY WITH EACH ADDITIONAL FEATURE

Epistasis is an approach fast emerging in genome studies [26]; its potential in inviting possible successes for diseases and disorders that show little to no genetic signal particularly concerning complex disease whose phenotypic presence can vary from individual to individual. Complex diseases present a particularly difficult challenge in their expression; with varying phenotypic expression that show strong association to environmental factors, the genetic composition remains ambiguous for many of these diseases. Epistasis may be able to shed some light on the genetic components causing the phenotypic expression of these diseases.

2.2.1 BIOINFORMATIC PROJECTS

We use SNP information for a variety of purposes including pattern recognition for classification, prediction and susceptibility. The following section introduces pinnacle projects in bioinformatics that have advanced the field and introduces the focus area of bioinformatics that influence the statistical components of the work. Bioinformatics can be categorised into two fields; representation and inference. While representation encompasses bioinformatics methods and techniques that aim to define the way in which genomic information is interpreted e.g. DNA structure, categorisation of coding regions, inference measures the associations within the structure e.g. relationships between phenotype and genotype, pathways that interconnect gene by functional expression or regulation. The field of genomics and bioinformatics is constantly expanding and improving as evidenced by the international projects that have rocketed the field into a new era:

The Human Genome Project (HGP) was an international effort that pooled resources and skilled minds to fully sequence and map all the genes in human genome. Completed between 1990-2003, its distinguished contribution has encouraged and strengthened the field of genomics by providing a reference map for researchers to improve their research [62]. The HGP outlined approximately 20,000-25,000 genes in the human genome.

The HapMap Project defined what is now referred to as ‘tag-SNPs’, these SNPs allow researchers to reduce their SNP feature set from the original 10 million to a set of SNPs that are representative of a haplotype. A haplotype is a set of SNPs that are in high LD with each other, therefore can be represented as a group of SNPs with just 1 SNP that will provide the same results [63]. This project successfully aided researchers by reducing analysis from exhaustive to the necessary components.

The 1000 Genome Project focused on genetic variants with frequency >1% in populations studies to create the world’s largest public catalogue of human variation and genotype data [64]. This provided references for the variations being studied by researchers.

Each of these projects has had a significant effect on the genomic and bioinformatics community, giving rise to the development of more sophisticated software and techniques that ease the analysis process for researchers making the field of bioinformatics more accessible for geneticists.

Within this research, the focus of our efforts is dedicated to the inference category of bioinformatics, considering the utilisation, adaptation and development of inference models and methodologies that can aid in producing robust and efficient results. The methods and techniques of inference bioinformatics, both standard and state-of-the-art, are later discussed in the next chapter. Given the bioinformatics approach, further considerations are required to incorporate the genetic component of the project in considering biological pathways, expression and participant specification. The next section discusses the focus phenotype used within this research.

2.2.2 STUDY APPROACHES SUMMARY

Having explored the various study approaches applicable in genomic studies it is clear that each approach has its own advantages and limitations. Firstly, considered is the feature space approach that heavily contrasts between GWAS and candidate studies; the advantages of candidate studies draw their benefit from specialised knowledge of the phenotype/ trait that in itself is also a disadvantage. Prior knowledge of the phenotype/ trait results in limitations of the analysis and while genetics is still advancing there are many concepts and functionalities that are ambiguous and unknown. GWAS presents a solution to this limitation but at the cost of a high FDR, overlooking rare alleles and the requirements of a large sample [24]. GWAS is better suited to complex and common diseases that are hypothesised to lend their causality to 'common disease, common variant', referring to the phenomena that common diseases will be caused by a large number of common alleles [65].

To consider the subject specification, again there is a vast difference between the familial studies and GWAS. Familial studies can be used in conjunction with candidate studies depending the phenotype and study design and GWAS can be used in conjunction with Familial studies but is commonly performed using case-control unrelated subject cohorts. Familial studies introduce an approach that responds to fundamental genetic information that suggests that mutations are inherited from parentage. This study type has proved very significant in the genomics community and generally guides the research of the disease depending on the results. The limitations of this study approach lie within the fundamental requirements, the subjects. Common approaches in familial studies use either twins (monozygous or dizygous), trios (mother, father, offspring) or full-sisters, of which the genetic similarity is highest; this restricts the participant recruitment stage as if one of the duo/trio are not willing to participate, all participants must be removed. This also provides a very specific overview of a limited participation and does not apply to the general population.

Epistasis is an emerging field that has seen a rise in interest as prevalent evidence points to systems or networks of functional variants interacting to produce a phenotypic response. However, given the lack of success for clinical incorporation, there is some scepticism around the area [66]. Having discussed the approaches of genomics the thesis now introduces the technological impacts that have influenced and advanced the field.

2.3 Breast Cancer

Cancer is a global concern, with prominent mortality rates across the board demonstrated by its current position as second leading cause of death in the United States, 2016 [67]. Cancer covers a range of related diseases which initiate in different areas of the body, most commonly originating in the breast, prostate, colon and rectum [68]. The pathology of cancer is caused by 'defects' in the function of Apoptosis, also commonly referred to as Programmed Cell Death (PCD) [69]. Apoptosis refers to a process in which cells are instructed to deconstruct themselves; a necessary function of the body which particularly concerns the gastrointestinal tract, immune system and skin [70]. Excessive PCD can lead to diseases and disorders such as neurodegeneration and ischemia, while a lack of PCD can lead to diseases concerning the autoimmune system, famously, cancer. As the most frequently occurring cancer in women, breast cancer is a major health concern in our current society. On a global scale, breast cancer represents a broad spectrum which appears to be more prevalent in developed countries [71].

In 2016, 11,563 deaths were reported due to Breast Cancer, with increasing incidence rates that resulted in approx. 55,000 new cases in 2015, for England alone [10]. The current survival rate for Breast Cancer in England is 78%, however this is highly related to screening practices that are in place for quick diagnosis, ensuring treatment is started as soon as possible [10]. The symptoms of breast cancer vary, and quite often are due to common occurrences in the body that are unrelated to the development of cancerous cells. Current campaigns urge women to regularly check the size, shape and feel of breasts to be aware of changes that are associated with breast cancer. Lumps, breast pain, changes in skin colour and texture, abnormal discharge and inverted or sunken nipples encompass the most common symptoms associated with Breast Cancer [72].

Breast cancer is most curable in its early stages which emphasises the importance of the screening processes that are in place. Diagnosis of breast cancer is most commonly conducted using imaging techniques including mammograms and ultrasound [73]. Diagnosis of breast cancer normally adheres to a 'two-week wait' protocol that insists that suspected cancer patients are first seen by a specialist within 2-weeks [74]. With this protocol in place, ~90% of cases with known stage are diagnosed with early stage breast cancer (Stage 1 & 2, discussed later) [10].

Breast Cancer is divided in to 4 stages that are based upon the TNM staging system. The TNM system uses information about the tumour size, node spread and metastasis status to assign a stage to a case. Tumour size refers to the size of the tumour present in the patient. Node spread considers the presences of cancer cells in lymph nodes in the surrounding area of the cancer site. Metastasis refers to status at which the cancer has developed in other areas of the body. Table 2.1 outlines the staging system for breast cancer [10]. Stages assigned to breast cancer cases are based on the varying status of TNM. Table 2.2 outlines the breakdown of the breast cancer stages with reference to the TNM staging system [10].

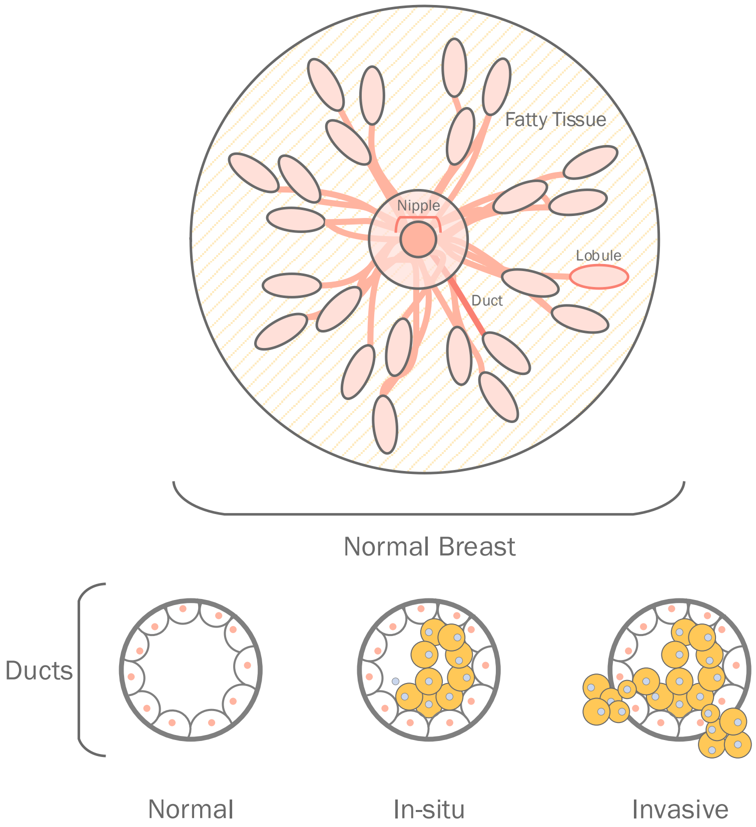


FIGURE 2-9: OCCURRENCE OF 'IN SITU' AND 'INVASIVE' BREAST CANCER WITHIN THE LOBULES.

Taking into consideration the stages outlines for breast cancer, it should also be noted that there are 2 types of breast cancer which refer to the status and spread of cancer cells in the breast. ‘Ductal carcinoma in situ’ refers to the state of breast cancer when cancer cells have not yet spread beyond the lining of the duct or lobules [32]. In contrast, invasive breast cancer is the state in which the cancer cells have spread to the surrounding area; this type of cancer encompasses the majority of breast cancer stages and primarily concerns breast cancer states that have developed a tumour status [32].

TABLE 2.1: TNM STAGING CLASSIFICATION CRITERION

TNM	Description
Tumour	***How big is the primary tumour and where is it located?***
T0	No Evidence of Cancer in the breast
Tis	Carcinoma In-Situ. Cancer is confined to ducts and/or lobules.
T1	x < 20mm
T2	20mm > x < 50mm
T3	x > 50mmm
T4	Tumour has spread into the chest wall and/or skin, or is inflammatory breast cancer
Nodes	***Is there evidence of cancer in any lymph nodes? If so, how many and where?***
N0	No cancer found or areas of cancer < 0.2mm
N1	Cancer has spread to between 1 and 3 axillary lymph nodes and/or the internal mammary lymph nodes.
N2	Cancer has spread to between 4 and 9 axillary lymph nodes or the internal mammary lymph nodes but no axillary lymph nodes.
N3	Cancer has spread to 10 or more axillary lymph nodes. Or it has spready to lymph nodes located under the clavicle/ collarbone, it may have also spread to internal mammary lymph nodes.
Metastasis	***Has cancer spread to other parts of the body? If so, how much and where?***
M0	Cancer has not metastasized
M1	The is evidence of cancer in other areas/ organs of the body.

TABLE 2.2: DIAGNOSIS OF BREAST USING TNM STAGING SYSTEM

	T	N	M	Cancer Type
Stage 0	Tis	N0	M0	Non-invasive(in-situ)
Stage 1a	T1	N0	M0	Invasive
Stage 1b	T0/T1	N1	M0	Invasive
Stage 2a	T0	N1	M0	Invasive
	T1	N1	M0	Invasive
	T2	N0	M0	Invasive
Stage 2b	T2	N1	M0	Invasive
	T3	N0	M0	Invasive
Stage 3a	T0/1/2/3	N2	M0	Invasive
	T3	N1	M0	Invasive
Stage 3b	T4	N0/1/2	M0	Invasive
Stage 3c	T(All)	N3	M0	Invasive
Stage 4	T(All)	N(All)	M1	Invasive

CONTRIBUTING FACTORS TO BREAST CANCER

Disease prevention, progression and diagnosis rely on further contributing information that have been highlighted in research for their association and relationship with Breast Cancer. The majority of risk factors related to breast cancer are previously established in research from as early as 1970's [8]; with the focus of research changing direction to embrace the advances of molecular and genetic impact. The following sections outline associated factors of Breast Cancer.

2.3.1.1. AGE

Incidence rates in breast cancer are highly associated to the age of the patient. The probability of developing Breast Cancer within the next 10 years increases as women age with the median age at ~61 [75]. Figure 2-10 provides an overview of the increase in probability of developing cancer over a period of 10 years, based on age [75]. However, this factor likely a covariate to the remaining factors that relate to the reproductive, growth and hormone production.

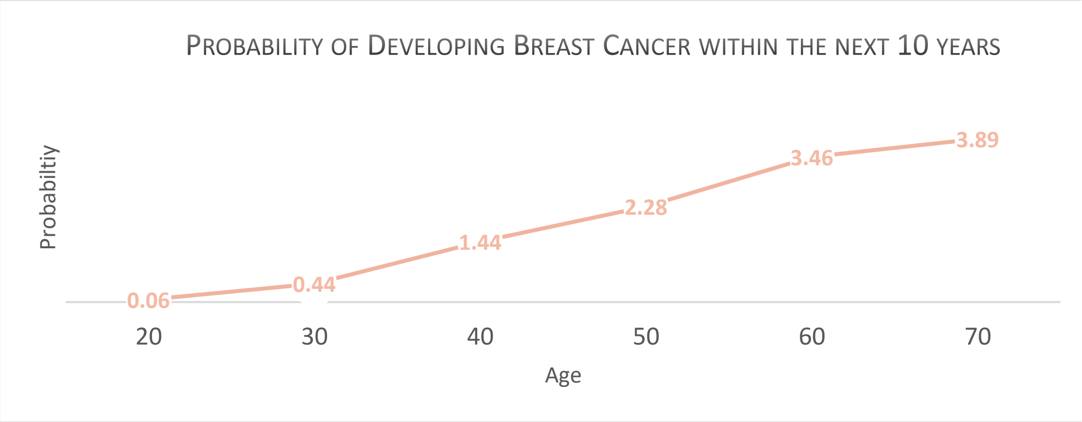


FIGURE 2-10: INCREASE OF PROBABILITY IN DEVELOPING BREAST CANCER OVER 10 YEARS BASED ON AGE

2.3.1.2. AGE AT MENARCHE & MENOPAUSE

Concerning menstruation, early start and late menopause are associated with an increased risk in Breast Cancer. Early age-at-menarche is associated with hormone receptor positive (HR+) cancer (explained later) [76]; high risk established in women whose menstruation starts before the age of 11 with a relative risk of 3 [67]. Natural menopause after the age of 55 presents a 2-fold risk than women whose natural menopause start before the age of 45 [77]. Further to this, women who undergo bilateral oophorectomy before the age of 35 have a decreased risk of developing breast cancer than their peers who experience natural menopause [77].

2.3.1.3. Parity Factors

In earlier years, focus of risk factors in breast cancer were associated with the relationship status of women, single or married [8]. This was later overshadowed by risk increase being associated with nulliparity (no pregnancies), or women whose first birth was at a late age [78]. Women who have their first child after the age of 30 are twice as likely to develop breast cancer as women who have their first child before the age of 20 [77]. A ratio risk (RR) of 3 is associated with women who have their first child past the of age 40, putting them in a high risk category [77].

2.3.1.4. Hormone-related Factors

Due to the sparse nature of cancer, and with that breast cancer itself, one occurrence of breast cancer may present differently to another due to molecular, histology or morphological tumour characteristics [32]. A prominent subtyping category for breast tumours is hormone expression pertaining to Oestrogen Receptors (ER), Progesterone Receptors (PR) and Human Epidermal Receptor 2 (HER2). Subtypes are guided by the presence of ER and HER2. ER positive (ER+) tumours are more common than ER negative (ER-) tumours, occurring in 30-70% of cases [32]. ER+ tumour present less aggressively than ER- tumours, with smaller tumours, low grade and lymph node negative. Major subtypes include Luminal A, Luminal B, Her2 and basal-like. Luminal A is the least aggressive form of cancerous tumour and presents the best prognosis while the remaining subtypes presents a worse prognosis that can include further complication such as cancerous cells in lymph nodes. Table 2.3 outlines the characteristics of each subtype [79].

Table 2.3: Characteristics of tumour Subtypes

Subtype	ER+/-	HER2+/-	Characteristics
Luminal A	+	-	Low levels of protein Ki-67[a]
			Tumour grade 1 or 2
			Slow growing tumour
			Best Prognosis
Luminal B	+	+/-	High levels of protein Ki-67
			Slightly faster growing than Luminal A
			Slightly worse prognosis than Luminal A
Basal-like	-	-	Common in women with BRCA1 mutation
			Common among young and African American women
HER2 enriched	-	+	More treatable with anti-HER2 drugs
			Lymph Node +
			Poorer tumour grade

[a] A protein that helps to control how fast cancer cells grow

2.3.1.5. Family History

A commonly utilised factor for diagnosis and risk in patients is family history. This factor presents a link between breast cancer and genetic predisposition resulting from inherited genes from parents. Women with a first degree relative that has been diagnosed with breast cancer at a young age (before 50) are at high risk of developing breast cancer as well; this results in a risk of 2-fold or more of developing the disease [32].

2.3.1.6. Genetics

Genetic predisposition to breast cancer is fast becoming a common practice in aiding both the diagnostic and preventative measures for Breast Cancer[80], [81]. One of the most commonly associated but rare genetic associations in breast cancer is the BRCA1 and BRCA2 genes; these are inherited genes that express a predisposition to breast cancer in 15% of familial cases; presenting a 50-85% increased risk in women. BRCA1 an BRCA2 presents the highest penetrance in familial cases of breast cancer, however several genes have been indicated to present a percentage of penetrance for familial breast cancer but does not explain all [5].

2.3.1.7. Other Related Factors

Additional factors relating to breast cancer include lifestyle factors such as diet, alcohol consumption, exposure to radiation, oral contraceptive use, hormone replacement therapy, smoking and geographical variation [77]. However, these factors do not present as much risk and are disputed in further studies [77].

Contributing factors for breast cancer range from environmental to demographics factors, however the most prominent focus for breast cancer is currently in the genetic association that highlights the underlying pathology of cancer. Genetic predisposition and classification of breast cancer is currently a focus given the increased survival rate and better prognosis associated with early diagnosis; therefore, the following sections focus on genetics and highlight the current research in the area of breast cancer.

2.4 GENETICS OF BREAST CANCER

Familial studies encompass the vast majority of successful genetic discoveries in breast cancer with emphasis being placed in the now well-known BRCA1 and BRCA2 genes. There are currently three established categories of mutations defined as high penetrance, moderate-risk and low-risk. These categories currently include a number of genes indicated in research including ATM, BARD1, BRCA1, BRCA2, BRIP1, CDH1, CHEK2, FANCM, MLH1, MRE11A, MSH2, MSH6, MUTYH, NBN, PALB2, PMS1, PMS2, PTEN, RAD50, RAD51C, STK11 and TP53 genes [5].

As a high penetrance mutation, the BRCA1 gene was first localised in 1990 by Hall et al. [76] who utilised logarithm of the likelihood ratio for linkage, or better known as 'Lod', to ascertain a likelihood ratio ranging from 2000:1 and 1.4×10^{6}:1 among the 23 tested families within the study [76]. From this, further studies were performed, leading to the discovering of the BRCA2 gene by Wooster et al. [82], using similar techniques. Table 2.4 provides approximate estimates for penetrance and relative risk of high and moderate penetrance SNPs [5].

TABLE 2.4: PENETRANCE LEVEL OF ESTABLISHED SNPS ASSOCIATED WITH BREAST CANCER

High Penetrance	**Gene**	**Incidence**
	BRCA1	82% lifetime risk
	BRCA2	82% lifetime risk
	PTEN	85% lifetime risk
	TP53	25% by age 74
	CDH1	39% lifetime risk of lobular breast cancer
	STK11	32% by age 60
Moderate Risk	Gene	Risk in Females (RR)[a]
	CHEK2	1.7
	BRIP1	2.0
	ATM	2.37
	PALB2	2.3
[a] RR; Relative Risk		

Given the advancements in technology and techniques in the area of genomics and bioinformatics, full genome scans are being utilised in studies to outline associations in breast cancer that are represented by much larger cohorts. While previous investigations have presented successes in the area of breast cancer, the limitations to the study are within the cohort size and therefore statistical power. Overall, there are further limitations that exist in the predictive power of Breast Cancer; the main concern being the diagnosis of Breast Cancer in patients whose case would not have become clinically evident [83]. This transpires to cases of diagnosis in patients whose tumours or abnormal cell growth deemed clinically relevant when advanced occurrence would not have resulted in a cancerous presence. The following sections discuss two types of GWAS commonly conducted in Breast Cancer.

2.4.1 Familial-based GWAS

Familial GWAS extends the techniques of Genome-wide studies to family-based populations; with the aim focused on breast cancer associated genetic components that are present as inherited biological occurrences. These techniques allow for larger cohorts and SNP samples in analysis. Further SNPs and genes are outlined for their association with breast cancer due to these large analyses. The FGFR2 gene has been outlined in multiple studies [84], [85] during which further genes, LSP1 and TRNRC9 and MAP3K1 were also outlined [85]. These genes have also been replicated in further studies [86], [87] with further evidence found in other populations [88].

Familial-based GWAS are at a disadvantage to familial-linkage studies as they use a feature of Family History Score to indicate the familial risk rather than analysing the relatives and the inherited haplotypes that are present in both parents and offspring. This restricts the study, as genetic susceptibility could not be measured based on the occurrence of genetic components in both parents and offspring. The largest current familial study undertaken in breast cancer was conducted by Easton et al, in which ~44,000 subjects were included using Family History Score > 2 that indicates the number of and degree of family members with cases of breast cancer [85]. The main advantage of Familial GWAS as opposed to linkage studies is the ability to obtain larger cohorts for analysis, improving the statistical power of the investigation.

2.4.2 Population-based GWAS

While many studies have been conducted in the area of familial occurrences of Breast Cancer, research into the sporadic occurrences of breast cancer in the general population remains less successful. Sporadic occurrences of breast cancer concern the development of the disease outside of the commonly associated inheritance from familial lines. A majority of Breast cancer cases (~66%) are considered to be sporadic occurrences [2]; these cases, while not affected by the established familial genetic mutations such BRCA1 and BRCA2, still adheres to the risk associations such as age and hormone-related factors. Further research has been conducted into sporadic occurrence of the disease in BRCA1 carriers, with promising results (~67% AUC) when utilising the blood signatures of white and peripheral blood cells with serum DNA. Thomas et al [89] produced a multi-stage study for population-based sporadic occurrence of breast cancer that outlined novel risk alleles in chromosomes 1 and 14. Focus of breast cancer in GWAS has been limited with the majority of studies concentrating on the pathology of breast cancer for suggestive genes and SNPs in candidate gene studies or the familial links that are prominent in breast cancer.

2.4.3 Epistasis Studies

Previously mentioned, epistasis association is a developing technique that investigates the role of multiple genetic signatures in respect to the disease, suggesting the interacting components produce the phenotypic expression commonly associated with the disease. Breast cancer has received a lot of attention using the epistasis technique within the past 10 years. However, having been subject to the limitations of epistasis, studies have been focusing their effort on smaller sets of biologically related gene or prior knowledge from previous studies [90].

The limitations of this study type result in many epistasis studies focusing on a dramatically reduced set of SNPs or using a limited 2-way interaction model that only considers the interactions of 2 SNPs in relation to the phenotype. A large-scale analysis of ~89,000 subjects and 75,380 SNPs previously identified via 9 GWAS studies encompassing 10,052 cases and 12575 controls was conducted using two-way SNP interactions [91]. This study yielded few SNPs that exceeded the genome-wide threshold of 1×10^{-8} but concluded more SNPs with 1×10^{-6}. Further studies have been conducted in association with Breast Cancer, using reduction parameters for SNP dimension such as pathway analysis. Pathway analysis considers the pathology of the disease and uses these genetic signatures to conduct an epistasis study. Using DNA repair, modification and metabolism related pathways, Sapkota et al [92] identified 2-way SNP interactions that yielded a result of $<7.3\times10^{-3}$, however this again uses a two-way interaction model which may not confer the risk that is associated with a group of interacting SNPs across genes or chromosomes.

2.4.4 Genetics of Breast Cancer Summary

The development of breast cancer in patients is still mostly undetermined, considering the most promising genes of high penetrance only explain ~5-20% of familial cases (~33%). Sporadic occurrences of breast cancer form the majority of diagnosed cases; therefore, they present a unique challenge as opposed to the familial occurrence of breast cancer. Breast cancer is defined as a complex disease, which suggests that there are multiple factors effecting the development of the disease in subjects; this may relate to factors such as environmental that result in mutations in genetic components. Environmental factors present a challenge in themselves, as it is a difficult task to identify/ measure the significance of association to factors in the environment that may seem related, but aren’t, or may seem unrelated, but are.

2.5 DISCUSSION

Explored in this chapter were the fundamental basis of knowledge that build the core aspects of this thesis. The fundamentals of genetics introduce and concentrate on the chain of events that lead to phenotypic expression; providing a summary as to the importance of the chosen biological material for analysis in this thesis. The choice of phenotype is due to the expanse of publications that indicate a genetic component to Breast Cancer; the purpose of this thesis is in validating a methodology therefore choosing a phenotype that has a genetic importance. An additional reason for the use of Breast Cancer is its prominence in media and research, which as a result has led to many studies in the area increasing the choice of data and the sample size.

Study approaches concern the study design and influence the intention of the analysis. Previously discussed in 2.2 are the limitations and advantages of each approach. Incorporating the chosen phenotype directs the choice of approach; sporadic cancer concerns the development of the phenotype outside of familial relatedness therefore excluding the familial study approach. To consider the approaches GWAS and Candidate that approach a problem similarly but contrast in the inclusivity of features, a gap in publications that concern GWAS with sporadic cancer provide an area of interest for this thesis. Further to this, the restrictive nature of the candidate study approach does not complement the second approach of Epistasis. Epistasis is a progressive approach that has been developed in response to the hypothesised phenomena of genetic interactions, a core subject of this thesis. With the combination of GWAS and Epistasis, the limitations of epistasis are complemented by the process of GWAS. The computational complexity of Epistasis is a drawback of the approach that forces a range of solutions to be considered, one of the most common and best performing solutions is feature filtering for dimensionality reduction. Performing dimensionality reduction with the GWAS approach reduces the feature set to significant features and therefore significantly reduces the dimensionality of the data. However, there are still limitations that need to be addressed including a high false discovery rate and SNP selection, this is discussed throughout the remaining chapters.

Chapter 3: TECHNIQUES & METHODS

In the previous chapter, the core aspects of the thesis were discussed and outlined the study design. To summarise this, with a focus on sporadic occurrences of Breast Cancer using a chained GWAS to Epistasis approach, presented is an advantageous combination that are complementary. Given the methodology built from the research so far, three general topics of analysis need to be addressed: Quality Control, Association Analysis and Inference Methods.

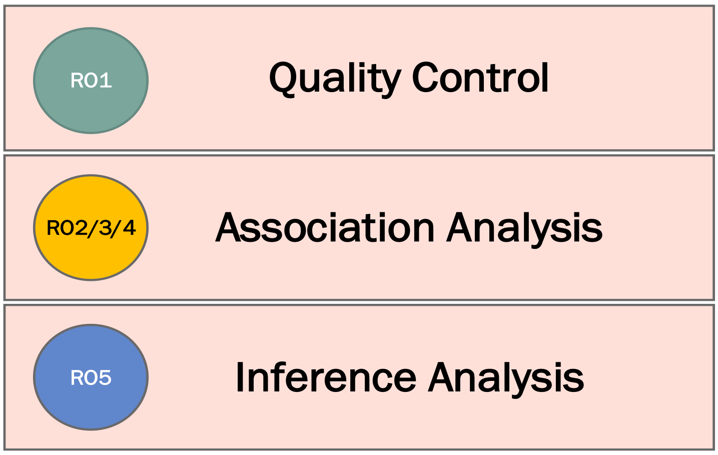


FIGURE 3-1: DEPICTION OF THE SECTIONS OF TECHNIQUES AND METHODS DISCUSSED WITHIN THIS CHAPTER ALONG WITH THE RESEARCH OBJECTIVES BEING NAVIGATED.

Bioinformatics is a substantial field that has been benefitted by continuously advancing techniques and adaptations that suit a variety of biological material including SNP's; this results in an abundance of techniques and methods that are applicable to the current study design. A factor that is critical in this methodology, is the use of techniques and methods that are interpretable and therefore will be a demanding aspect to drive the decisions of this chapter.

3.1 QUALITY CONTROL

The following section supports ***RO1***; outlining the standard processes that are considered and applied in data quality control for genetic material, with focus on SNP data. Methods and techniques outlined in this section are standard practice techniques that are used to remove both subjects and SNP features that can cause bias, obstruct or mask signals, or produce false positive results [93].

3.1.1 HETEROZYGOSITY

Heterozygosity is a parameter often measured in the early stages of genetic variation studies [94][95][96]. This measure is used in this capacity as an indication for inbreeding or severe effects in populations to ensure that individuals with reduced or excessive rate of heterozygous genotypes are identified for removal [97]. The presence of these individuals could indicate contamination, inbreeding, outbreeding or poor genotype calling in the sample. Plots in Figure 3-2 are produce using the number of non-missing genotype (N_{NM}) and the observed number of homozygotes (O_{HOM}).

$$\text{Heterozygosity} = \frac{N_{NM} - O_{HOM}}{N_{NM}} \qquad \text{EQ. 3-1}$$

Plotting this, the x-axis represents the proportion of missing genotypes for an individual while the y-axis represents the observed heterozygosity. 'Rule of thumb' indicates that individuals that deviate outside of 3σ are excluded from further analysis [97]; however, this threshold will be dependent on study factors, such as cohort size and focus disease.

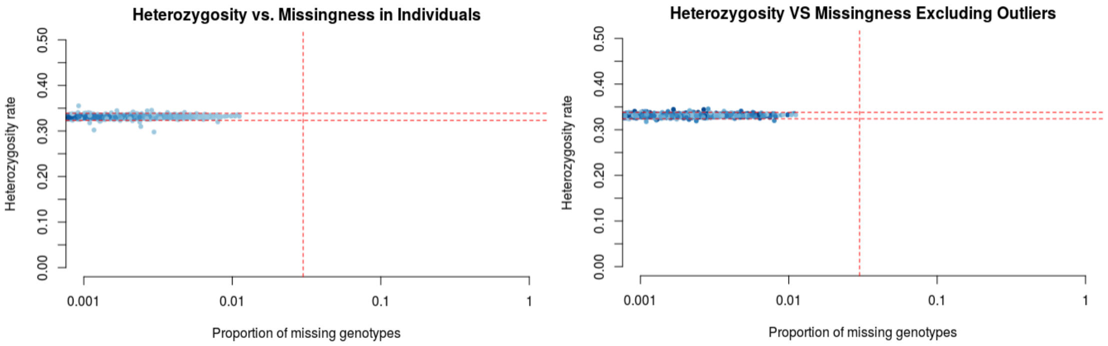


FIGURE 3-2: HETEROZYGOSITY × MISSINGNESS IN GENOTYPE PLOTS

Example heterozygosity rate plots for removing outlier subjects (a) original data (b) outliers removed. Threshold lines on x-axis indicate the maximum level of missingness for genotypes in individuals, while the y-axis shows the outlier threshold for heterozygosity rate separated by 3σ from average.

3.1.2 Sex Inconsistencies

Sex inconsistencies is common method adopted in GWAS [98][99][46] that refers to a discrepancy between the recorded sex of a subject and the heterozygosity rate of X chromosome [97]. There are distinct differences between the genetics of males and females, in reference to the sex chromosomes (XX/ XY), leading to an evaluation method which considers the heterozygosity in the X chromosome, indicating a heterozygosity rate of >0.8 for males and <0.2 for females [97][93]. Therefore, subjects whose heterozygosity rate falls with the threshold of 0.2 < x > 0.8 will be excluded from further analysis. The heterozygosity rate of individuals, in respect to the X chromosome, is analysed using fixation indices, or more commonly known as F-statistics, a statistical analysis to measure the expected level of heterozygosity against the observed level using the following equation [100]:

$$\frac{\text{X Chromsome Heterozygosity}}{\text{Homozygosity Rate}} = \frac{\sum_{i=1}^{n} p^2 + q^2}{n} \qquad \text{EQ. 3-2}$$

As seen in Heterozygosity, a coefficient result of <0.2 is expected for females while males are expected to have a coefficient of 1, however leniency is adopted to account for genotyping error discrepancies [93]. For individuals that are flagged for issues between the recorded sex and F coefficient, the coefficient should be scrutinised to ensure that an error has not been made during the record of information. While sex checks are only essential to the analysis process when basing features or control on the sex of the individuals, they also provide a unique processing control that considers issues that exist in gametes such as Turner and Kleinfelter syndrome or mosaic individuals [101]; sex chromosome anomalies that while present normal phenotypically, the genotypic presence is abnormal.

3.1.3 Relatedness & Duplicates in Subjects

Generally in clinical studies, the relatedness of individuals will be recorded and noted for research purpose [97]; however in cases where either subjects are unaware of attending related subjects in the study or are unaware of their relation status to other subjects in the study; subjects are excluded to avoid a bias analysis. Similarly, any duplicate cases, case in which the genetic material of a person is included in the same more than once, will be removed. Using case-control cohort, as demonstrated in this study, measures must be put in place to ensure that individuals included in the study are unrelated. When testing for relatedness in individuals, pairwise identity-by-descent (IBD) is used [102] to tests pairs of subjects to compare the reported relatedness of individuals against the proportion of loci with which two individuals share one, two or zero alleles. As provided in Table 3.1, relatedness in subjects is measured by the number of matching allele states i.e. Z score = 2 when individuals both show the same homozygous state, AA.

TABLE 3.1: REPRESENTATION OF RELATEDNESS MEASURED BY IBS

Subject-1	Subject-2	IBS State/ Z
AA	AA	2
AA	Aa	1
AA	aa	0

Calculated during IBD in PLINK is the PI_HAT value; a value that can be demonstrated as [102]:

$$PI\_HAT = P(IBD = 2) + \frac{1}{2}P(IBD = 1) \qquad \text{EQ. 3-3}$$

This value provides an overall estimate of the relatedness of 2 individuals and as such can be used as exclusion criteria using the following thresholds [93]:

TABLE 3.2: RELATEDNESS PI_HAT SCORE WITH REFERENCE TO DEGREE OF RELATION

Relatedness	PI_HAT Value
1st Degree (Siblings)	0.5
2nd Degree (Half Siblings)	0.25
3rd Degree (Cousins)	0.125

Applying an upper threshold of 0.125 will exclude outlier subjects that appear to have relations up to and including 3rd Degree (Cousins) within the cohort.

3.1.4 DIVERGENT ANCESTRY

Diverging Ancestry is another problem that occurs in case-control data; individuals report their ancestry and are analysed based on these clustered groups. Ancestry in genomic analysis is important given the bias that can occur and false positive association which arise due to spurious associations that are a result of differences in ancestry rather than case-control [103]. It is established that the susceptibility to immune-related disease varies wildly between populations [104]. These differences can cause health disparities due to phenotypic diversity pertaining to the genetic differences that are evident between populations. Therefore, the exclusion of individuals that deviate from the population cluster can aid in creating a representative set of features and observations for further analysis. As a standard practice element of genetic analysis studies [98][46] ancestry divergence is commonly performed using a Principle Component Analysis (PCA) technique as outlined below.

For this study, PCA based approach is used with the HapMap3 reference panel for YRI, CEU and CHB+JPT referring to and using tools outlined in [93]. Figure 3-3 provides a visualisation of the initial PCA model which is represented by YRI (Yoruba, Ibadan, Nigeria) (Green), CHB (Hans

Chinese, Beijing, China) + JPT (Japanese, Tokyo, Japan) (Purple) and CEU (Utah residents with Northern & Western Europeans ancestry from CEPH collection) (Red), case and control subjects from the acquired data cohort are represented as black and blue, respectively.

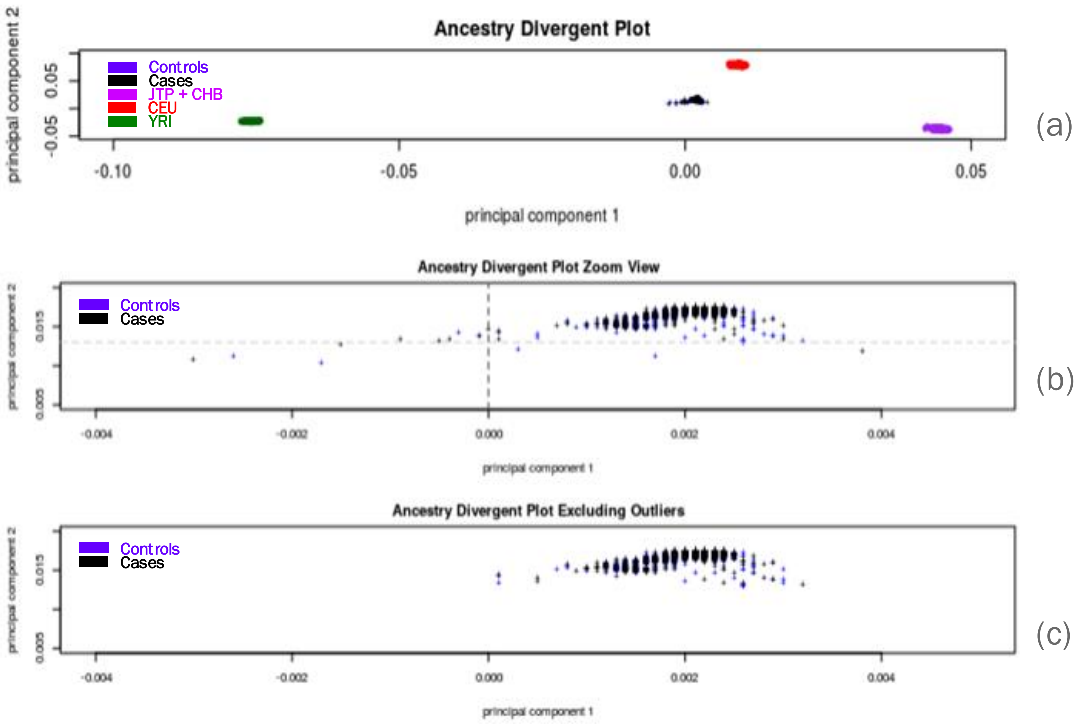


FIGURE 3-3: ANCESTRY DIVERGENCE PCA PLOT EXAMPLE SHOWING THE DISTRIBUTION OF INDIVIDUALS IN RESPECT TO PCA1&2.

(a) Overall distribution in relation to populations JPT+CHB, CEU and YRI. (b) Zoom view of the distribution with threshold lines associated to the x and y axis (c) Distribution after outliers have been removed.

Figure 3-3 shows the acquired data clusters centrally in the plot indicating its separation from the outlined populations. The second plot provides a zoom view of the population plot to inspect the distribution, which shows there is a minor spread that extends outwards. Threshold are set to remove these outliers to improve the cohort reliability.

3.1.5 LINKAGE DISEQUILIBRIUM PRUNING

Linkage Disequilibrium (LD) pruning refers to a process that removes SNPs based on the correlative effect between two loci. The purpose of LD relates to the genomic population structure, considering the evolution of the genetics within. The genome is separated into Chromosomes (although the initial use of LD was analysed cross Chromosomes, this is no longer necessary as it is considered that associative haplotypes are inherited via chromosome blocks). Within the chromosome, each pair of loci are analysed using varying parameters of window size, step size and correlation threshold [21]. From here, any loci that present a LD greater than a specified threshold are eliminated. Commonly the process of LD adheres to the following equation [21]:

$$D_{AB} = P_{AB} - P_A P_B \quad \text{EQ. 3-4}$$

Where P_A refers to the proportion of allele A, P_B refers to the proportion of allele B and P_{AB} refers to the proportion of allele A and allele B occurring together. Linkage Disequilibrium is in occurrence if:

$$D \neq 0 \quad \text{EQ. 3-5}$$

The very basics of an LD pruning analysis will concern a window size, a step size and a threshold [102]. The window size refers to the number of SNPs considered at one time for analysis, while the step size indicates the number of SNPs that are used to shift a window of focus SNPs for analysis. In this motion, the analyses will advance across the chromosome until the full length of SNPs within the chromosome have been analysed. The threshold is a given measurable quantity with which the LD between two loci will exceed if exclusion is necessary; this threshold can use various correlation equations to determine the basis of LD including D’, r 2 and Variance Inflation Factor (VIF)[102]. At this point, we will be focusing on the standard approaches, r^2 and VIF, most commonly applied in literature. The following equations outline the details of r^2 and VIF [21]:

$$r^2 = \frac{D^2}{P_A(1-P_A)P_B(1-P_B)} \quad \text{EQ. 3-6}$$

$$VIF_i = \frac{1}{1-R_i^2}$$
$$\text{where } i = 1\ldots\text{n} \quad \text{EQ. 3-7}$$

The threshold r^2 provides an analysis in terms of the pairwise genotypic correlation while VIF focus on the variance dependent on the collinearity found between the loci [102]. Each of these candidates rely on different threshold inputs ranging from 0-1 (r^2) [21] or 1 and above (VIF) [105], depending on the stringency of the analysis. Of course, the more stringent the threshold, the more SNPs will be excluded, while this would leave the remaining candidate dataset with predominantly independent loci, this may also exclude SNPs that are highly associated with the disease. In essence, the use of LD within an analysis concerns the evolutionary path of human genomics and the effects on the genotyping information within [106]. We want to eliminate any genetic data that may present itself as ‘noise’ within our dataset and as such aim to provide the most promising and prominent genotypes associated with the focus disorder. A common threshold of $r^2 < 0.6$, or $r^2 < 0.8$ are applied in studies [95][98].

3.1.6 GENOTYPE IMPUTATION

Genotype Imputation is currently considered common practice in GWA analysis. Imputation is an applied method that will enrich a genotype dataset with the most similar genotype information given in the reference sample set [107]. Firstly, it is important to note that in common applications of imputation; missing values are imputed based on given values already present. Genetic imputation uses the values present in the data to impute new columns of data; that is, the imputation of data into genomic datasets will add new SNPs that were not present previously [107]. Given the introduction of data that is imputed using probabilistic methods for 'new' data, it is important that the accuracy of these methods be fully explored, not only for the accuracy of the data being imputed but also for its necessity in the methodology proposed. Due to this, required sizes of cohort for statistical power in genomic studies when sequencing the genetic data are expensive, so in order to reduce this cost and allow a more feasible genomic analysis, researchers adopt the imputation technique.

The imputation process considers the genome within Chromosomes (Chromosome 1, Chromosome 2, etc), from there haplotypes are generated from tag-SNPs located within the genotype information provided [107]. Haplotypes are set genetic determinants located on a Chromosome, simply put this is a group of genes that are inherited together from a single parent. This could be a pair or set of genes that have been inherited together. From here, this information is used to impute genotypes for an individual based on the most similar reference haplotype(s) to insert markers/ alleles from the corresponding reference haplotype into the sample Haplotypes [108].

While imputation introduces an inexpensive and quick way of populating genomic feature sets, there are implications and disadvantages to the method. The samples included in reference datasets are not known, however they are presumed to be predominantly healthy individuals, which introduces questions such as; how accurately these reference sets can predict the missing genotypic information for an affected disorder genotype set? Imputation is only as good as the provided sample set; this refers to the imputation of markers based on the current data that is already sequenced within the dataset. Therefore, there are a few considerations that should be explored in order to determine whether imputation would beneficial to the study; these considerations include the current sequenced information, the reference dataset being used and the sequencing chip that the genotypes were sequenced with [109].

The provided dataset for imputation will only provide further genotype information based on the correlative tag-SNPs already available, therefore if there are tag-SNPs missing within a haplotype, imputation cannot be performed in that haplotype e.g. the APOE locus associated with Alzheimer Disease cannot be imputed from the Affy 500k chip [110]. This introduces the issues that concern sequencing chip options; while it is not always possible to choose which chip is used for the sequencing of data, there are advantages and disadvantages between them; the main and most significant difference is the outcome of sequenced information that differs between them.

Finally, the reference dataset is a consideration that is most likely to affect the results of the data. Various datasets can be utilised dependent on the software in use. For example, PLINK utilises the HapMap reference dataset [102]. The datasets each provide a different sample of individuals however, some are shared e.g. HapMap reference and 1000 genomes share some individuals [110].

Another consideration concerning the process of imputation is the software that is used to perform the imputation. There are many open source options readily available for use however; we are considering the main open source options for imputation. MACH [111], IMPUTE2 [112] and BEAGLE [113] are in competition; each proving advantages and disadvantages when compared using criteria such as the memory consumption, runtime, prediction quality and error handling. The use of this software should be considered based several factors such as dataset format and size, reference panel for imputation, operating system used for analysis and the output from each software for association analysis. As demonstrated in [114], imputation's correct rate reaches >95% when using many different software's, of which the previously mentioned are some of the best performers.

3.1.7 Threshold Measures

Threshold Measures are standard practice applied as a base or final control process that removes individuals and SNP features. These quality control measures are important and are performed as a standard practice in any genomic study that utilises SNP or related data [99][89][96]. Figure 3-4 provides a visual representation of the analysis of the below threshold measures.

3.1.7.1. Missingness in Individuals (MIND)

This threshold considers the missingness per individual for genotypic data. When we analyse the data further, we need to include individuals whose data is mostly, to entirely complete [115]. This threshold will remove any individuals whose data has a missing rate higher than the provided threshold. For this threshold we used a measure of 0.01, when ensures that every individual included in the dataset has at least 99% of their genotypic data.

3.1.7.2. Hardy-Weinberg Equilibrium (HWE)

The Hardy-Weinberg Equilibrium is a simple equation that is used to discover the probable frequency of genotype in a given population and track the evolutionary changes from generation-to-generation. When performing an association analysis or further testing in a genetic cohort, a control group is used as a representative sample of the given population and therefore if there are evolutionary migrations in the set, it could cause bias or incorrect results in the next stage. The Hardy-Weinberg equation is demonstrated as:

$$\begin{gathered} p = AA + \frac{1}{2}Aa \\ q = aa + \frac{1}{2}Aa \\ p^2 + 2pq + q^2 = 1 \end{gathered} \tag{EQ. 3-8}$$

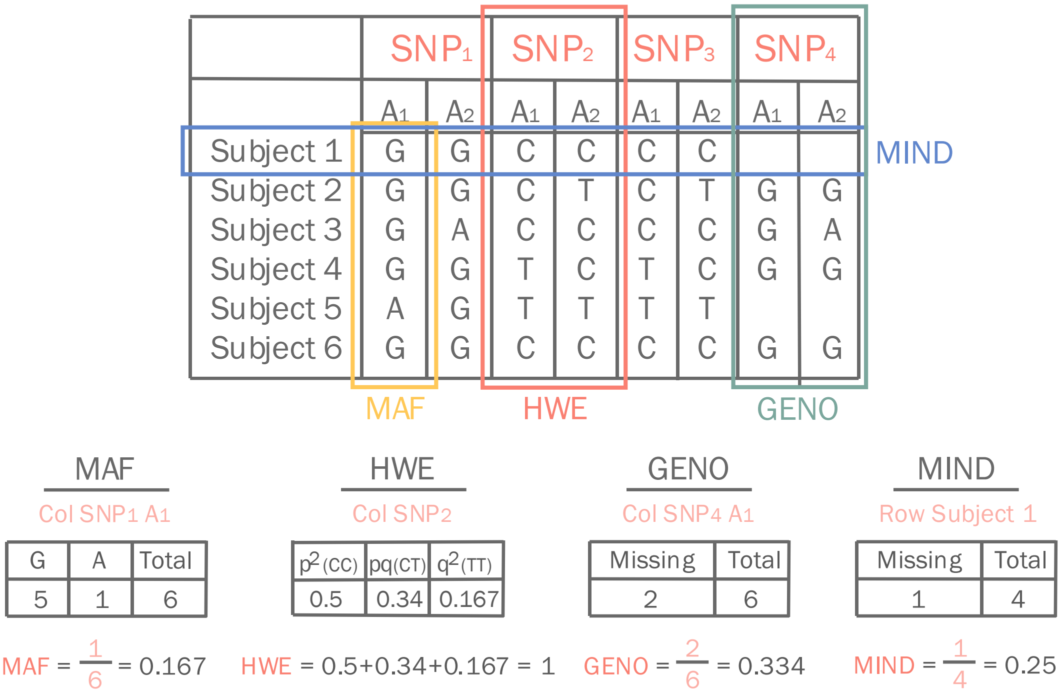
Standard approaches will use thresholds that range from 10^{-5} to 10^{-6} in case-control studies [115]. This process identifies issues that can relate to cryptic relatedness, genotyping error or population admixture and more [115].

3.1.7.3. *Genotype Call Rate (GENO)*

Genotype rate (Geno) threshold concerns the marker genotyping efficiency relating to the SNP assays performance rate; in particular, this considers the percentage across all individuals pertaining to the missingness of SNP information [115]. This may reduce the marker set but will also improve the reliability of the results. Standard practice employs a threshold between 0.01 to 0.05 lower limit that will exclude any markers that have more than 1-5% missing information, less than 99-95% call rate confidence.

3.1.7.4. *Minor Allele Frequency (MAF)*

Minor Allele Frequency considers the alleles of each SNP across all subjects; if this SNP is present in less than the specified threshold, then it is removed [97]. When we use the MAF threshold, we consider the presence of the minor allele as the potential risk and therefore we would like to reduce dimensionality by removing any SNPs that are unlikely to yield results. e.g. a MAF of $r^2 = 0.01$ in a dataset of 500 subjects would result in SNPs whose minor allele is not present in more than 1% of the cohort, 5 people, being removed from set.



	SNP_1		SNP_2		SNP_3		SNP_4	
	A_1	A_2	A_1	A_2	A_1	A_2	A_1	A_2
Subject 1	G	G	C	C	C	C		
Subject 2	G	G	C	T	C	T	G	G
Subject 3	G	A	C	C	C	C	G	A
Subject 4	G	G	T	C	T	C	G	G
Subject 5	A	G	T	T	T	T		
Subject 6	G	G	C	C	C	C	G	G

MAF

Col SNP_1 A_1

G	A	Total
5	1	6

$MAF = \frac{1}{6} = 0.167$

HWE

Col SNP_2

p^2(CC)	pq(CT)	q^2(TT)
0.5	0.34	0.167

$HWE = 0.5+0.34+0.167 = 1$

GENO

Col SNP_4 A_1

Missing	Total
2	6

$GENO = \frac{2}{6} = 0.334$

MIND

Row Subject 1

Missing	Total
1	4

$MIND = \frac{1}{4} = 0.25$

Figure 3-4: Visual representation of analysis conducted for threshold measures [A]

[A] Column highlighted in yellow provide the input for MAF calculations. Columns highlighted in pink provide the input for HWE, with p^2 referring to genotype CC (3/6), pq referring to genotype CT or TC (2/6) and q^2 referring to genotype TT (1/6). Columns high in green provide the input for GENO. Columns highlighted in green provide the input for MIND.

3.1.8 *Quality Control Summary*

Quality control is an essential process that can reduce bias, improve signal strength, reduce false positive signals but more importantly, it addresses issues that prove to be prominent pitfalls when analysing and processing data. Quality Control processes Heterozygosity rate, Sex Inconsistencies, Relatedness and Duplicates in Subjects, Divergent Ancestry and Threshold Measure all provide advantageous effects for utilisation in the proposed methodology. By removing individuals and SNPs that do not surpass the standards set by these methods, the dataset is being reduced to a representative set of SNP features and subject cohort that are more likely to present underlying genetic signals in association with the phenotype.

Techniques that have not been adopted into the proposed methodology are Imputation. With the advantages that benefit a GWAS study primarily associated with the increase in feature set, the expansion of this dataset directly effects the primary objective analysis, epistasis. By increasing the number of SNPs included in the analysis, not only would the computational complexity be increased but also the time taken to perform the analysis. Further to this, imputation imputes features directly linked to tag SNPs already established in the dataset; these features are unlikely to produce signal effects that differ dramatically from their original source but would introduce more ‘noise’. Therefore, for the current methodology, imputation would cause more issues than would introduce benefits. The choice of methods and techniques to process the feature dataset are selected to produce a feature set optimised concerning reliability and robustness for further analysis in the proposed methodology.

3.2 Association Analysis

Association analysis is a very broad term that encompasses a variety of approaches from statistical filtering to relationship modelling using univariate and multivariate data. In the context of these research, the following section outlines a variety of methods that perform association analysis in relation to genomic data with reference to RO2, RO3 and RO4. Further to this, discussed are techniques and methods that are used for correction in analysis for limitations and issues that arise in particular from the application of such methods to large datasets, such as genomic data.

3.2.1 *Univariate Analysis*

Univariate analysis encompasses methods that measure associations between variable X and response Y, in the context of the research an example would be to measure the association of one genetic variant to the expressed phenotype. As previously mentioned, the analysis of the association is not a definitive answer to a question and therefore is commonly used as a supporting method to either statistically filter data or to indicate potential relationships for further analysis. This section outlined methods and techniques for the exploration of RO2.

To discuss one of the most prominent univariate methods, standard GWAS is a widely used approach that has been indicated as a powerful alternative to traditional linkage-studies [112]. GWAS has previously been discussed in 2.2.1 and at this point it is important to outline that GWAS approaches vary between studies and in order to outline techniques that are related to the current approach, the scope of GWAS is defined as the following:

> *A genome-wide association study is an approach that involves rapidly scanning markers across the complete sets of DNA, or genomes, of many people to find genetic variations associated with a particular disease.*
> *- National Human Genome Research Institute* [116]

In standard practice GWAS as is commonly conducted, a univariate analysis of the data will be performed using a permutation test on a 2x2 or 2x3 contingency table (refer to Table 3.53.2.1.1.) that transforms the data to manipulate the focus. This will commonly be performed by fisher’s exact test, chi-squared [13][102] or logistic regression [84]. Further adaptations of the GWAS approach include simulation techniques to deduce empirical statistics of causal variants based on control haplotype frequencies [117][118] and incorporating multiple traits [119]. Further to this, due to the release of datasets and information, access to this information is much more accessible and has given rise to a GWAS adaptation called ‘meta-analysis’ that use data from multiple studies to gain insight, particularly in complex diseases [120].

3.2.1.1. Genetic Models

The following section discusses the association models that are commonly utilised during univariate analysis. Due to the nature of SNP data we must consider the variable presence of features in both allele and genotype form, further explained in 0. Due to the different combinations that are available for testing purposes, there are a number of models that can be used. An **Additive model** is used to determine the disease penetrance of a given SNP, that is, the risk of disease in subjects carrying a given genotype. This is assessed using code 0, 1 and 2, which corresponds to the γ-fold risk increase with each additional genotype presence. Table 3.3 presents the assigned risk codes and corresponding genotypes. Similarly, the **dominant model** also measures the disease penetrance of a given SNP but rather than using the genotype states separately, it combines Aa and aa to produce a model which assesses the risk of genotypes that do not contain a dominant allele. The **recessive model** again measures the disease penetrance as demonstrated by the dominant model, but rather than measuring the disease penetrance of the SNP in subjects that do not hold a dominant homozygous genotype, the focus is instead on recessive homozygous genotype status.

TABLE 3.3: DATA TRANSFORMATIONS FOR DISEASE PENETRANCE MODELS

	Penetrance		
	AA	**Aa**	**aa**
Additive	0	γ	2γ
Dominant	0	γ	γ
Recessive	0	0	γ
Multiplicative	0	γ	γ^2

Allelic model uses an OR calculation and utilises the allele form of the SNP by measuring the 'Odds of Disease'. The 'Odds of disease' is the probability that a disease is present compared to the probability that a disease is absent; while this cannot be directly measured, the 'Odds of Exposure' can be; this uses the frequencies of exposure in case and control. The allelic model measures the association between the Odds of Disease in subjects with dominant allele A over the odds of disease of subjects with the recessive allele a. This is represented in Table 3.4 to Table 3.5 where a, b, c, d and T refer to the values within the table.

TABLE 3.4: ALLELIC MODEL CODE REPRESENTATION

Allele 1		**Allele 2**	
A	a	A	A
1	0	1	0

TABLE 3.5: ALLELIC MODEL OR

	Allele 1	**Allele 2**	
Case	a	b	a + b
Control	c	d	c + d
	a + c	b + d	T

The **Multiplicative Model** produces an OR similar to the allelic model, assuming that the genotypes represent in increasing risk with each additional minor allele [121]. Similar to the additive model and commonly referred to as the log-additive model, the multiplicative model assumes a greater risk from risk homozygotes as demonstrated by the n^2 risk increase. It should also be noted that there are some inconsistencies in the literature regarding the disease penetrance models [122], the above mentioned disease penetrance models are chosen to optimise the investigation of risk alleles [123]. Further to the above-mentioned models is the commonly used logistic regression that builds upon the additive model. Logistic Regression is explained further in the next section.

3.2.1.2. Logistic Regression

The logistic regression model is a renowned statistical method that commonly utilises binary predictors to estimate the parameters of a model, in simple terms, it provides a co-efficient that indicates, based on the inputted data, that the presence of a risk factor increases the odds of a given outcome by factor x.

The logistic regression curve is given as:

$$P = \frac{e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}}{1+e^{\alpha+\beta_1 x_1+\beta_2 x_2+...+\beta_n x_n}} \qquad \text{EQ. (3-9)}$$

To give us probability, P, with coefficients β_i using predictors, x_i.

The logistic regression model can be utilised to produce coefficient values for a variety of data types including categorised (multinomial logistic regression), ordinal (Ordinal Logistic Regression) and continuous (Linear Regression), the use of these models depends on the dependent variable data type. Further varieties of data types can also be used within the analysis using logistic regression by employing adjusted log-odds ratios e.g. Binary with continuous feature. The logit model produces the coefficients, β. Each coefficient is generated from one feature across all observations. In this instance, the following derivations produce β coefficients based on binary features.

$$logit(p(x)) = \log\left(\frac{p(x)}{1-p(x)}\right) = \alpha + \beta x \qquad \text{EQ. (3-10)}$$

Log-Odds is used to estimate the value of β (See 3.3.1 for more information on Odds Ratio).

$$\begin{aligned} \beta &= \log it(p(1)) - \log it(p(0)) \\ &= \log\left(\frac{p(1)}{(1-p(1))}\right) - \log\left(\frac{p(0)}{(1-p(0))}\right) \\ &= \log\left(\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}\right) \\ &= \log(OR) \end{aligned} \qquad \text{EQ. 3-11}$$

For this, we firstly generate α by using ‘unexposed’ binary observations. This is the proceeded with β using the ‘exposed’ binary observations which provide a nominal factor value of how much the odds of given outcome are increased by the presence of this risk factor. Where x = 0, β coefficients are not produced, providing a value for the intercept, the expected mean value of Y.

$$\begin{aligned} x &= 0(unexposed), \quad logit(p(0)) = \alpha + \beta(0) = \alpha \\ x &= 1(exposed), \quad logit(p(1)) = \alpha + \beta(1) = \alpha + \beta \end{aligned} \qquad \text{EQ. 3-12}$$

Logistic regression presents an option to geneticists to incorporate the effects of covariates such as gender, smoking, lifestyle and age. This controls for any bias that these covariates may have on the data particularly when concerning complex traits.

3.2.1.3. Genomic Thresholds

Within genomics there is a constant debate surrounding the threshold measures of significance [124],[125],[14]. These thresholds are advisable in genetic analysis studies, particularly SNP studies, to outlined significant results however the debate discusses the issues that are present with a threshold that is too conservative and those that are present with a threshold that is too lenient. A standard process in GWAS is to use the genome-wide significance threshold that is based on Bonferroni correction [19] under the assumption of a feature set of 1 million SNPs. As a result of plentiful publications claiming significance in genetic markers; the genome-wide significance threshold was set to 1 x 10^{-8} [125]. This ensured that any studies undertaken would adhere to the threshold and only results that showed a significantly high genetic variant would prove reliable for further study and replication.

However, with many studies this does not prove to be the most effective solution. [125] introduces research which explores the p-values of genotype-phenotype associations of previous studies investigating the potential of replication between the widely-accepted threshold of 1 x 10^{-8} and the borderline threshold of 1 x 10^{-7}[19]. Associations falling within these thresholds are deemed to be borderline associations. Depending on the study, borderline associations may be included in the final set of candidate variants as research suggests that variants that fall into the borderline category may also show significance to the phenotypic trait [125]. The results of this study suggested that borderline associations can prove significantly replicable for numerous phenotypes.

Many papers [126][19][127] have also suggested that the restriction of the genome-wide significance threshold could lead to the suspension of research concerning genetic associations to demonstrate true significance with a given phenotype. However, this claim is also disputed in numerous papers which suggest that lower thresholds do not produce significant associations [125]. Publications discussing the threshold of significant associations review the threshold based on the empirical measure of p-values. This paper explores the potential for true association in genetic variants when adopting various threshold measures for the genome-wide significance line. Considering the controversy around the subject in relation to p-values [128], [129], further thresholds are explored in relation to multiple testing in section 3.2.3 .

3.2.2 Multivariate Analysis

Multivariate analysis refers to the statistical analysis of features X_1, X_2, X_3, … X_n and their combined effect on response Y e.g. to measure the statistical significance of an age, height and a particular genetic variant in reference to an expressed phenotype. Multivariate analyses pose a particularly challenging problem when using large datasets as the number of features that are included can become very computationally expensive. Multivariate analysis encompasses a variety of approaches that consider hypothesis regarding covariates, environmental factors and interactions. In order to remain within the scope of this research, the following section focuses on the multivariate analysis approaches that can and are used for interactions, pattern detection and relationships between features. This section outlined methods and techniques for the exploration of RO4.

Epistasis as previously discussed, is an increasing presence in genomic studies. The potential of combinatorial genetic signals as significant features is also more likely in complex diseases given the varying phenotypic expression of symptoms in these diseases. Many complex diseases are deemed ‘umbrella’ terms that cluster the most common symptoms into one disease while the varying of additional symptoms suggests potential interaction among the underlying genetic aetiology. The complexity of performing an epistasis study lies in the combination arrays required to test all possible combinations, an exponential increase in computation complexity and time. The following sections discusses the two main approaches to perform epistasis; Pairwise-Interactions detection and limitless-arity.

3.2.2.1. Pairwise SNP Interaction Detection

Due to the prominent issue of computational complexity in epistasis approaches, a solution that is presented for many techniques to consider the interactions within the genome is to instead use a pairwise interaction method. Using this, an exhaustive approach can be successfully performed without high performance computing hardware, which analyses and measures probability of association to phenotype between every pair combination within the input features as visualised in Figure 3-5 considering 4 SNPs, analysis for interactions with SNP 1.

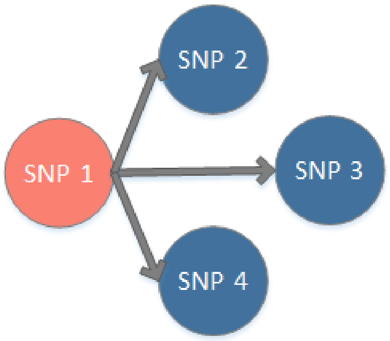


FIGURE 3-5: PAIRWISE INTERACTION ANALYSIS REPRESENTATION

While the computational complexity of the model is significantly reduced by limiting the combination factors to 2, the problem still exists given the size and required representations of the data. There are numerous solutions that exist to approach this problem.

PLINK employs a logistic regression model that corrects and adjusts for multiple testing using the Bonferroni multiple-test correction. While this method adjusts for the errors introduced from high dimensional data, the consequence is the reduction of type 1 errors and the increase in type 2 errors given the conservative approach of the Bonferroni. Further to this, this method is costly in time and therefore the logistic regression model has been denoted for its unsuitability in handling genome-wide datasets [13]. While logistic regression has been outlined as inappropriate for use in genome-wide datasets, variations on the model have been employed successfully in Least Absolute Shrinkage and Selection Operator (LASSO) [130].

Boolean Operation-based testing and screening (BOOST) transforms the data into binary representation to improve time and space efficiency by using language that is closer to machine code [131]. The operational measures are conducted using contingency tables and applying Fisher's exact test and using a non-iterative approximation of the log-likelihood ratio, Kirkwood superposition approximation (KSA) [13].

3.2.2.2. LIMITLESS ARITY DETECTION

Multifactor Dimensional Reduction (MDR) is a non-parametric and genetic-model free data mining strategy used with discrete data for the prediction of discrete outcomes. This machine learning model approaches the area of epistasis using feature extraction to define a new attribute through a process called 'constructive induction' by pooling (combining SNPs as genotypes). Using Multi-locus genotypes, it produces a ratio of case and controls for each genotype and compares this to the overall ratio of case and controls, assigning a status of either high, G_1, or low risk, G_0, depending on whether the defined ratio of the genotype exceeds that of the overall ratio.

MDR relies on the transformation of the representation space for easier classification using ML models to detect attribute dependencies [132][2]. Since its introduction in 2001, the MDR method has been further developed to incorporate additional techniques and models [133] to improve robustness [134], data flexibility[135] and adaptations to concede to various data imperfections such as missing data and status imbalances [136].

While there has been a vast contributing community for the improvement and development of MDR techniques, limitations still exist. In its native form, the consistent issue in epistasis of computational complexity is still present. While MDR provides a method for non-parametric and genetic-model free performance, the computational intensity still performs at an exponential time expense [13] and further to this, the arduous task of performing MDR on more than 2 SNPs requires the steps to be repeated for each model size [13].

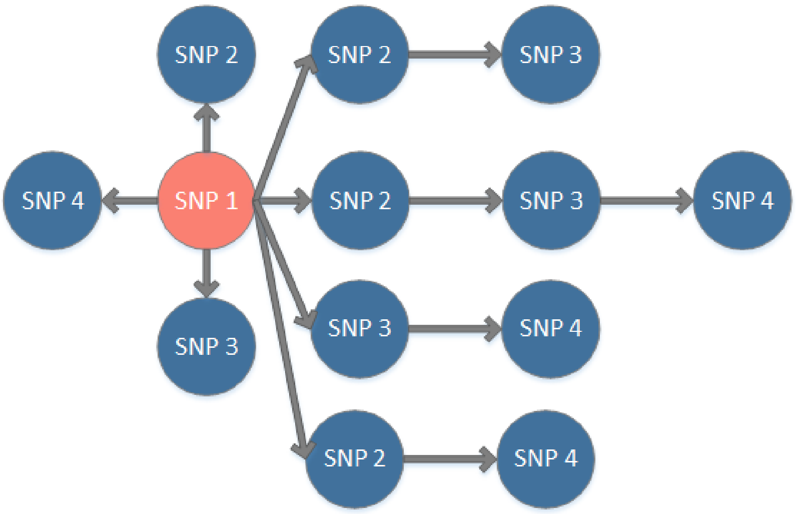


FIGURE 3-6: LIMITLESS-ARITY INTERACTION ANALYSIS REPRESENTATION

While association rule-mining (ARM) is a widely used method in ‘market-basket’ research, the fundamental applications are transferrable to genomic data, and in particular SNP data. The original application of these algorithms was to detect frequent patterns in purchased items [137]. ARM is based on the ‘apriori’ algorithm that identifies frequent ‘itemsets’ that are then used to generate rules. ARM is used in a variety of field including the automated detection of unusual soil moisture probe response patterns [138], and darknet big data [137]. There are also adaptations such as the frequent pattern growth algorithm using a prefix tree that recursively eliminates branches of itemsets to store minimum supports that generate the association rules [137]. Further to this, ARM has previously been used in a genomic capacity to complement and modify the GWAS process [139]. While ARM presents a method to perform an exhaustive search for interactions in genomic data, computational expense and the fundamental practice of unsupervised learning can create issues in both the analysis and result interpretation.

Software program LAMPlink [140] derives its name from methods employed, 'Limitless Arity Multiple-testing Procedure'. This developed algorithm provides a method of detecting significant associations using a limitless number of features. The benefit of this method is the vast reduction in computational complexities by using ‘itemset mining’ to remove redundant SNP combinations before an exhaustive search is performed. For further information on the techniques in LAMPlink, refer to section 5.5.1 . The limitations of LAMPlink exist in the requirements of using limited genetic models, dominant and recessive. This contrasts with the benefits provided by the MDR technique but is superior in addressing the prominent issue of computational complexities.

As previously mentioned, and utilised in the above software, ‘itemset mining’ is a limitless arity option that falls within the area of association rule learning for data mining. Previously designed for market basket analysis, its uses have extended into many fields, including genetics [141]. ‘Itemset mining’ approaches the problem using an unsupervised machine learning which benefits from a hypothesis-free and assumption-free method. Additionally, it also provides a method in which all combinations of SNPs can be investigated, but as previously mentioned, this is at the cost of computational complexity and further to this, unsupervised learning methods are used to analyse and extract patterns of information from the data; this results in all patterns being investigated which may not always be associated to the phenotype.

3.2.3 Correction Methods

Due to the nature of genomic data, with large datasets with variability in the representation, there are problems that arise when analysing information such as this. Within this section, some of the most prominent issues in association analysis for genomics are discussed and correction methods that are produced to modify or adapt the analysis results to better represent the true nature of the results. This section outlined methods and techniques for the exploration of RO3 and RO1.

3.2.3.1. Multiple Testing Problem

Univariate association analysis considers a feature, X, against a response, Y. This analysis provides a value of measure to indicate the probability that a feature, X, deviates from the null hypothesis. The null hypothesis in the case of given example is ‘SNP A’ is not significant to the occurrence of breast cancer in the given cohort’; therefore anything that deviates from a p-value of 1 (less than 1) implies that there is an increased occurrence of ‘SNP A’ in response to subjects status, case and control. During common data analysis investigations, a p-value of 0.05 (5%) suggests that feature, X, is very significant to response, Y; however, when investigating genomic data, a more conservative threshold is considered due to the vast number of features that are tested.

In standard GWAS, the number of SNP features considered are commonly in the range of 300,000 to 500,000. In order to put this into context, if the number of features to be tested is 8 the following calculation provides an approximate estimation for the probability of observing at least one significant result (S) by chance:

$$\begin{aligned} &P(S \geq 1 | N) = 1 - P(S = 0) \\ &P(S \geq 1 | 8) = 1 - (1 - 0.05)^8 \\ &P(S \geq 1 | 8) \sim 0.34 \end{aligned} \qquad \text{EQ. 3-13}$$

So, even with only 8 features there is a 34% chance of observing a least one significant result, S. Using the common 300,000 features that are consistently used or exceeded during genomic studies, an estimated probability of 1 signifies that at least one false positive result will present itself. This is known as the multiple testing problem as the more features that are included in the test, the greater the probability of observing at least one significant result and with this, are likely to observe what is referred to as 'type 1' and 'type 2' errors.

Family Wise Error Rate (FWER) and False Discovery Rate (FDR)

When analysing data, results consist of a set of p-values that are calculated probabilities based on a null hypothesis. These p-value results will provide a value between 0 and 1 that denote the significance of the feature; values closest to 0 are considered very significant results with the most common threshold being applied at 0.05, in other words, a 5% chance of false positive result.

'FDR is the rate that significant features are truly null'[1]

The FDR is statistical measure that aims to correct or reduce the effects of Type 1 and Type 2 errors [142]. FWER refers to the probability of making at least one Type 1 error [143]. Type 1 errors occur when false positive results are present and type 2 errors occur when there are false negative results present; in other words, results with no significance are deemed significant and results with significance are not considered when they should be. Table 3.6 demonstrates the variables considered in FWER and FDR. While FWER refers to the control of *V* where a test that is truly null has been labelled significant, FDR refers to the control of *V* and *T*, where V remains the same and *T* refers to tests that reject the null hypothesis but are labelled non-significant.

TABLE 3.6: TYPE 1 AND TYPE 2 ERRORS DEMONSTRATION

	Null True (H_0)	**Alternative True (H_1)**	**Total**
Declared Significant	*V*	*S*	*V+S*
Declared Non-significant	*U*	*T*	*U+T*
Total	*V + U*	*S+T*	*N*

p-values produce results based on a false positive rate; however, it has been heavily discredited for its lenient measurements which lead to a large number of type 1 errors. There have been numerous techniques which attempt to correct the issue by making adjustments to the resulting p-values of a given statistical test such as Bonferroni Correction [144], Local False Discovery Rate [145] and q-value [51].

BONFERRONI CORRECTION

Bonferroni Correction is a technique that adjusts the significance threshold in response to the number of features in the test [19]. Therefore,

$$f(x) = (1-\alpha)^{\frac{1}{n}}$$
$$where; \alpha = \frac{\sigma}{n} \qquad \text{EQ. 3-14}$$

However, this adjustment technique is very conservative; while it corrects for the number of type 2 errors, it is at a disadvantage to type 1 errors that increase dramatically depending on the number of features [146]. While we may be ensuring a reduction in the number of false positive results, type 2 errors have also disregarded in X amount of results that contain genuine results and effects for the given hypothesis. The Bonferroni method can also extended the Bonferroni inequality, or Boole's law which suggests that the probability of at one event happening is either equal to or less than the sum of the probability of the individual events [147].

Q VALUE

The q-values [14] consider the level of uniform distribution in the given set of results to correct and adjustment the probability measurements based on threshold, t.

$$FDR(t) = \frac{area\,under\,\hat{\pi}_0}{total\,area} \qquad \text{EQ. 3-15}$$
$$\hat{\pi}_0 = proportion\,of\;\; features\,that\,are\,truly\,null$$

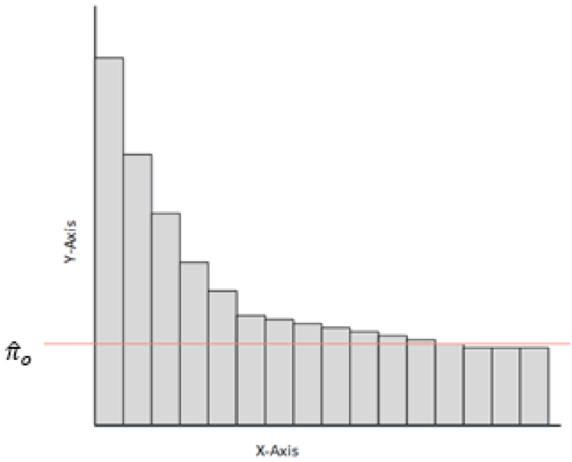


FIGURE 3-7: REPRESENTATION OF THE Q-VALUE THRESHOLD

Threshold $\hat{\pi}_o$ is demonstrated by the red line in Figure 3-7. This provides a more realistic measurement for FDR and as such balancing the potential of type 1 and type 2 errors. In comparison, performing 1000 tests using a threshold of 0.05, p-values would yield a likely result of approx. 100 false positives. In contrast, q-values would only yield approx. 5% false positive from the significant results; so, if 100 significant results were found, the number of likely false positive is reduced to 5. With q-values, each estimated q-value is either greater than or equal to its actual value; this approach is more conservative (but not overly so) which is desirable when choosing features for further analysis [14][148].

3.2.3.2. POPULATION STRATIFICATION

Genomic control (GC) is an adjustment method for population stratification that controls for the presence of population structure. Population stratification is used to control for the presence of systematic difference that exist in a population based on the allele frequencies that suggest there exists sub-populations within the cohort [115]. As previously mentioned, the majority of association models used within genomic studies use the chi-squared statistical measure to produce a p-value measure. GC utilises the chi-squared results (X^2) and assumes that a constant inflation factor (λ) is in effect across the population, therefore each results is adjusted using [15];

$$GC = \frac{X^2}{\lambda}$$

$$\lambda = \frac{median(X_1^2, X_2^2, ..., X_L^2,)}{0.456}$$

EQ. 3-16

eq. 3-16 [115] produces the genomic inflation factor, where L represents the number of null results. Using the median of L, null X^2 feature results to adjust by the expected median that when based on 1df is 0.456. This inflation is then used as a constant to adjust the total results.

One of the visual tools for quality control that is commonly used is Quantile-Quantile plot, or more commonly referred to as, Q-Q plot. This visual tool plots the observed values of p against the expected values (null hypothesis) where it is expected that no deviation exists toward the left-side of the plot, while minor deviation exists at the right-side. Genomic control is most notable in its adjustment difference when visualised using Q-Q plots, as demonstrated below;

As can be seen from Figure 3-8, GC adjusts the results so that the deviation from the null hypothesis is corrected.

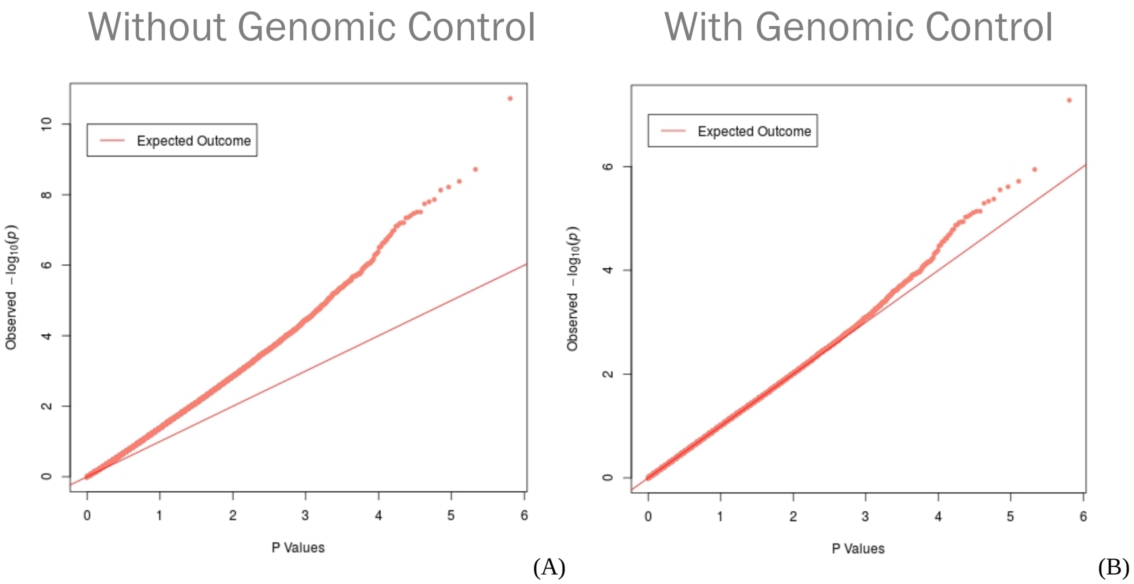


FIGURE 3-8: DEMONSTRATION OF THE EFFECT OF NO GENOMIC CONTROL (A) AS OPPOSED TO USING THE CORRECTION METHOD GENOMIC CONTROL (B).

As the demand for genetic material is becoming more prevalent in the field of genomics, with larger cohort sizes required for analysis to improve the statistical power of results, this can, and normally does, result in anomaly cohort clusters. There are as a result of differences due a number of factors, primarily, environmental exposure that can include lifestyle choices, urban or rural habitation, weather exposure, etc. This is due to the increasing distance required to obtain subjects for studies across multiple sites, states and even countries particularly concerning rare diseases and even complex diseases which require larger cohorts to compensate for the vast phenotypic differences that can be present. Reliance on control for ancestry alone will not control for sub-population occurrence therefore GC performs the necessary adjustments to control for this.

3.2.4 Association Analysis Summary

Association analysis is used to determine a p-value estimate for each feature. However, due to the structure of SNPs, the interpretation of these values can be evaluated using a variety of models as outlined in section 3.2.1 .

Apparent in Figure 3-9 is the lack of diversity between the allelic and logistic models, the -log$_{10}$ p-values presented in these plots vary slightly however the overall structure of the data remains similar among the models. Logistic regression provides a less optimistic results but does not have the added advantages of increased statistical power, however, when approaching genetic data as a singular process, this method provides the most realistic results.

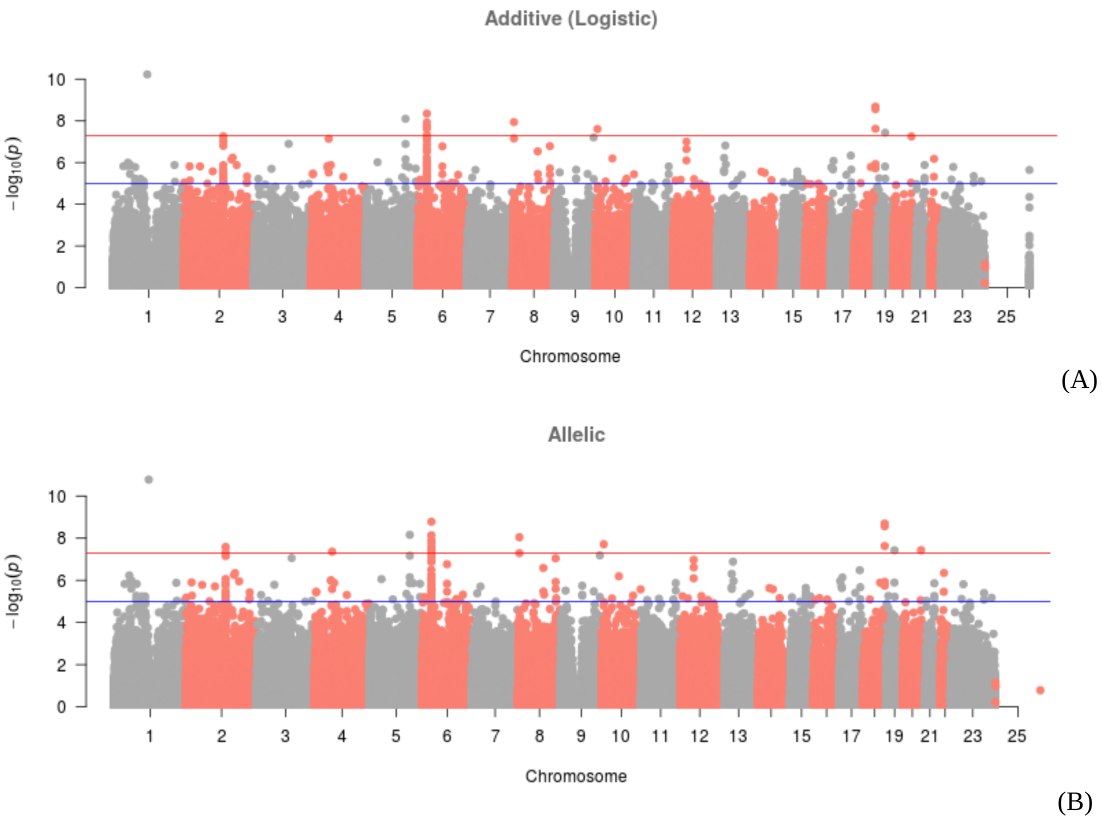


FIGURE 3-9: COMPARISON OF GENETIC MODELS (A) ADDITIVE (LOGISTIC) (B) ALLELIC

DOMINANT, RECESSIVE AND GENOTYPIC METHOD DEMONSTRATED IN Figure 3-10 present similar results to allelic and logistic with the majority of the structure remaining the same, but differences in the p-values show the effect the model choice can make.

Although all these models produce an estimate based on a different approach, only two of these models can utilise genomic control. Genomic Control is standard practice in current genomic studies given the increasing cohort sizes that are available and demanded to produce reliable results that effectively represent the population. Therefore, there is an increased probability of sub-populations existing with the data; as previously discussed, these sub-populations can present themselves due to environmental factors that commonly differ between countries, states and classes of individuals. It is not always possible to obtain all data to control for these subpopulations, therefore genomic control provides a control estimate for the differences that exist in the data.

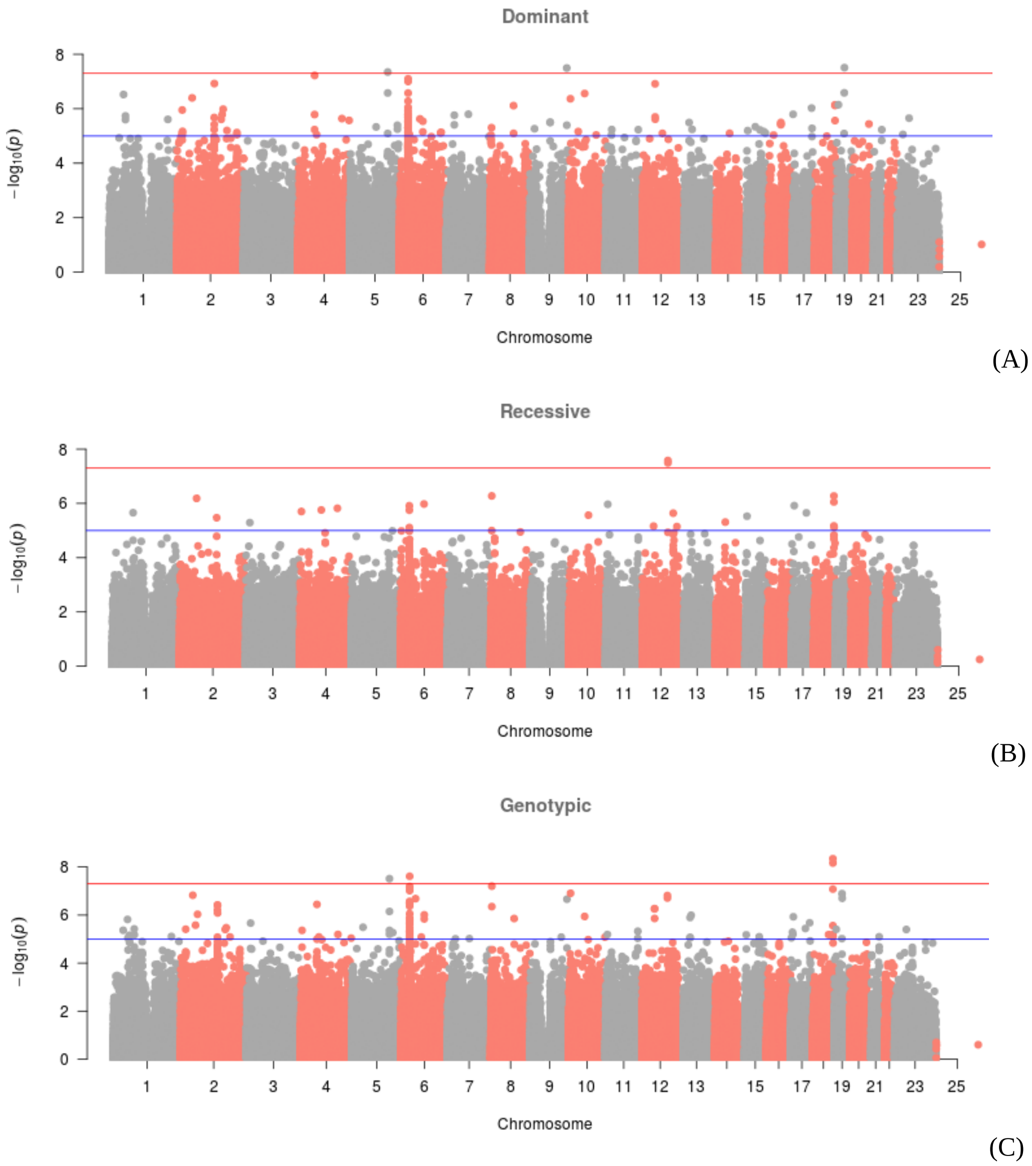


FIGURE 3-10: COMPARISON OF GENETIC MODELS (A) DOMINANT (B) RECESSIVE (C) GENOTYPIC

The adopted association model used within this methodology is allelic analysis as it increases statistical power over genotypic models [48], this is an important quality that will be beneficial to the feature extraction phase. The use of the allelic model, while beneficial, is at the expense of analysing the genotypic presence of genotypes, however, throughout the remainder of the methodology the genotype state is the primary format used to analyse the statistical significance of genetic features.

Within this research, a combination of univariate and multivariate approaches is used within stages for feature filtering and analysis. During this research, the use of multiple testing is explored, and further demonstration is provided in section 3.2.3 . Genomic control presents a particularly advantageous adjustment method that is an accepted standard in current studies in order to adjust for the issues that population structure can cause as demonstrated in Figure 3-8.

Multivariate analysis is a particularly important aspect of this research given that one of the most challenging aspect of an epistasis approach is the computational constraints and hardware demands. While there are several solutions to epistasis, the restraints prove particularly restrictive. To adhere to the requirements of the research, the multivariate analysis considers a limitation in hardware specifications with the need for a timely return, preferably using an exhaustive search. As previously outlined, the stages of the methodology utilise a combination of univariate and multivariate analysis in order to reduce the representative dataset to the most prominent SNPs. This leads to an improvement in time performance and increases the selection of software that can be considered, particularly concerning limitless-arity choices as this takes into consideration the possibility of interactions between SNP that extend beyond a pair-wise nature. To compare the multivariate limitless-arity methods, the limitations of the MDR method exist in the computation time expense and the continuous iterative process that it demands for more than 2 SNPs. The limitations of the ARM method exist in the unsupervised approach that takes no guidance from the classification labels; this could lead to outlined interactions that are unrelated to the phenotype. The limitations of the LAMPlink software exists in the number of SNPs that can be input as the computation time expense exponentially increases. Given the above methods and their limitations, the choice of multivariate analysis technique that would be best suited to the current research is LAMPlink as the limitations of the method are complemented by previous stages of using univariate analysis and further to this the implemented ‘itemset mining’, an ARM technique, systematically reduces the number of combinations to be tested, overall providing a faster option.

3.3 Inference Analysis Methods

The analysis of SNPs is an important process that not only aims to analyse and produce an estimated measure of associative significance but to also represent that information in an interpretable form that can be used to outline relationships and association that exist in real-world terms. Through this stage it is appreciated that there are a substantial number of techniques, models and approaches to apply to the current data type [149]–[151]. In order to scale the literature to within the scope of our project and in particular for the purposes required the following section targets analysis techniques that are used in context of computing the probability that a variant and/or combination is truly associated with the phenotype [1] and measure of association. Further to this, one of the most demanding requirements in the health field currently, with the rise in techniques and technology, is classification, particularly for the inclusion of decision support systems (DSS) [152]–[154]. This section outlined methods and techniques for the exploration of RO4 and RO5.

3.3.1 Statistical Techniques

Statistical techniques are used throughout genomics, therefore, to reduce the scope of this section, the statistical techniques that are considered for this are those that represent the relationship in terms of association or the measure of associations. To consider first statistical techniques that test for association, the considered techniques are frequently used within the broad term of data analysis with confirmed applications in genetic analysis. The second considers the relationship that exists and provides a measure for how associated one variable has to another.

3.3.1.1. Test of Association

Tests of association are commonly conducted using contingency tables, taking a variety of approaches in how to analyse the data. This then presents a choice of test depending on how the data is represented within each approach. The following tests are standard practice in data analysis and are also commonly applied in genomics.

Fisher’s Exact Test

The Fisher’s exact test is a commonly used tool to assess the statistical significance between variants in the capacity of GWAS [103], epistasis [30] and is also adopted as a permutation test in standard software for genetic analysis [102]. Fisher’s exact test identifies the exact difference between the null and alternative hypothesis [155]. Fisher’s exact test is recommended for sample sizes < 1000 while the technique introduced next is recommended for sample sizes > 1000, however, fisher’s exact test is applicable in both cases [155]. Refer to Table 3.8 in regard to Eq. 3.16.

$$Fishers\,Exact\,Test = \frac{[a+b]![c+d]![a+c]![b+d]!}{[a+b+c+d]!a!b!c!d!} \qquad \text{EQ. 3-17}$$

PEARSON'S CHI-SQUARED TEST OF INDEPENDENCE

The Chi-Squared test (X^2) is an approximation method [155]. This technique models and measures the departure from independence in a non-parametric test making no assumptions about normality or homogeneity of variance [156]. The X^2 test is a transparent and interpretable method that has previously been used in studies to assess heterogeneity [47] , QC [101], GWAS [51] and for statistical association [47]. These tests are noted to have considerable power [157] but they are limited given their inability to consider covariates and as such are susceptible to population stratification.

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} \qquad \text{EQ. 3-18}$$

The main difference that exists between the Fisher's exact test and X^2 test is due to the approaches. While fisher's exact test will give an exact difference between null and alternative hypothesis, the X^2 will provide an approximation. As only an approximation relies on the assumption that the probability of observed binomial frequencies can be approximated by the continuous X^2 distribution, which is not always correct and creates some error.

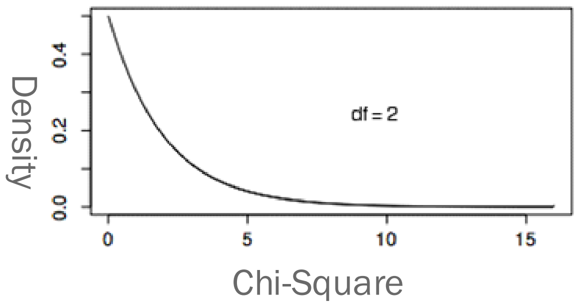


FIGURE 3-11: X^2 DISTRIBUTION (2DF)

Under circumstances that it is required, an adaptation of the X^2 test is the Yate's correction for continuity [158]. To prevent overestimation of statistical significance when analysis small data (cells of contingency table contain small values), Yates' suggested a -0.5 correction between the observed and expected values in the contingency table, therefore reducing the chi-squared value and increasing the p-value.

COCHRAN-ARMITAGE (CA) TREND TEST

An adaptation of the infamous X^2, CA introduces a robust directional adjustment that focuses on the suspected result [12] and can be used without the pre-processing of HWE [159].

$$T^2(x)=\frac{[n^{-1}\sum_{i=0}^{2}x_i(sr_i-rs_i)]^2}{\operatorname{var}[n^{-1}\sum_{i=0}^{2}(sr_i-rs_i)]}$$ EQ. 3-19

TABLE 3.7: CA EXAMPLE 3X2 CONTINGENCY TABLE

	NN	NM	MM	Total
Exposed	r_0 (p_{10})	r_1 (p_{11})	r_2 (p_{12})	r ($p_{1.}$)
Not Exposed	s_0 (p_{20})	s_1 (p_{21})	s_2 (p_{22})	s ($p_{2.}$)
Total	n_0 ($p_{.0}$)	n_1 ($p_{.1}$)	n_2 ($p_{.2}$)	n (1.0)

referring to Table 3.6, which can be derived as [160];

$$T_{CA}=\sqrt{\frac{n}{r(n-r)}}\frac{n(r_1+2r_2)-r(n_1+2n_2)}{\sqrt{n(n_1+4n_2)-(n_1+2n_2)^2}}$$ EQ. 3-20

with reference to a 2x2 contingency table as provided in Table 3.7.

The limitations of this method exist in the nature of the approach, as an adaptation that aimed to focus on the alternative hypothesis, the test can overlook associations that exist outside of the null and alternative hypothesis.

3.3.1.2. Relative Risk

Measures of association apply a value to an association which suggests the relationship that exists between them e.g. in the presence of genotype A there is 1:1.2 risk that subject A will have the disease. The following section explores the common measures of association in genomic studies referred to as the relative risk.

Odds Ratio (OR)

The odds ratio is a common method that is widely used across many if not all fields that require statistical analysis. It is also a method that is commonly used in genetics [47] to outline the measure of association for a variety of conclusions particularly in case-control studies to measure, for example, the risk presented for a genotype in a sample population present for a phenotypic trait [161].

$$OR=\frac{\text{odds of disease among exposed}}{\text{odds of disease among unexposed}}=\frac{p(1)/(1-p(1))}{p(0)/(1-p(0))}$$

EQ. 3-21

$$OR=\frac{a\times d}{b\times c}$$

eq. 3-21 shows the equation of the odds ratio in relation to a contingency table, for demonstration in Table 3.8.

TABLE 3.8: OR EXAMPLE 2X2 CONTINGENCY TABLE

	Case	Control
Exposed	a	b
Not Exposed	c	d

While the odds ratio is a simple interpretable test that is devoid of bias, there are still limitations as with any method. The prevalence ratio is not approximated which can result in misleading assumptions of the information [162] and information is hidden when analysing dichotomized continuous measure, leading the reduce statistical power and amplification of the measurement error [163].

RISK RATIO (RR)

Complementary to odds ratio, the risk ratio provides the multiple of risk of the outcome in one group compared to another[164]. This is in contrast to odds ratio that provides an estimate of the odds of disease between exposed and non-exposed individuals which can yield much higher values that the risk ratio e.g. in cases with an incidence rate of 20%, an OR of 10 corresponds to a RR of <4 [163].

$$RR = \frac{prevalence}{incidence} = \frac{a/a+b}{c/c+d} \qquad \text{EQ. 3-22}$$

In reference to case-control studies, the incidence rate is not measurable given that the study is split (normally evenly) into exposed and unexposed individuals, however the OR can be transformed into RR.

$$RR = \frac{OR}{1 - \frac{c}{c+d} + \frac{c}{c+d} * OR} \qquad \text{EQ. 3-23}$$

3.3.2 CLASSIFICATION TECHNIQUES

Classification techniques are advancing as more data becomes available and more focus is drawn to these methods. Machine learning is a technique that uses algorithms and statistical techniques to produce a model which can 'learn' information about a given set of data to provide us with the resulting correlations (or otherwise)[165]. ML can be used in a variety of capacities including for feature selection, extraction, classification and regression [166]. There are several methods of machine learning which can be applied to various types of data, however choosing the correct ML method is key to discovering valuable information from the given data. There are a number of desired outcome features that are normally used to select the most appropriate Machine Learning method; classification and regression provide a focus for machine learning methods for classifying data etc. In terms of machine learning methods, there are two main areas which are supervised and unsupervised learning.

3.3.2.1. Supervised Learning

Supervised Learning uses previous information, demonstration or what would primarily be training data to produce a basic framework which can be used in a number of circumstances: Classification and Regression [167]. This type of learning can be explained using a simple example:

A child is given a pile of sweets and is told to identify the different classes; either chocolate or gummy. The child will firstly take a selection of the sweets and eat them to find out whether or not they are chocolate or gummy. Based on their physical appearance and the previous knowledge of which appearance corresponds to which class of sweet they will then divide them between the two classes.

3.3.2.2. Unsupervised Learning

Unsupervised Learning does not use any previous information to produce a framework of information. In cases where unsupervised learning is used, the researcher will likely be looking for patterns of information in the data that are not lead by any given classification bias [168][169]. An example of unsupervised learning:

A museum would like to introduce events on different floors of the museum. They would however like to address their target audience. The museum has collected a list of attendants to the museum with a large amount of variable data e.g. age, sex, group attendance (e.g. family, couple), exhibits visited, exhibits enjoyed, clothing worn, hair colour, eye colour, gift shop purchases, etc. Using this data, they want to identify if there is a pattern among the visitations of the varying types of people.

Associations will be drawn from the data which could either lead to successful or unsuccessful decisions of events on the each of the floors. The data could produce an association that families that are attending are more likely to both visit and enjoy the insect and bug section of the museum which could indicate an event associated with this exhibit advertised towards children could produce increased attendance. However, an association could also be made between the dinosaur exhibit and people with blue eyes; a dinosaur event advertised towards people with blue is less likely to be a success. This method produces results free of any bias; however, they can also result in a lot of false positive results. Unsupervised learning can be detrimental to the findings of genetic association studies as with such a large dataset with a large set of features there are bound to be false positive associations frequently[170]. The following methods use different techniques to apply ML analysis which include feed-forward, back-propagation, cost/loss functions, risk minimisation, feature selection etc. These methods will address classification of case/ control subject data to produce a best accuracy estimation based on the given features in the data.

3.3.2.3. DECISION TREES

Tree methods can be applied to both classification and regression-based problems. Most if not all decision tree models derive from logic and statistics and generally provide interpretable models for inductive reasoning using supervised learning.

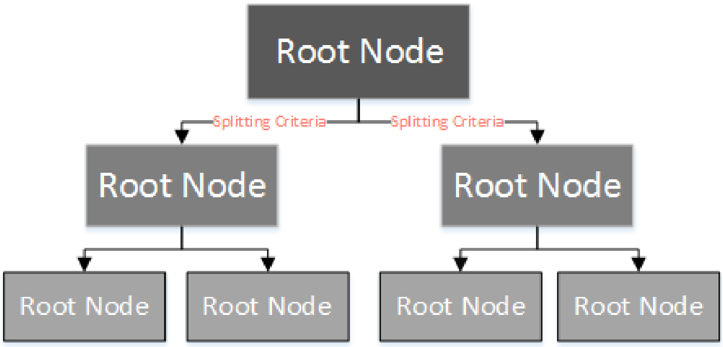


FIGURE 3-12: REPRESENTATION OF DECISION TREES

The Random Forests (RF) algorithm [171], or Random Decision Forest, works as a large collection of de-correlated trees. It is an ensemble approach and a type of 'bagging' technique. An ensemble approach combines a group of 'weak learners' to produce a 'strong learner'; the effect of this aims to improve performance by reducing noise in the set. Bagging techniques normally have low variance, and therefore, as previously noted, are less prone to influence of 'noise'. The RF method starts with the initial tree, S, which is the randomly split into several subsets of the tree, $S_1, S_2 \cdots S_n$, including feature $A_n \cdots n_n$, depending on the number of features in the given set, as demonstrated in (10). From here the RF classifier uses the results from each of the subset trees, acting a 'voting system' to give an overall predictive estimate of the results.

Chi-Squared Automatic Interaction Detection (CHAID) is one of the most popular statistical decision tree methods. It is a multivariable algorithm, based on Pearson's Chi-square statistic that identifies the strongest interaction association to the dependent variable (Phenotype). The dependent variable of a CHAID tree is required to be categorical but the independent variable can be categorical or metric data types [172]. The rules are generated using X^2 test for categorical variables and F-test statistics for continuous variables and Bonferroni is used to adjust the p-value. CHAID is suited for large datasets and commonly used in marketing.

Classification and Regression Trees (CART) is a technique very similar to CHAID with the main difference being that CART use binary splits, while CHAID uses multiway splits. While there has been some discussion around the area, authors of the CART method argue that the binary split are preferred as they improve predictive performance [173]. This predictive performance however may be at the cost of ease in interpretability as trees can grow large with predictors used many times.

3.3.2.4. Neural Networks

Neural Networks encompass a collection of algorithms that base their systems on biologically inspired networks, in particular, the brain. It is composed of a large network of interconnecting edges that work together to produce a response from input variables. This can be used for classification or regression.

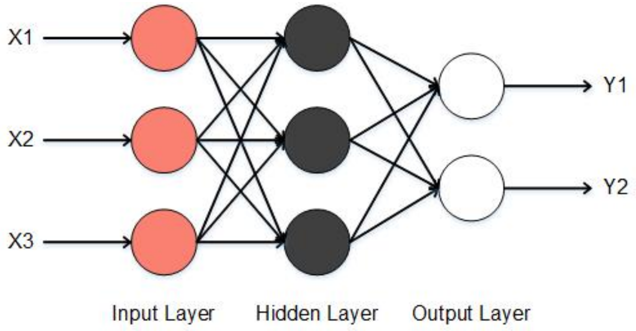


Figure 3-13: Representation of a Neural Network

Multi-layer Perceptron (MLPs) is a type of Neural Network which is based on the biological axons of the brain. If we consider a single Perceptron, we can use this to model linearly separable problems; this occurs when data can be separated by one single threshold or a line if we consider a graph. However, a lot of problems are not linearly separable, as is the case with the current study; in these cases, Multi-layered Perceptron model can be used which builds numerous Perceptrons to classify and interpret inputted data that could not be processed by a single neuro.

Stacked Autoencoder (SAE) is a deep learning method that builds upon the standard neural network framework to produce a larger system to analyse complex data structures. SAE builds itself using a series of Autoencoders that are simplified versions of an MLP, one hidden layer and one output layer, that have an equal number of output layers to input layers. Each hidden layer of the autoencoder, ‘encodes’ the information and forms the input layer of the consecutive autoencoder to build the stacked autoencoders. The purpose of this model is to ‘recreate’ the information from the input layer in the output layer using compressed values as the hidden layer will force this when smaller [174]. While stacked autoencoders have predominantly focused efforts in the area of market basket analysis, its uses have recently been extended to pattern recognition for epistasis [175].

3.3.2.5. Support Vector Machines (SVM)

SVM's are a firm favourite in the machine learning community and while this category stems from one main algorithm, the adaptations that have been developed categorise and extenuate its uses beyond linearly separable data structures.

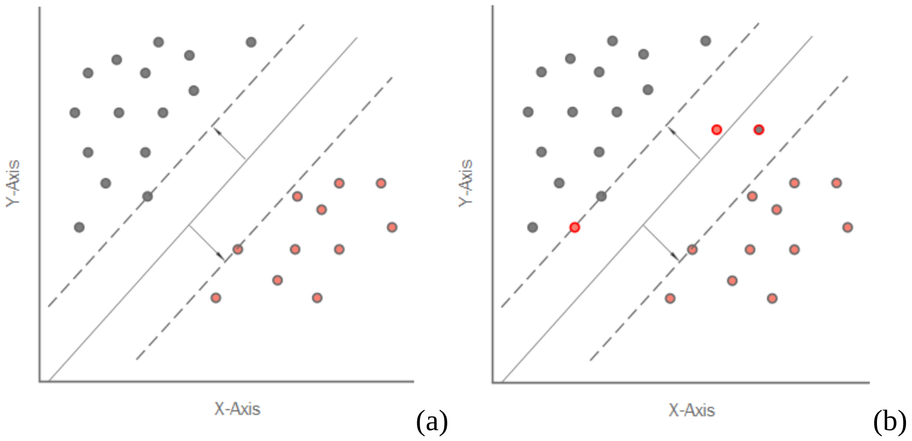


Figure 3-14: Visualisation of Linear SVM (a) Hard Margin (b) Soft Margin

Linear SVM's consider data structures with a linear approach. The aim to find the 'maximum margin hyperplane', that is using two hyperplanes to bound the categories or clustered regions with the largest distance between them, the maximum margin hyperplane then lie halfway between them as demonstrated in Figure 3-14. If the input data is linearly separable, a 'hard margin' will be used that supports the hyperplanes based on the most outer coordinates which then determine the margin hyperplane that differentiates the two. If the input is not linearly separable due to outliers, then Linear SVM's can still be used, the hyperplanes will allow for some noise, but this will come at a cost for misclassification.

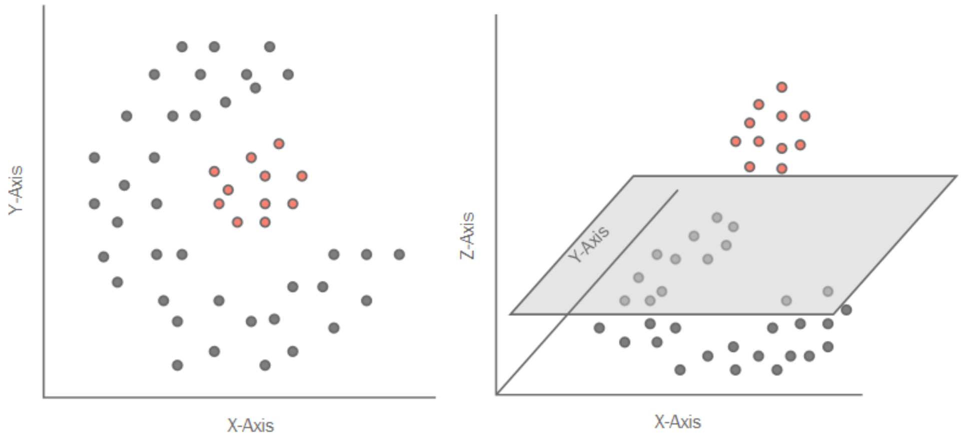


FIGURE 3-15: VISUALISATION OF NON-LINEAR SVM

Non-linear SVM's considers data structures that not linearly separable, in this instant they will use a 'kernel trick' that will use a small dimensional space and transform it into a large to infinite dimensional space using a function called a 'kernel'. There are a few kernels that can be used including linear, polynomial linear, radial based function that use varying functions to instantiate the process.

SVM's are strong candidates in many complex problems concerning data classification and regression due to their flexibility towards a variety of data structures. Further to this, it holds a particular advantage when the researcher is unaware of the data structure. However, SVM's are mathematically complex and computationally expensive [176], which is a limitation in small dimensional data.

3.3.3 INFERENCE METHODS SUMMARY

Inference methods are split into two categories, the statistical and classification approaches. To assess the inference methods, there are number of criteria that need to be fulfilled; the test of association, measure of association and classification. The techniques outlined for test of association are standard techniques that are widely used in not only the genomics field but overall in data analysis; the limitations of these methods present a minor challenge that by using multiple, verification is introduced with a variety of approaches. The choice of methods is chi-squared and fisher's exact test, both of these methods are standard, and it is expected that they should reach similar conclusions to significance of association, under circumstances where they are not, further investigation can be flagged. Additionally, the use of Yate's continuity correction will be adopted for tests of association that require it combat the issues that can be apparent in Chi-Squared. The exclusion of CA trend test is due to the specific nature of focusing on the alternative hypothesis, the benefits of this technique can be recreated in a similar fashion with the chosen techniques.

Measure of association provides a value to indicate the relevance of the relationship. Outlined are two methods that represent this information however only one method can be use under the study design. The OR is a standard method that is again widely applied across all areas of data analysis, however its disadvantages in this field are due to misinterpretation. The RR technique addresses this but cannot be directly calculated from case-control data but can be calculated from the OR, therefore in order to provide a measure that better represents the information, the RR is also adopted.

Classification of data is a wide field that is continually receiving updates, adaptations and new methods and techniques to perform the analysis. Each method uses its own approach to data; previously discussed are some of the most prominent categories of machine learning. Previously outlined was the critical factor of interpretability, previous methods demonstrate this factor, but it becomes a defining factor in classification. Neural network and SVM methods are strong candidates in classification due to their successes when being applied to many different data types. While the successes of these methods are well documented, the interpretability is lacking, the advantages of interpretation lie in the ability to outline areas of adjustment, correction and interest that cannot be identified through the output of a model. Therefore, the most interpretable approach is decision trees, such as CHAID and CART. As the most interpretable methods outlined, the differences between these two define their approach. CHAID method uses multi-split chi-squared technique while CART uses binary splits that can obstruct the interpretation by outputting large trees that can be difficult to navigate and outlined areas of interest.

The methods and techniques outlined in this section encompass the 'belly' of the methodology and are chosen in response to the study design and a crucial factor of interpretability. The next section summarises the findings of this chapter.

3.4 DISCUSSION

Within this chapter, various methods and techniques have been discussed that service and analyse the dataset to improve and gain insight from the information available. Chosen quality control procedures are chosen in response to the study design and perform a necessary service that improve the representation of the dataset by removing bias and erroneous data.

Univariate association analysis again introduces a variety of approaches to genetic data that concern the transformation and processing of the data taking into consideration the specific relevance of genetic information in the form of allelic and genotypic formats. An allelic model has been outlined as the model of choice given the increase in statistical power. Multivariate analysis outlined the choice of software LAMPlink that complements the study design and also incorporates techniques that address the main concern of multivariate analysis in the context of genomics, computational complexity.

While each method has its advantages, there are always limitations. Previously discussed are the limitations of GWAS due to the large feature set used in analysis; this opens up the method to issue of False Discovery Rate which incorporates False Positives and Negatives. In order to combat this, there are a number of techniques that can be used. A specialised technique for genomics is genomic control, this specifically controls for population structure in the dataset which is addressed during the association analysis to reduce inflated values as a result of existing population structure. Further to this, to address value inflation due to large feature sets, a common approach is to use multiple testing techniques.

Within this research, one of the issues that is prevalent is the use multiple testing; this is applicable once the association analysis has been performed to adjust the values to provide a realistic representation of the information. While many techniques have been introduced that range from lenient to conservative, a concern of this thesis is the adjustment of these values after they have already inflated. The concept of probability suggests that out of the features analysed there is likely to be at least 1 false positive, and in the size of data used in genomics, this is likely to be far more. Therefore, adjusting the values will control for inflation of features that represent their true significance to the phenotype, while features that are inflated due to a biased representation of the data will retain significance. Therefore, the problem needs to be addressed before the association analysis is conducted to reduce the presence of these biased features. This is discussed further in section 3.2.3 .

Inference methods are used to provide an evaluation of the relationships and significant of features within the given sample population. Previously discussed is the importance of interpretability which has influenced the choice of techniques and methods in this section. Multiple methods are chosen for test of association and measure of association in order to validate, verify and improve representation of the information. Chi-Squared and Fishers' exact approach data analysis from different perspectives, this can lead to a small margin of error that can skew the results and therefore by using both, they can verify the result of the other. Further to this, the choice of OR and RR will lead to better interpretation of the resulting information. Defined previously are classification methods, the use of these methods is aimed at statistical feature selection by outlining best performing combinations of features. As a result, the most qualified methods that also adheres to the interpretability requirement are decision trees, in particular, the CHAID analysis, explained in section 5.6 .

These outlined methods are the working engine of the proposed methodology and as such are important considerations to a successful application. The complexity of genomic data and the transformations required to better represent the data based on the techniques being used can lead to some confusion. To better understand and describe this, the next chapter will discuss the representation of the data and to describe the dataset used to evaluate the method.

Chapter 4: DATA DESCRIPTION

Due to the complexity of genetic data, there are many representations and transformations that can be used to approach the data based on different hypothesis e.g. the presence of a recessive allele will increase the risk of the variant. Further to this, in order to represent the varying approaches, the data can be transformed in a variety of ways; this is further explained during this chapter. This chapter is provided for use as a reference going forward to the next chapters. This chapter contains descriptions and examples of the data based on genetic form and representation during analysis. This includes the genetic representation, binary transformation and the format of the PLINK input data that are used to perform the association analysis. Additionally, during this chapter the DRIVE dataset for breast cancer subjects is outlined and described including characteristics of the data and the available information.

4.1 GAME-ON: DRIVE: DATA DESCRIPTION

Subject genotypes were attained from repository platform, Database of Genotypes and Phenotypes (DBGaP). Data was collected under the Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative that funded 5 projects, one of which was the Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) [11] project that focused its efforts in breast cancer for the systematic discovery and replication of additional common genetic variants. These variants were assessed for their biological significance and from this, developed evidence-based assessments of the clinical validity of prediction algorithms in practice.

Genotyping was conducted by the Center for Inherited Disease Research (CIDR), Centre for Cancer Genetic Epidemiology, University of Cambridge, and the National Cancer Institute. The following studies contributed germline DNA from breast cancer cases and controls:

Study	**Inc.**	**Ref**
Breast Oncology Galicia Network **(BREOGAN)**		[177]
Cancer Prevention Study 2 **(CPSII)**	X	[178]
Copenhagen General Population Study **(CGPS)**	X	[179]
Melbourne Collaborative Cohort Study **(MCCS)**	X	[180]
Multiethnic Cohort **(MEC)**	X	[181]
Nashville Breast Health Study **(NBHS)**		[182]
Nurses Health Study **(NHS)**	X	[183]
Nurses Health Study 2 **(NHS2)**	X	[183]
Polish Breast Cancer Study **(PBCS)**		[184]
Prostate Lung Colorectal and Ovarian Cancer Screening Trial **(PLCO)**		[185]
Studies of Epidemiology and Risk Factors in Cancer Heredity **(SEARCH)**		[186]

Swedish Mammographic Cohort **(SMC)**		[187]
The European Prospective Investigation into Cancer and Nutrition **(EPIC)**		[188]
The Sister Study **(SISTER)**		[189]
The Two Sister Study **(2SISTER)**		[189]
Women's Health Initiative **(WHI)**	X	[190]
Women of African Ancestry Breast Cancer Study **(WAABCS)**		--

After quality control was conducted, genotype information remained from 7 combined studies, more information is provided below for each:

Cancer Prevention Study 2 (CPSII): The Cancer Prevention Study II (CPS-II) is a prospective mortality study of approximately 1.2 million American men and women [178].

Copenhagen General Population Study (CGPS): Initiated in 2003, participants from Copenhagen were randomly invited using the Danish Civil Registration System. Participants completed a questionnaire, a physical examination and provide blood samples for DNA extraction. Statistical analyses of cancer risk were determined from the Danish Civil Registration System. All participants gave written informed consent. The study was approved by the Herlev and Gentofte Hospital and by a Danish ethical committee (H-KF-01-144/01) and was conducted according to the Declaration of Helsinki. Since 1987, is has been compulsory for all physicians by law to register cancer diagnoses in Denmark in the national Danish Cancer Registry which records approximately 98% of all cancers in Denmark [190]. Diagnoses of invasive cancer were made using the seventh or tenth editions of the WHO International Classification of Diseases [191].

Melbourne Collaborative Cohort Study (MCCS): A cohort of 41,500 subjects were collected between 1990 and 1994, aged 40-69. Cases of cancer were identified by matching to cancer registries and death indices [192].

Multiethnic Cohort (MEC): The Multiethnic Cohort Study of Diet and Cancer (MEC) was funded by the National Cancer Institute (NCI) in 1993. The cohort is comprised of more than 215,000 men and women primarily of African American, Japanese, Latino, Native Hawaiian and Caucasian origin. The study focused on examining lifestyle risk factors, including genetic susceptibility in relation to causation of cancer. Each member completed a 26-page questionnaire as well as biological specimens (blood or urine). Cancer status was confirmed using cancer registries established by state statute in Hawai`i and California [181].

Nurses Health Study (NHS): The Nurses' Health Study (NHS) was established in 1976. The focus of this study was to investigate the potential long-term consequences of oral contraceptives.

Nurses were used for subjects given their knowledge about public health assuming their ability to provide complete and accurate information regarding diseases. The study was carried out across the following states of USA: California, Connecticut, Florida, Maryland, Massachusetts, Michigan, New Jersey, New York, Ohio, Pennsylvania, and Texas. Biological specimens consisted of blood and urine samples. [183]

Nurses Health Study 2 (NHS2): Established in 1989, a continuation from the previous study to included younger nurses that has started using oral contraceptives during adolescence (a maximal exposure during early reproductive life). The study was carried out in the following states: California, Connecticut, Indiana, Iowa, Kentucky, Massachusetts, Michigan, Missouri, New York, North Carolina, Ohio, Pennsylvania, South Carolina, and Texas. Biological specimens consisted of blood and urine samples. [183]

Women's Health Initiative (WHI): The WHI entered postmenopausal women into four clinical trials (n = 68 132) and an observational study (n = 93 176) at 40 US clinical centers. Participants were aged between 50 and 79 years and had anticipated 3-year survival. Exclusion criteria of prior hysterectomy or breast cancer. Participants were required to have a mammogram not suspicious for breast cancer less than 2 years before entry was required and should also be taking oestrogen plus progesterone or no hormone therapy [190].

The genotype sequencing chip used to sequence the data for the subjects from the above studies was the OncoArray Illumina Chip that was designed with a GWAS backbone of ~250K SNPs for the Illumina HumanCore, with a total of ~533K. The remaining SNPs outside of the GWAS backbone were focused on variant findings and pathways indicated for the 5 most prominent cancers; breast, ovarian, prostate, colon and lung [193].

Table 4.3 provides demographic information regarding the country in which the study was conducted. Further information regarding demographic information in regards to the final set of participants used and clinical information pertaining to the dataset acquired is available in Table 4.1 and Table 4.2.

TABLE 4.1: CLINICAL VARIABLES AVAILABLE WITH THE DATASET

Variable	Variable Description
Subject ID	Unique identification number for each individual
Study Name	Acronym identifier for study
Study Country	Country of the study
Status	Case-Control category
Sex	Gender of Participant
Age	Age at Interview
Age of Diagnosis	Age at Diagnosis
ER Status	Estrogen Receptor status of tumour
ER Status Source	Source of ER Status
Sex by Genotype	Identified sex by genotype

TABLE 4.2: ADDITIONAL EXCLUSION CRITERIA

Exclusion Criteria	Reason
Individuals with AA ancestry	To reduce population structure bias
Individuals aged <40	To remove early onset breast cancer
Individuals without category invasive breast cancer	To include only individuals with confirmed breast cancer status

TABLE 4.3: DEMOGRAPHIC FEATURES BY STUDY

	Australia	Cameroon	Denmark	Nigeria	Uganda	USA
CGPS	-	-	2140	-	-	-
CPSII	-	-	-	-	-	6103
MCCS	1693	-	-	-	-	-
MEC	-	-	-	-	-	1169
NHS	-	-	-	-	-	3404
NHS2	-	-	-	-	-	3525
WAABCS	-	125	-	442	62	-
WHI	-	-	-	-	-	9618

TABLE 4.4: OESTROGEN RECEPTOR (ER) STATUS BY STUDY (EXCLUDING CONTROLS)

	Negative ER	Positive ER
CGPS	172	1029
CPSII	102	2044
MCCS	189	662
MEC	24	483
NHS	205	935
NHS2	225	1063
WAABCS	63	21
WHI	670	3918

With Breast Cancer incidence rates primarily effecting the female population, the study cohort is made up entirely of female participants (n = 28,281). Of these participants 14,435 subjects were cases and 13,846 were controls. Age ranged from 20 to 98 (μ = 63) based on a sample of 27,585 with age ranging from 20 to 92 (μ = 65). Estrogen Receptor Status has not been utilised in this study (See 2.3.1.4. for more information on Oestrogen Receptors). Cases were split into 3 histology types, invasive (12,412), in-situ (1,506) and unknown (517) of which, individuals of interest in this research are invasive histology type. Invasive breast cancer status regards cancer cells that have at least 'spread' to the surrounding breast tissue.

4.2 DATA REPRESENTATION

The following section outlines the representation of genetic data and the binary transformations used having converted information into a format for statistical manipulation. Firstly, consider the structure of bi-allelic SNPs that are used within this study; these SNPs can either be represented as alleles, their primary format, or genotypes, the combined state of the alleles, as presented in Table 4.5 and Table 4.6.

TABLE 4.5: ALLELE REPRESENTATION

SNP X	
Allele 1	Allele 2

TABLE 4.6: GENOTYPE REPRESENTATION

SNP X		
AA	Aa	aa

Further to this, SNPs also contain dominant and recessive alleles, or can also be commonly referred to as major and minor alleles, respectively, as represented in Table 4.7. This describes the genotypic expression found within the focus population; therefore, a dominant or major allele will describe an allele that is present in the majority of the population. However, it should be noted that while the dominant allele will commonly be expressed in the majority of the population, the main purpose of this term is to describe the effect that the allele has on the phenotypic expression. A dominant allele will generally mask the contribution of the recessive allele.

TABLE 4.7: DOMINANT AND RECESSIVE ALLELE REPRESENTATION

SNP X			
Allele 1		Allele 2	
Dominant	Recessive	Dominant	Recessive
A	a	A	a

Combined these alleles form one of three genotype states as demonstrated in Table 4.6, that can also be described in terms of a noted state as represented in Table 4.8

- When both alleles are dominant (AA), this is referred to as dominant homozygous,
- When both alleles are recessive (aa), this is referred to as homozygous recessive.
- In the case, allele 1 is dominant and allele 2 is recessive (Aa), this is referred to as heterozygous

TABLE 4.8: HOMOZYGOUS AND HETEROZYGOUS GENOTYPE REPRESENTATION

SNP X		
Dominant Homozygous	Heterozygous	Recessive Homozygous
AA	Aa	aa

These states and terms are used throughout the remainder of the methodology.

4.3 BINARY TRANSFORMATION

When transforming data to a binary or ordinal state, we take the approach that a recessive allele presents as the risk allele and therefore each additional recessive is additional risk. The following tables show the numerical representation of genotypic states in regard to the varying models that are commonly adopted in genetic studies, and further to this, are used throughout the proposed methodology.

Table 4.9 assumes that in the presence of 1 recessive allele, the risk increase 1-fold as is represented in a Heterozygous state (Aa). Further to this, a genotype state with a presence of 2 recessive allele, Homozygous Recessive state, the risk increases 2-fold.

TABLE 4.9: ADDITIVE MODEL CODE REPRESENTATION

AA	Aa	aa
0	1	2

Table 4.10 presents the assumption that any presence of a recessive allele will increase the risk 1-fold which also results in Heterozygous and Homozygous Recessives states being combined when conducting analysis such as permutation tests.

TABLE 4.10: DOMINANT MODEL CODE REPRESENTATION

AA	Aa	aa
0	1	1

Table 4.11 presents the assumption that only state, Homozygous Recessive which contains 2 recessive alleles, results in a 1-fold risk. Similar to the dominant model, the recessive model will normally combine Homozygous Dominant and Heterozygous states in analysis such as permutation tests.

TABLE 4.11: RECESSIVE MODEL CODE REPRESENTATION

AA	Aa	aa
0	0	1

Lastly, Table 4.12 presents the allelic model that considers the genetic information in an allelic state rather than genotypic as can be observed from the difference in representation between Table 4.5 and Table 4.6. For this model, the assumption presents a binary transformation between dominant and recessive allele states, similar to previous models, the recessive allele will be regarded as the risk allele and therefore will be assigned as 1.

TABLE 4.12: ALLELIC MODEL CODE REPRESENTATION

Allele 1		Allele 2	
A	a	A	a
0	1	0	1

The above models do not encompass all of the utilised genetic models in the literature but do represent the models that are adopted throughout the methodology. For further information regarding other genetic models, see 3.2.1

4.4 PLINK DATA FORMAT

The following section provides representation of the PLINK binary files that are input into PLINK to perform statistical analysis. There are 3 files associated with a cohort set: .bim, .bed, .fam. The .fam file provides information about subject characteristics and identity codes as presented in Table 4.14 and Table 4.13.

TABLE 4.13: FORMAT DESCRIPTION OF .FAM FILE

Column Name	Description
IID	Individual ID
FID	Family ID indicates if individual belongs to a family set.
PID	Paternal ID indicates the IID of the individual's father.
MID	Maternal ID indicates the IID of the individual's mother.
SEX	The recorded sex of the individual (1=male, 2=female)
Phenotype	The case status of the individual (-9/0 = missing, 1= unaffected, 2 = affected)

TABLE 4.14: FORMAT OF .FAM FILE

IID	FID	PID	MID	Sex	Phenotype
Sub123	Fam01	Sub124	Sub125	2	2
Sub124	Fam01	0	0	1	1
Sub125	Fam01	0	0	2	2

The .bim file provide information about SNP characteristics that are used to consider the location, assigned reference/ name of the SNP and allele representation.

TABLE 4.15: FORMAT DESCRIPTION OF .BIM FILE

Column Name	Description
Chrom	Chromosome of marker
Var ID	Variant ID
Pos	Position in morgans or centimorgans
BP	Base-pair coordinate (Refer to 2.1.1
Allele 1	Corresponding to clear bits in .bed; usually minor
Allele 2	Corresponding to set bits in .bed; usually major

TABLE 4.16: FORMAT OF .BIM FILE

Chrom	Var ID	Pos	BP	Allele 1	Allele 2
2	rs1045485	0	202149589	A	G
13	rs1799944	0	32911463	C	T
17	rs28897696	0	41215920	G	A

The .bed file is represented in machine language that unfriendly for readability in humans but contains a readable format of 8-bit codes that correspond to genotype code, further to this it also maps the information between the .fam and .bim files.

TABLE 4.17: FORMAT DESCRIPTION OF GENOTYPE DATA IN .BED FILE

Genotype Code	Description
00	Homozygous for first allele
01	Missing genotype
10	Heterozygous
11	Homozygous for second allele

4.5 DISCUSSION

During this chapter, data representations and transformations were described for reference in further chapters when analysing data due to the complex variation of data hypothesis. Further to this, the breast cancer DRIVE dataset was discussed outlined ~28K subjects for analysis from a multitude of originating countries, and a wide range of age. Already discussed are some exclusion criteria including age < 40 and individuals of African American ancestry. At this point, the focus of the analysis is not considering the tumour histology in order to retain a larger subject sample size. Also outlined, the criteria of invasive breast cancer histology; participants that have confirmed cancers cells that have spread to the surrounding breast tissue. This ensures that the patients are under the influence of a developed cancer that is more likely to show a genetic link. The next chapter introduces the proposed methodology, describing the steps of the process as well as indicating the reasons for each decision.

Chapter 5: Proposed Methodology: RaSaR

Random Sampling Regularisation (RaSaR) is a proposed methodology that aims to improve selection criterion for epistasis, reduce false discovery and cater for non-intensive computational requirements by prefiltering features using regularisation (See 5.4). The following selection outlines the stages of the methodology and explains the processes involved, the reason for their selection, how they benefit the method and what input and output is provided for each stage. Cascading processes rely on information from the previous process in order to perform their function, therefore introducing the first stage, Quality Control that prepares the data for further analysis to remove bias and erroneous data. Cohort Extraction and Association analysis provide the details for the main deviation of this methodology from standard case-control. Feature selection introduces the threshold choice and the regularisation technique that presents a main part in false positive reduction objective (RO2). for this research and explains how the success of the previous stages is presented in this stage. Having prepared and selected the most promising features according to this methodology, the epistasis stage will outline relationships that pass a significance threshold. Finally, inference analysis will expand, analyse and represent the information to infer the relationships that exist in the outlined combination within this sample population. To demonstrate the methodology Figure 5-1 provides a visualisation that encompasses the stages of the methodology and shows the transfer of data and information during the process.

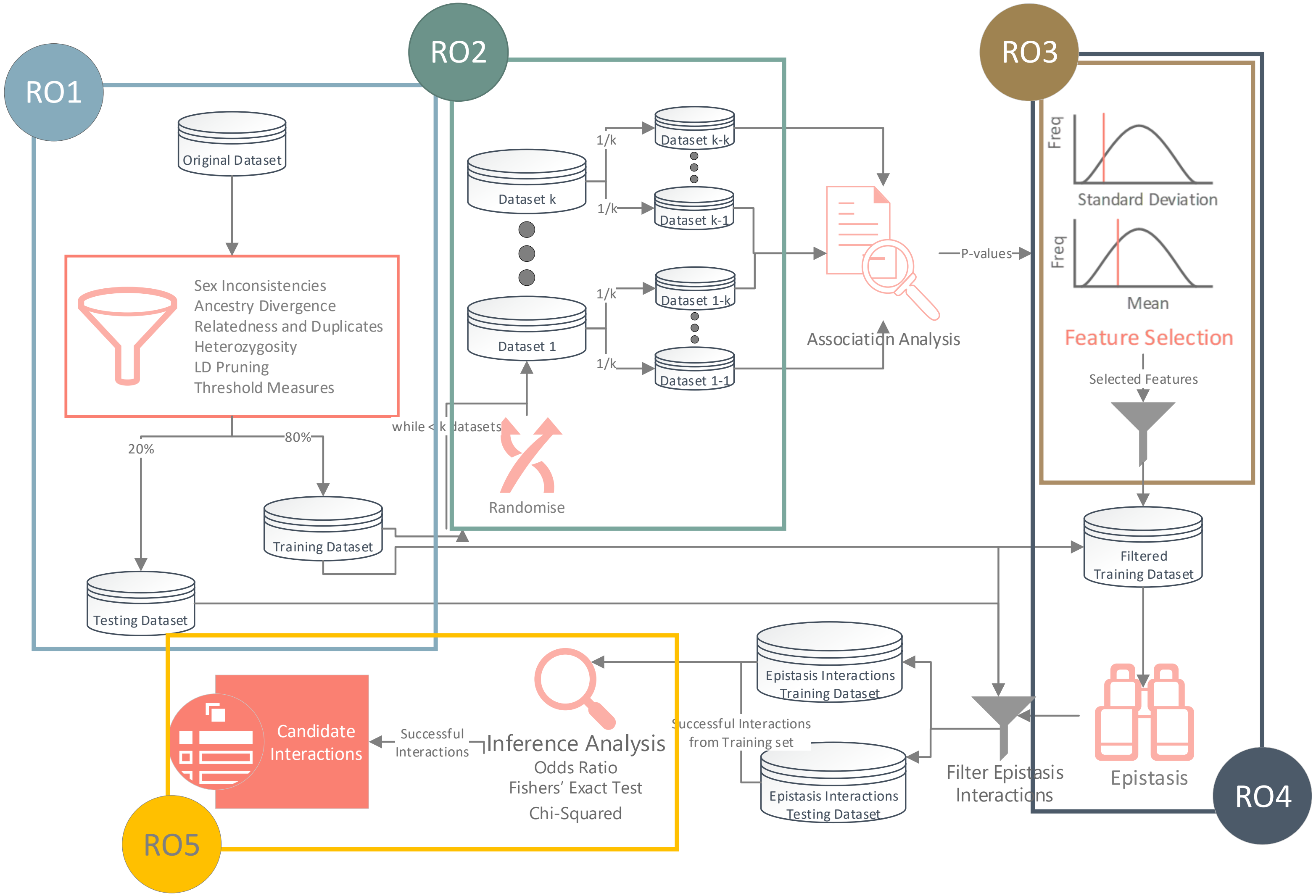


Figure 5-1: Visualisation of the proposed methodology

5.1 STAGE 1: QUALITY CONTROL (QC)

QC is a pre-processing stage to remove erroneous data and information that may cause bias in the study. This stage is standard practice in the genomic research and can vary among studies. The proposed methodology adheres to standard practices in GWAS [115][93][97]; adopted in many studies [136][52] employing standard processes; ancestry divergence, sex inconsistencies, heterozygosity, relatedness and duplicates in subjects, LD pruning and common threshold measures, MAF, GENO, MIND and HWE. Conservative threshold measures are applied to create a reliable dataset that is devoid of missing values and information that could cause errors later.

TABLE 5.1: QUALITY CONTROL PROCESS THRESHOLDS

Process	**Threshold**
Ancestry Divergence	0.6
Sex Inconsistencies	$0.2 < x > 0.8$
Heterozygosity	$\bar{X} + 2\sigma$
Relatedness and duplicates in subjects	$x < 0.125$
LD Pruning	$x < 0.8$
MAF	$x > 1\%$
GENO	$x > 99\%$
MIND	$x < 1\%$
HWE	$x < 1 \times 10^{-4}$

The processes in Table 5.1 vary in threshold leniency in order to accommodate for the epistasis approach. While sex inconsistencies along with relatedness and duplicates in subjects adheres to the standard thresholds, the remaining processes are variable and should respond to data structure and study design. Heterozygosity is commonly confirmed based on visual and/or statistical output; for this particular threshold, the measure was based on the observed structure to streamline results, refer to 6.1 for further evidence. Similarly, ancestry divergence is based on the visual representation and thresholds are decided in response to this, refer to 6.1 for further evidence. LD pruning was processed using a lenient threshold in order to retain a representative SNP dataset while removing features to reduce noise. Threshold specified for MAF is lenient given that only 1% of minor alleles need to be present to retain the feature; this decision was in response to the study design, as these processes will be dependents for an Epistasis approach both common and rare alleles could present as interactions. Further to this, the remainder of the threshold measure are based on standard practice methods [93]. For more information regarding the processes involved in the QC stages, refer to section 3.1 .

5.2 Stage 2: Random Cohort Sampling

To respond to the common issues that are present in GWAS, the cohort extraction is a preliminary stage in RaSaR to extract random cohorts of individuals that can represent a real-world sample cohort for analysis. Using this method, the data is prepared to explore in later stages if a constant effect is common across significant SNPs (which it should be if it is truly associated). Further to this, the effect of population structure is evident in many studies; this is a difficult problem to solve unless the study data has been obtained from a purpose-built clinical study. Therefore, the randomisation of the data can disperse the effect of the population structure among the cohort to reduce the effects.

Having performed QC, excluding any outlier subjects and/or SNPs, in a standard GWAS the remaining subjects are used to perform an association analysis to provide a resulting set of SNPs with their corresponding p-values to indicate the probability of significance to the phenotype. However, in order to consider the varying presence of the SNPs across clusters of individuals, the following step produces sets of subjects that consider both the potential effects of separating the cohort and randomising the subjects within.

The first stage is to randomise the subjects by phenotype status and split the dataset into n sections depending on the number of desired folds used to perform the analysis, i.e. n = 3. The training cohort is split into 3 subsets of randomised individuals with proportionate levels of cases and controls. This is then repeated 2 more times, resulting in a total of 9 subsets of which all derive from the original training cohort but contain varied subjects for further analysis in the next stage. By varying to subjects randomly, the probability of obtaining a false positive across all subsets of the training cohort is dramatically reduced.

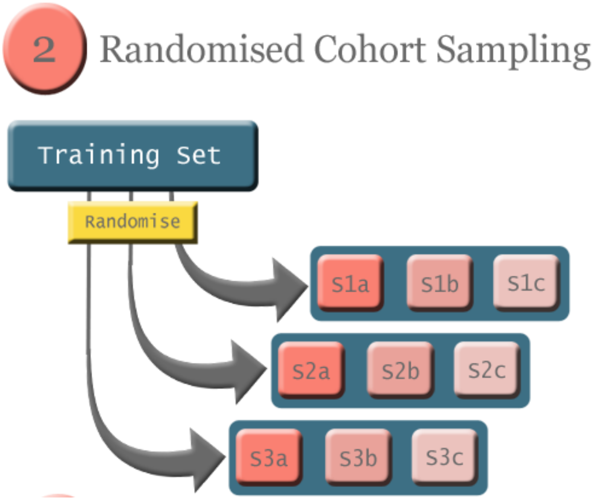


Figure 5-2: Representation of Random cohort sampling

This stage introduces the initial deviation from the standard methods employed in GWAS. Standard methods input the full set of observations and features into association analysis to evaluate the probability of significance between feature X and response Y; however even in large datasets, bias is likely to occur given the size of the feature set. Some GWAS are utilising multiple datasets to validate results however this can cause additional problems such as differences in population structure established in the initial study as well as difficulty in attaining a further dataset particularly for disease or disorders that are difficult to genotype large cohorts in case status. Therefore, this stage produces cohort sets for a method that in inspired by the 'weak learners to strong learners' [166] approach as seen in machine learning algorithms such as Random Forests. The main purpose of this approach is to use a number of weak models to produce a strong model using a 'voting system'. To elaborate further, 9 p-values produced for one feature will dramatically reduce the chances of producing false positives as all cohorts must be constructed in such a way that by chance a bias has occurred in each sample, while not impossible, it significantly reduces the probability. During this stage, a sample size is defined depending on the size of the dataset and the number of folds being analysed. From here, the phenotype is established, and random permutations are used to generate cohorts of the size defined which adhere and retain the phenotype distribution.

5.3 Stage 3: Association Analysis

Association analysis models vary in the outcome information, this puts importance on choosing the most appropriate model for the approach. As further analysis is used to investigate the information beyond this point, there is affordance to use the allelic model analysis and gain the benefits of the increased statistical power. During this stage, multiple testing is not utilised but will be addressed in later stages, however genomic control is used to control for population structure.

5.4 Stage 4: Feature Selection

This stage uses the results of the association analysis to produce a subset of features that show significant association to the given phenotype. The results from the association analysis are combined to produce a mean GC-value and the corresponding standard deviation which will provide information as to how much the value is shifting across the subset cohorts. This will provide information as to whether the SNP is consistently associated with the phenotype or is falsely associated with a sample of subjects. Continuously mentioned in literature, is the 'statistical power' of a study, within genomic studies it is generally accepted that the bigger the cohort the less likely a SNP will show false associations; this is due to the normalisation of data with the addition of more observations. While this is true, consider that the number of features that are tested during genomic studies is large and as a result the likelihood of producing a false positive is also increased. By splitting the cohorts into n*n sections, our sample size is improved by n times.

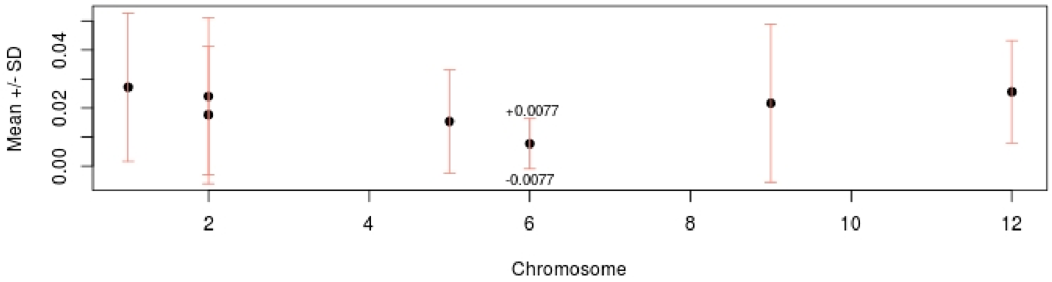


FIGURE 5-3: REPRESENTATION OF RASAR FEATURE SELECTION

Having produced the mean GC-value and associated standard deviation, σ, between all subsets of the cohort, a cut-off point is applied. Depending on the desired outcome, a conservative threshold can be applied which will indicate the most significant SNPs consistently associated among the feature set. Alternatively, a more lenient threshold can be achieved by using a larger σ value. The μ and σ value regularise the SNPs to aid in outlining candidates, therefore the consistency of the SNP should still be significant among the tested cohorts. Demonstrated in Figure 5-3 is the aim of the process, by choosing features that adhere to a particular σ threshold.

This stage continues the deviation from the standard approaches in GWAS, introduce a new feature selection method that uses the consistency of SNP presence as the dominant driver for selecting a candidate feature set. This produces regularised SNP selections that are consistently associated to the phenotype, employing threshold based on both the standard deviation and mean across the resulting GC-values. The likelihood of producing a false positive result at this stage is significantly reduced, as the subsets of cohorts should regularise the resulting values, bypassing the issues associated with chance probability signals. However, it should also be noted that a strong false positive feature signal that presents significantly across the cohort is still likely to remain at this stage.

During this stage, we have chosen to benchmark the results of the analysis against standard methods and robust techniques to compare the outcome information (See 6.4 for more information on the chosen benchmarks). Refer to 3.2.3 for further explanation on correction methods.

5.5 STAGE 5: EPISTASIS

We have chosen to use LAMPlink [140] due to the benefits of limitless arity with the additional benefit of speed. Acknowledging the use of a dominant model leads to sacrificing potential combinations, the purpose of the method is to explore the effects of using random sampling regularisation to produce a set of resulting candidates while reducing FP and FN error rates. The method relies on established open source software programmes with which the underlying statistical methods can be interpreted and understood for reproducibility and understanding of the methods employed. One of the most common programmes for epistasis analysis is LAMPlink, an open source software programme that adapts its methods from the standard and well-established programme PLINK. The following section outlines the statistical methods and techniques employed by LAMPlink to produce epistasis results.

The purpose of this stage is to sift through the combinations of SNPs to outline potentially significant relationships for further analysis. The use of a software programme that employs a limitless arity approach produces exhaustive results that investigate the relationships that exists between all combinations (excluding those eliminated during reduction techniques) and the focus phenotype. This benefits the methodology as it considers a larger feature set that would otherwise be impractical to explore via normal statistical techniques. In the next section LAMPlink is discussed further.

5.5.1 LAMPLINK

A software programme called LAMPlink derives its name from methods employed 'Limitless Arity Multiple-testing Procedure'. Provides a method of detecting significant associations using a large number of features. Generally, epistasis programmes will perform epistatic interaction tests using two-way feature sets e.g. PLINK; this significantly reduces the exploratory power of epistasis by by-passing the potential for component clusters of 3 or more features. LAMPlink tests the potential of every possible combination while reducing the number of tests performed by adjusting the number of SNPs based on [85] complexity correction. This significantly reduces computational complexity and also reduces the time-consuming process that is generally associated with epistasis approaches. The following section outlines the process of LAMPlink.

5.5.1.1. LAMPLINK GENETIC MODELS

LAMPlink uses a binary representation for SNPs that can be formatted using two models. Similarly, is previously introduced, the dominant and recessive models measure the significance of a SNP in terms of the presence of recessive alleles in either heterozygous or recessive homozygous form. The following tables represent the binary transformations adopted in LAMPlink for 2 and 3 SNPs although more feature combinations are analysed.

TABLE 5.2: LAMPLINK DOMINANT MODEL REPRESENTATION FOR 2 SNPS

	BB (0)	**Bb (1)**	**bb (1)**
AA (0)	0	0	0
Aa (1)	0	1	1
aa (1)	0	1	1

TABLE 5.3: LAMPLINK RECESSIVE MODEL REPRESENTATION FOR 2 SNPS

	BB (0)	**Bb (1)**	**bb (1)**
AA (0)	0	0	0
Aa (1)	0	0	0
aa (1)	0	0	1

TABLE 5.4: LAMPLINK DOMINANT MODEL REPRESENTATION FOR 3 SNPS

	BB (0)	Bb (1)	bb (1)	
AA (0)	0	0	0	**CC (0)**
	0	1	1	**Cc (1)**
	0	1	1	**cc (1)**
Aa (1)	0	0	0	**CC (0)**
	0	1	1	**Cc (1)**
	0	1	1	**cc (1)**
aa (1)	0	0	0	**CC (0)**
	0	1	1	**Cc (1)**
	0	1	1	**cc (1)**

TABLE 5.5: LAMPLINK RECESSIVE MODEL REPRESENTATION FOR 3 SNPS

	BB (0)	Bb (1)	bb (1)	
AA (0)	0	0	0	**CC (0)**
	0	0	0	**Cc (1)**
	0	0	0	**cc (1)**
Aa (1)	0	0	0	**CC (0)**
	0	0	0	**Cc (1)**
	0	0	0	**cc (1)**
aa (1)	0	0	0	**CC (0)**
	0	0	0	**Cc (1)**
	0	0	1	**cc (1)**

Each model analyses the data in difference capacity and therefore both models are important to consider, although note that the dominant model is more effective in producing results given that there is a limited number of possible combinations to be tested in recessive models.

LAMPLINK STATISTICAL ANALYSIS

Statistical analysis of the binary representation is performed using one of two techniques, fisher's exact or Chi-squared test. This can be changed depending on the parameter passed into the program at the time of analysis. For further information about fisher's exact and Chi-Squared test, please refer to section 3.3.1 .

LAMPLINK MULTIPLE TESTING, LD AND COMPUTATIONAL COMPLEXITY

As previously mentioned, the multiple testing problem concerns large dataset analysis as the probability of producing a false positive is almost certain. Therefore, LAMPlink uses the complexity inequality technique [140] that employs the assumption that probability of at least one significant result showing is less than or equal to the sum total of the probabilities taken individually. Using the Bonferroni inequality FWER upper bound, the probability of producing one or more false discovery, the tests performed are separated into testable and untestable categories that assigns the possibility of producing a false positive into possible and not possible, respectively. This is to reduce the number tests included as the untestable test will not increase the FWER value. Further to this, LAMPlink also use a technique of ‘itemset mining’ to parse through large sets of features to extract combinations that commonly occur across the population. These techniques reduce computational complexity and time. An additional process that can be employed is to eliminate SNP combination that contain SNPs in high LD to each other. Using LD at this later stage will allow for the creation of combinations that may be reliant on one SNP that would have previously been removed during LD pruning but within a combination of SNPs is relevant leading to a significant combinatorial collection of features that may have a relationship with the focus phenotype.

LAMPLINK OUTPUT

LAMPlink produces several files that contain the significant combinations of SNPs along with complementary information for SNPs. Analysed SNPs result in singular, two-way and multiple combination SNP outcomes. When selecting significant SNP combination, not only the probability of the combination should be considered, but also the significance of the singular entity. If one of the singular entities results in a higher significance than the total combinations, then this may indicate that the combination is only significant due to the presence of the singular entity. However, this may also indicate that only a small population of the cohort is affected by the combination, therefore the significance is reduced but still significance.

TABLE 5.6: OUTPUT FOR .LAMP FILETYPE

Output	Description
COMBID	ID number for combination
Raw_P	p-value before adjustment via Bonferroni Inequality FWER
Adjusted_P	p-value after adjustment via Bonferroni Inequality FWER
COMB	Resulting SNPs for significant combination < significance

TABLE 5.7: OUTPUT FOR .LAMPLINK FILETYPE

Output	Description
CHR	Chromosome location of SNP
SNP	SNP name
A1	Minor allele nucleotide represented as A, G, C or T
A2	Major allele nucleotide represented as A, G, C or T
TEST	Test performed
AFF	Genotypes in cases
UNAFF	Genotype in controls
CHISQ	Chi-squared statistic
DF	Degrees of Freedom used for test
P	P Value statistic for test
OR	Odds Ratio value
$\mathbf{COMB_i}$- $\mathbf{COMB_n}$	Columns denoting n Combinations with binary representation for the presence of SNP in each combination.

The output file of LAMPlink lists results whose adjust p-value < significance threshold. These results can contain combination of 2 to n SNPs, with n being the total number of SNPs input into the analysis if there exists a significant relationship between them. LAMPlink also produce singular entities that can be used for further analysis but also allow for the comparison of combination against singular entity adjusted p-values. Although not an elimination factor, if a singular entity has a higher adjusted p-value than its counterpart combination, caution should be taken as this could indicate that the combination's significance is as a result of the significance of the singular entity. It is important to note that even if a combination has a lower p-value that it's singular entity, there could still exist a subpopulation that shows increased significance (as later demonstrated in results).

Figure 5-4 demonstrates an example of the **Petal Plot Policy (PPP)**, with each additional red petal that measures as more significant in its singular form than the interaction, the less confidence is found in the interaction as indicated by the colour of the centre of the petal plot. Further to this, confidence should also be lowered based on the number of variants that exist on the same chromosome. The following equation can be used to determine a confidence value for each interaction.

$$PPPconfidence = \left(\frac{r}{n} \times 0.5 \right) + \left(\frac{n-s}{n} \times 0.5 \right) \qquad \text{EQ. 5-1}$$

Where *r* represents the number of petals with significance greater than the interaction, *s* represents the number of petals located on the same chromosome and *n* represents that total number of petals.

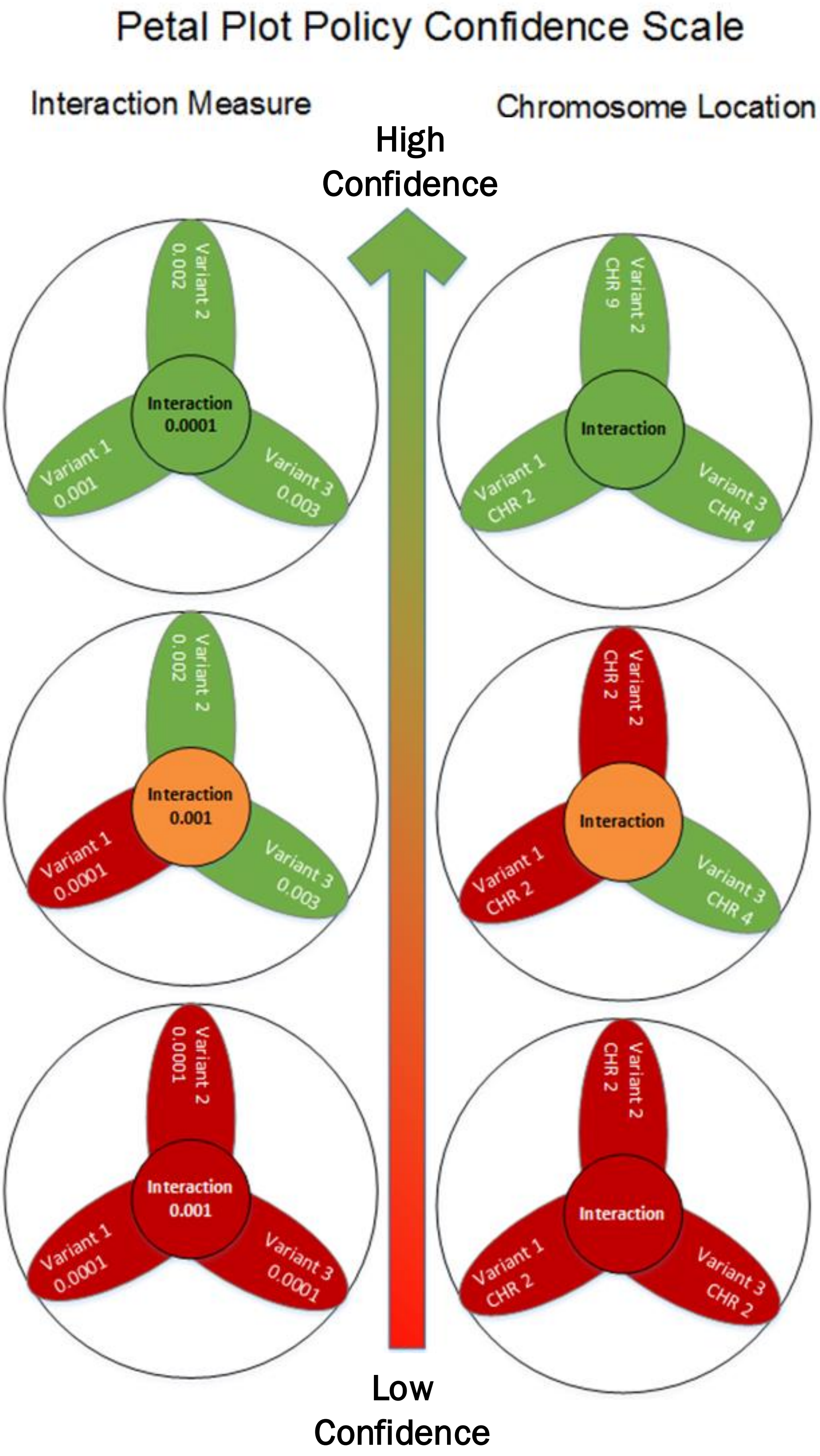


FIGURE 5-4: PETAL PLOT POLICY CONFIDENCE SCALE [A]

[A] Using two different metrics, the interaction measure considers whether the SNP p-value is greater than the interaction p-value and if so, the petals are represented as red. The chromosome location marks petals red if there are 2 or more petals that sit within the same chromosome.

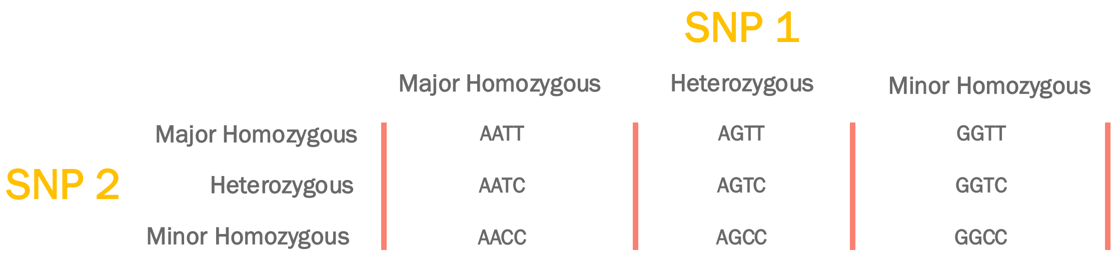
5.6 Stage 6: Inference Analysis

During this stage, the relationships outlined by LAMPlink are further analysed. As LAMPlink is only used to outline the potential relationship, this stage is used to expand the relationships outlined and to further analyse them to confirm or disregard the findings. During this stage, relationships that adhere to the petal plot policy will be extracted from the training set, with allelic and genotypic states combinations explored; as all combination states would be exhaustive to perform manually, the following models will be performed during this process:

Table 5.8: Model Descriptions for Inference Analysis

Model	Description
Dominant	All genotype states that contain a minor allele (homozygous recessive and heterozygous) will be measure against homozygous dominant states.
Recessive	All genotype states that contain a major allele (homozygous dominant and heterozygous) will be measured against homozygous recessive states.
Additive	Genotypic stages homozygous dominant, homozygous recessive and heterozygous will be measure against one another.
Cross-Genotype state	Every combination that is present in subjects for the explored genotype combination will be measure against the remainder.

Cross-Genotypic refers to the analysis of singular interaction states between SNPs as demonstrated in Figure 5-5 using inputs SNP 1 (Major Allele (A) Minor Allele (G)) and SNP 2 (Major Allele (T) Minor Allele (C)) (See 4.2 for more information on major and minor alleles). This allows for an exhaustive search of the genotype states for the interactions. All of these interaction states would be tested to produce a value that would indicate whether they were statistically significant. By splitting SNP states across genotypes, observation of a small cohort with high association combination may be possible.



SNP 1

SNP 2		Major Homozygous	Heterozygous	Minor Homozygous
	Major Homozygous	AATT	AGTT	GGTT
	Heterozygous	AATC	AGTC	GGTC
	Minor Homozygous	AACC	AGCC	GGCC

Figure 5-5: Cross-Genotype Demonstration

The inference of information is important in order to conclude assumptions and present association and relationship that show significant in the research. For this process multiple methods are used, firstly considered are the methods that will be used to statistically determine information based in the data provided. Two methods have been chosen for testing the significance of an association: Fisher’s exact test and X^2. These methods are tried and tested across many field and disciplines and are widely used in the genomics field. By using 2 methods, the limitations of each method can be outlined by validating the other as they should come to very similar conclusions as to the significance of an association.

Two methods have been chosen for the measure of association. As the data used within this research is case-control, a measure of the risk ratio is not possible without an odds ratio score. While an odds ratio score will provide an estimate of the measure of association, this can be misinterpreted therefore the risk ratio will compensate for this. The next process is to test the significant association found with a testing set of data that has not been used so far. This indicates whether the association is consistent in other samples and also removes any associations that have been outlined as false positives.

Presented in Table 5.8 are the models used to analyse the information, of these models there are combinations of missing information that should be explored before any conclusions can be drawn. Previously discussed in section 3.3.2 are classification techniques that in this instance will be used as a feature selection process to outline the most significant combinations of variants. To demonstrate this, Figure 5-6 show the states that have been analysed by this point (example State 1) but the analysis has not taken in consideration the potential that two or more alleles (example State 2) could be more significant regardless of the information that is present as a genotype.

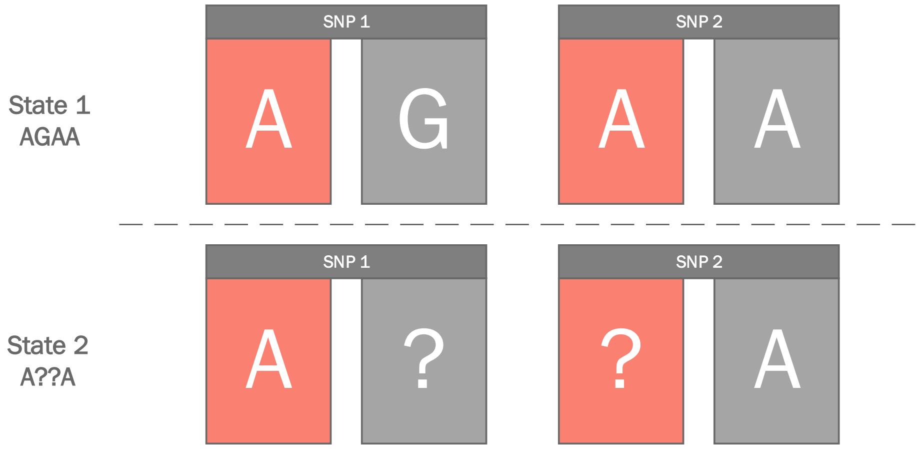


FIGURE 5-6: SNP STATE A??A [A]

[A] Pink represents SNP allele 1, Grey represents SNP allele 2

To perform that analysis, a CHAID model is used. The CHAID tree offers an interpretable option for exploratory analysis of variables using standard methods that are tried and tested. Using the ‘trunk’ of the tree to demonstrate the response variable and the categorical percentages, the independent features are analysis to outline the feature with the greatest impact to the response variable. The categories of the feature with the greatest impact are then analysed for their significance to the response variable with a threshold of alpha < 0.05 using the chi-squared test of association and the complexity correction method for type 1 errors. If the variable remains significant it is added to the tree. This is an iterative process that is performed with each feature that is added to the tree until the number of response variables is too low to reliably produce a split.

CHAID parameters allow for additional adaptations and corrections such as cross-validation, the ability to define the minimum cases for node split, the maximum number of nodes and the probability for merging and splitting. Within the proposed methodology, the parameter of cross validation has been utilised and no further parameters defined in order to outline even small patterns of significance. For the purpose of this methodology, only categorical variables will be used, but CHAID is also applicable to continuous variables of which the values are categorised into bins. Using the combinations output during epistasis stage, further analysis is required to evaluate its reliability, replicability and the existing statistical relationship between the feature set and the phenotype. In this research, the definition used to measure the reliability of the outlined combinations refers to the statistical confirmation using standard statistical methods Odd’s Ratio (OR) and Fisher’s Exact Test which use training and test sets to replicate and confirm or reject the outlined interactions (See 3.3.3 and 6.6 for more information on the statistical measures).

5.7 Discussion

Discussed is the proposed methodology, each aspect of this methodology has carefully considered and serves a purpose for filtering and extraction of features of significance approaching the problem for the detection of epistasis. The methodology aims to address a series of issues including:

Reducing FDR	Stage 2 introduces the first stage that aims to address the issue of FDR by tackling the problem before it occurs. Finalising in Stage 3 using multiple association analysis to build a picture of how genetic component behaves among varying sample populations.
Computational expense	Exhaustive Epistasis searches requires extended time allowances or require substantial hardware for processing. This method aims to outline a representative set of SNPs that do not require substantial hardware but is a small enough set of representative SNPs that epistasis can be performed on in a reasonable time constraint.
Improvement of epistasis detection	One of the most challenging problems in epistasis is the detection of interactions while accounting for the influencing pitfalls of FDR, computational complexity and the statistical filtering that is commonly used to reduce this. This method aims to outline the most prominent SNPs for epistasis from a large feature set that can commonly become lost in the expanse of information.
Concise identification of interaction combinations	Further to the identification of SNP for epistasis, the aim of this method is to concisely outline combinations that show significance with the phenotype. Commonly many combinations will be outlined for significance with the phenotype due to FDR and SNP selection; the aim of this method is to combat these issues.

Further to this, the methodology addresses the problems defined for this methodology using interpretable methods and introducing procedure and best practice to optimise the use of this method. In order to verify the claims of this thesis, the next section applies the proposed methodology to the data described in section Chapter 5:, using standard approaches and thresholds to benchmark the success of the methodology.

Chapter 6: RESULTS

The following section presents the results obtained using the proposed methodology as outlined in Chapter 5:. In order to measure the effectiveness of the proposed methodology, alternate cases of standard practice have been provided with the resulting output when analysing for Epistasis. The processes and output of quality control are outlined providing visualise aids to represent some of the processes.

Stage 2 provides the outcome frequencies for the subsets of individuals that are to be used in the following stage. Stage 3 outlines the results from the association analysis, providing a comparison against standard case-control and visualising the vast difference between output values. Feature selection introduces the methods of evaluation that will be used to benchmark the performance of the methodology in comparison to using standard case-control methods.

This describes the processes of obtaining the features selected and an overview of the information. Stage 5 performs 4 separate epistasis analysis for features outlined from each method. Given that many of the same features have been outlined through each method with slight variations, the output combinations of each method have been combined to show the detected combinations from each case analysis. Stage 6 then uses the information from Stage 5 to expand the outlined interactions and analyse the information to infer the relationship that exist while testing the significance using separate datasets to observe retention in significance. This chapter concludes with an in-depth results discussion that outlines the results and further demonstrates the performance of the methodology.

6.1 Stage 1: Quality Control

QC was performed as outlined in the proposed methodology (See section 5.1). Plots provide a visualisation for QC outcomes and threshold decisions. Table 6.1 below records the number of removed observations and features after each step has been performed and provides a breakdown of the processes, how many SNPs or Individuals were available before QC, removed with each stage and remained after QC. The remaining dataset is comprised of 13,649 (7136 cases), (6513 controls) observations and 320,247 features, or SNPs. Additionally Figure 6-1 to Figure 6-3 provide visualisations of processes Heterozygosity, Relatedness and Duplicates and Ancestry Divergence to demonstrate the thresholds used to remove individuals during this process.

Table 6.1: Quality control process exclusion values

Process	Removed		Remaining	
	Subjects	Variants	Subjects	Variants
Before QC	***-***	***-***	***28281***	***528620***
Ancestry Divergence	675	-	27606	-
Relatedness and Duplicates	7750	-	19856	-
Heterozygosity	301	-	19555	-
Sex Inconsistencies	72	-	19483	-
Linkage Disequilibrium Pruning	-	116115	-	412505
Threshold Measures				
Missingness in Individuals	4399	-	15084	-
Genotype Call Rate	-	21561	-	390944
Hardy Weinberg Equilibrium	-	1269	-	389675
Minor Allele Frequency	-	69428	-	320247
Missing Phenotype	1355	-	13729	-
Exc. Criteria: Age < 40	80	-	13649	-
After QC			***13649***	***320247***

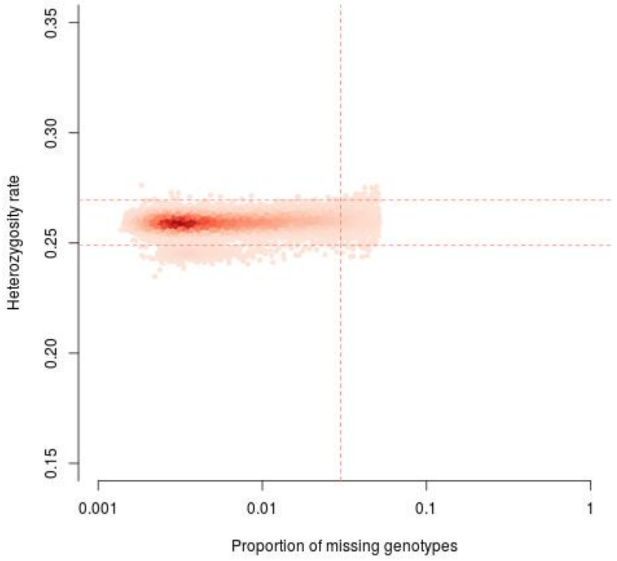


Figure 6-1: Heterozygosity density plot with exclusion thresholds at x and y axis. Horizontal red thresholds indicate 2σ from the density mean. Vertical red threshold shows outliers by missing genotype proportion.

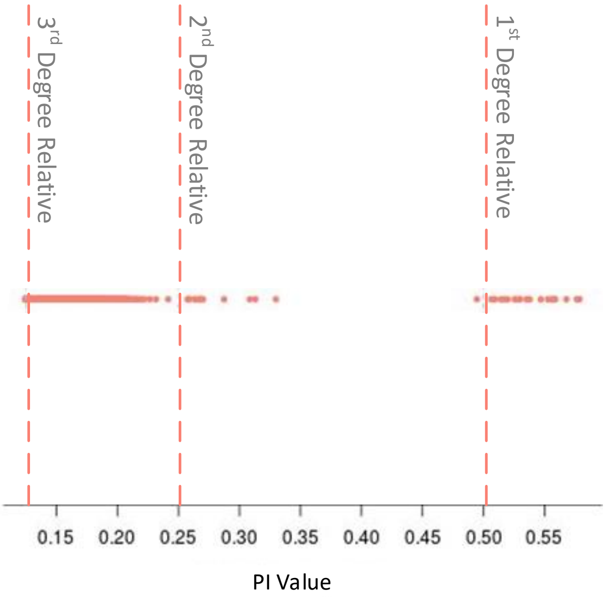


FIGURE 6-2: SUBJECTS EXCLUDED FOR RELATEDNESS OF FIRST, SECOND OR THIRD DEGREE.

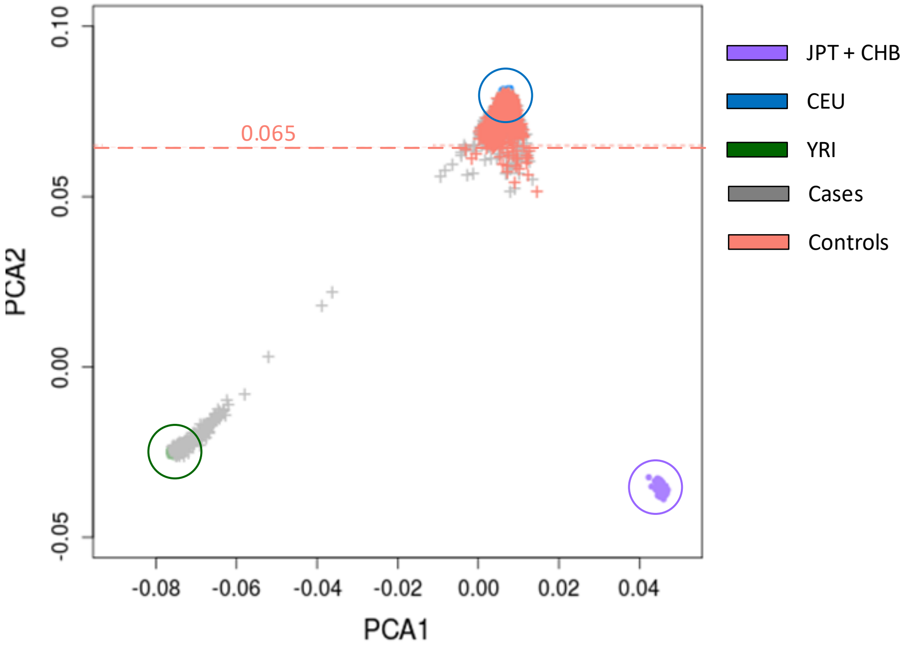


FIGURE 6-3: ANCESTRY DIVERGENCE USING 3 POPULATIONS TO PLOT THE ASSOCIATION.

6.2 Stage 2: Random Cohort Sampling

Figure 6-4 provide the frequency of case and control status subjects that were included in each fold. S includes the total case and control status subjects included in the training set. Within each table, the subset is identified by the fold number 1, 2 or 3 and the subset letter a, b or c. Therefore, S1$_a$ includes information for the first fold and the subject status distribution for subset a. S1$_b$ includes the subject's distribution between case and control for fold 1, subset b. S2$_a$ includes the distribution between case and control for fold 2, subset a.

$$
\begin{aligned}
S &= \{S1, S2, S3\} \\
S1 &= \{S1_a, S1_b, S1_c\} \\
S2 &= \{S2_a, S2_b, S2_c\} \\
S3 &= \{S3_a, S3_b, S3_c\}
\end{aligned}
$$

	Control (0)	**Case (1)**
S	5214	5706

		0	**1**		**0**	**1**		**0**	**1**
S1	**S1$_a$**	1702	1938	**S1$_b$**	1772	1868	**S1$_c$**	1740	1900

		0	**1**		**0**	**1**		**0**	**1**
S2	**S2$_a$**	1723	1917	**S2$_b$**	1789	1851	**S2$_c$**	1702	1938

		0	**1**		**0**	**1**		**0**	**1**
S3	**S3$_a$**	1701	1939	**S3$_b$**	1800	1840	**S3$_c$**	1713	1927

FIGURE 6-4: DISTRIBUTION OF CASE-CONTROL STATUS PER FOLD SUBSETS

These tables outline the distribution of each subset; taking into consideration both the effects of sample size and case-control status. Each subset is extracted from the original file to produce a binary replication with only the assigned subjects. This is then used within the next stage, Association Analysis.

6.3 Stage 3: Association Analysis

The following section outlines the results obtained from performing association analysis using 2 different approaches; standard case-control and proposed random sampling regularisation method. Both approaches were conducted using the same techniques only varying the input information. An allelic model, adjusted by genomic control, included remaining subjects and SNPs from the quality control stage with variation conducted for the Random Sampling Regularisation approach using cohort samples as reported in 'Random Cohort Sampling'.

Commonly in standard case-control approaches, QQ plots are used to indicate the success of the quality control procedure. Although this is not a definitive measure, it can provide an indication if are any problems in the data such as population stratification. Figure 6-5 shows an expected tail-end deviation from the null hypothesis, which is the expected values.

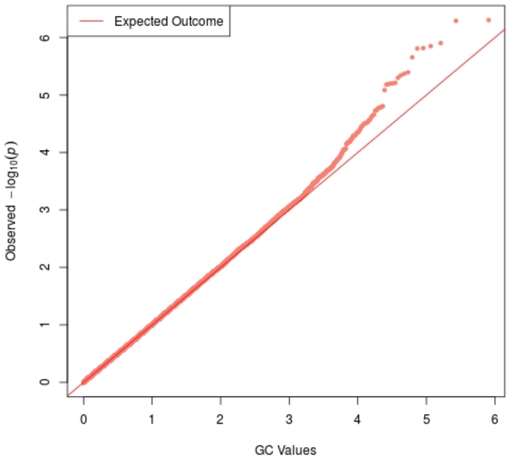


Figure 6-5: QQ-Plot shows the deviation from the null hypothesis line.
Deviation begins >3 which indicates that the quality control process was successful.

Figure 6-6 visualises the -log$_{10}$ p-values produced from the association analysis using standard case-control approach in a Manhattan plot. Values exceeding the blue threshold are indicated as suggestive significance. A standard approach in genomics is to use the genome-wide significance line (not visible on this plot), a more stringent threshold, however the results do not indicate the significance of any of the SNPs exceeds this threshold.

Figure 6-7 visualises the -log$_{10}$ p-values produced using the RaSaR methodology. Visible is the clear decrease in significance for all SNPs. As the mean of 9 p-values for each SNP is used to create the mean values, any sample p-values that show little significance for the SNP will reduce the mean value but will reduce the presence of False Positives based on chance. This figure is scaled to show the difference between the values generated from standard case-control process and random sampling regularisation method. Figure 6-8 provides a zoomed view of the random sampling regularisation method values.

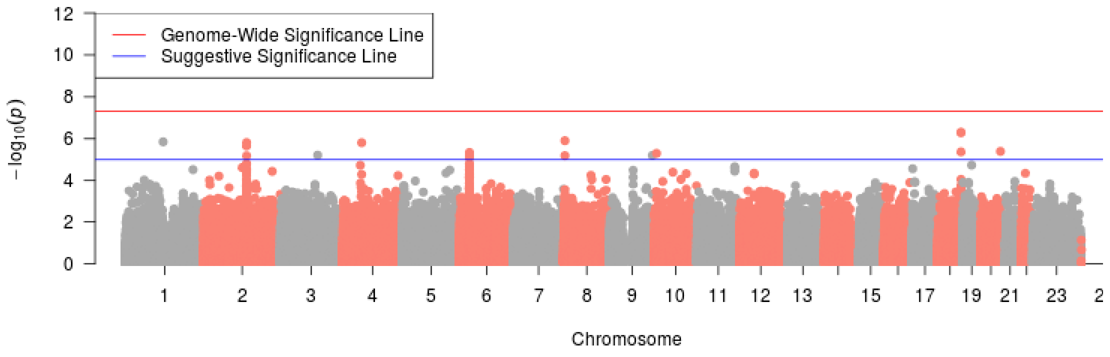


FIGURE 6-6: MANHATTAN PLOT FOR STANDARD CASE-CONTROL METHOD USING ALLELIC MODEL AND GENOMIC CONTROL.

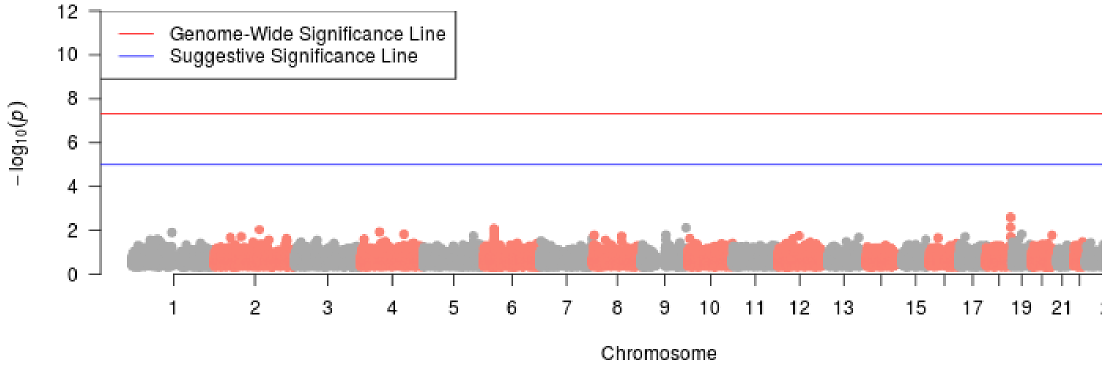


FIGURE 6-7: MANHATTAN PLOT GENERATED FROM MEAN OF 9 SNP -LOG$_{10}$ P-VALUES USING RANDOM SAMPLING REGULARISATION METHOD SCALED FOR COMPARISON TO STANDARD CASE-CONTROL

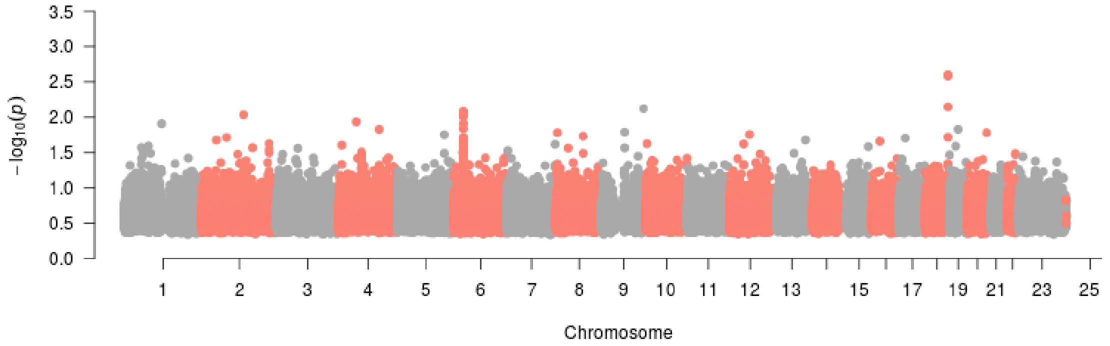


FIGURE 6-8: ZOOMED VIEW (FIGURE 6-7) MANHATTAN PLOT GENERATED FROM MEAN OF 9 SNP -LOG$_{10}$ P-VALUES USING RANDOM SAMPLING REGULARISATION METHOD

To further support this, Figure 6-9 uses the top SNPs values from chromosome 6 standard case-control approach (Figure 6-6) that exceed the suggestive significance threshold (See 3.2.1.3. for more information about genome-wide and suggestive significance thresholds) to represent the fluctuation in values when varying subjects between cohort samples. As demonstrated by Figure 6-10, the consistency of the top SNPs outlined by standard case-control methods fluctuate across the analyses but present strongly when using the full cohort.

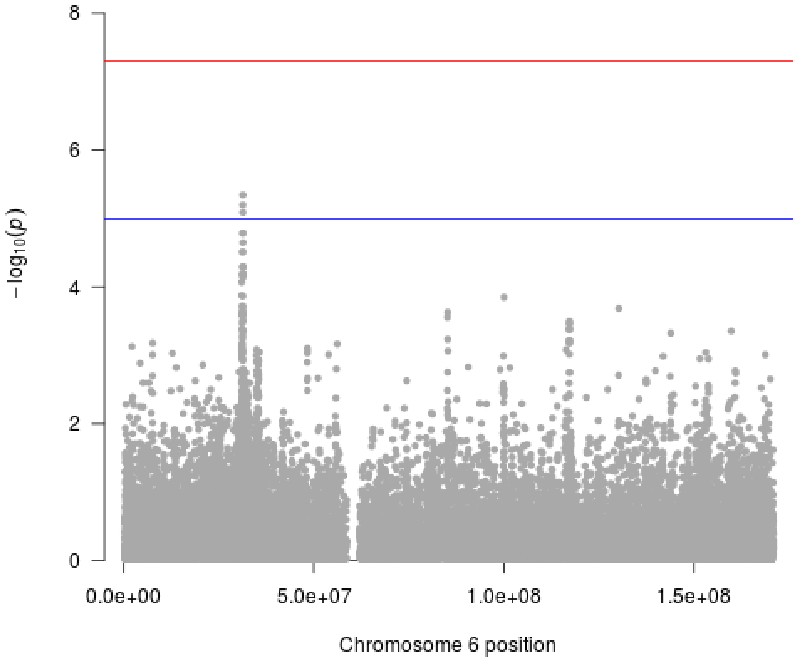


FIGURE 6-9: SAMPLE MANHATTAN PLOT SHOWING SNP -LOG$_{10}$ P-VALUES ACROSS CHROMOSOME 6 BY BASE-PAIR (BP).

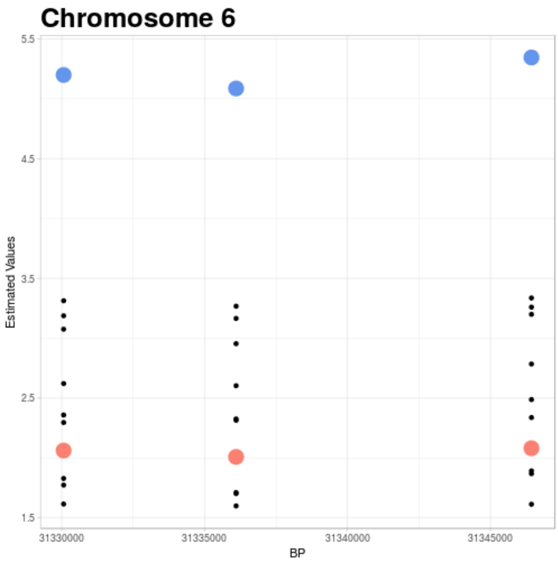


FIGURE 6-10: DOT PLOT COMPARISON OF STANDARD CASE-CONTROL VS. RANDOM SAMPLING REGULARISATION
The difference between the values produced using standard case-control methods (with genomic control), represented by blue points, and the values produced by random sampling regularisation. Each of the subset analyses are represented by black dots while the mean is represented in salmon.

6.4 Stage 4: Feature Selection

Producing a set of representative features that are most likely to indicate the presence of a significant relationship is one of the main challenges of this methodology. Benchmarking against standard approaches and correction methods will indicate the effectiveness of the methodology and could potentially outline areas of improvement. The following section outlines 3 additional cases of threshold choices that use multiple testing technique q-value (see section 3.2.3 to produce a feature set.

For benchmarking, a number of methods were considered including the genome-wide significance threshold (See 3.2.1.3.) and the p-value < 0.05, however each of these thresholds yielded a feature set that was either too small or too large to be used with the current set-up for epistasis. Figure 6-11 demonstrates the methods chosen in cases 2-3 long with the other considered statistical methods, with the novel methodology of this thesis being represented as case 1. As is visible, the genome-wide significance threshold was only able to yield 1 significant result which is not viable to perform an epistasis model with, and p-value < 0.05 yielded a feature set of >5000 which would increase the computational complexity of the methodology, requiring HPC.

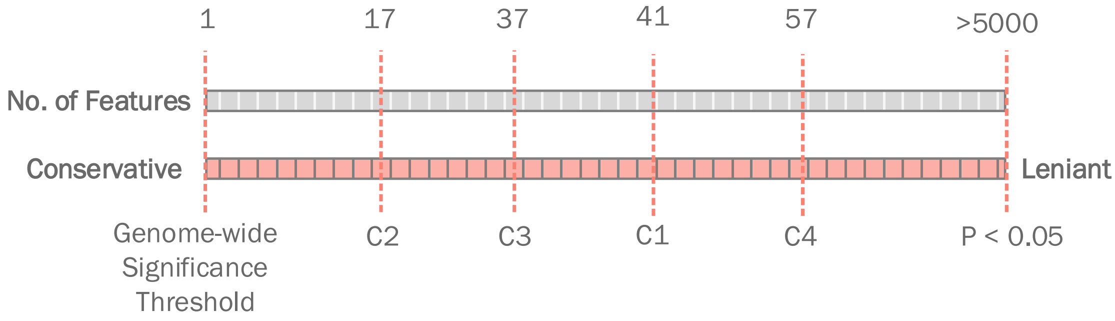


FIGURE 6-11: COMPARISON OF LENIENCE/ CONSERVATIVE METHODS ALONG WITH THE NUMBER OF FEATURES PRODUCED BY EACH. [A]

[A] Information on the bottom refer to the methods used including the ID for cases 1-4, as outlined in Table 6.2. Information on the top refers to the number of features outlined by each.

TABLE 6.2: THRESHOLDS AND FEATURES OF BENCHMARK CASES AND RASAR

ID	Adjustment Methods Applied	Threshold	Features
C1	RaSaR Cohort Sampling, RaSaR Feature Selection	$\mu < 0.05$, $\sigma < 0.025$	41
C2	q-value, Genomic Control	$\alpha < 0.3$	17
C3	q-value, Genomic Control	$\alpha < 0.4$	37
C4	q-value, No Genomic Control	$\alpha < 0.01$	48

Table 6.2 outlines the number of features and the thresholds used to filter those features. In order to decide the thresholds of the benchmarked cases and RaSaR feature selection methodology, the following sections outline the systematic approach.

6.4.1 CASE 2-3

The most conservative method out of the chosen benchmarks is the q-value in conjunction with the Genomic Control adjustment method. Therefore, in order to create a comparable set of features, case 2-3 were first explored. Table 6.3 shows the results from the multiple testing analysis.

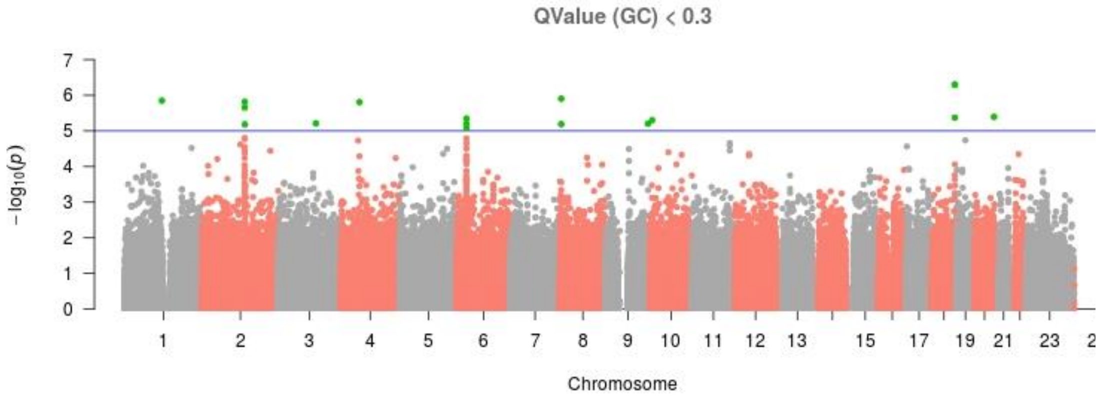


FIGURE 6-12: MANHATTAN PLOT SHOWING THE FEATURE SET SNPS HIGHLIGHTED IN GREEN WHEN APPLYING THE THRESHOLD Q(α) < 0.3 USING STANDARD CASE-CONTROL APPROACH (N = 17).

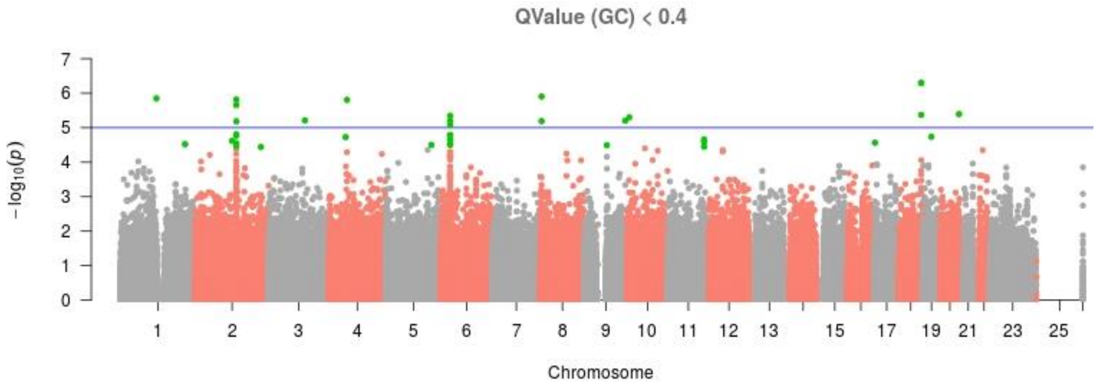


FIGURE 6-13: MANHATTAN PLOT SHOWING THE FEATURE SET SNPS HIGHLIGHTED IN GREEN WHEN APPLYING THE THRESHOLD Q(α) < 0.4 USING STANDARD CASE-CONTROL APPROACH (N = 37).

Visible in Table 6.3, there is a lack of significant results that are present for q-value. Given this, the common threshold $\alpha < 0.05$ has been increased to 0.3. While this is high, this only results in feature set of 17 SNPs, a low amount when analysing for epistasis. Selected features are visualised in Figure 6-12 highlighted in green.

TABLE 6.3: FDR VALUES FOR P-VALUES ADJUSTED BY GENOMIC CONTROL

	<1e-04	<0.001	<0.01	<0.025	<0.05	<0.1	<1
p-value	68	458	4207	10529	20868	41071	404980
q-value	0	0	0	0	0	0	405003
Local FDR	0	0	0	0	0	0	392897

As can be seen from Figure 6-14, the number of features that are considered significant are small, maintained a small increase in feature as the threshold is increased. Given the q-value correction method use the number of significance results and provide a probability of observing a false positive with that set, the threshold must be maintained under 0.5, which indicates an even probability of observing a false positive.

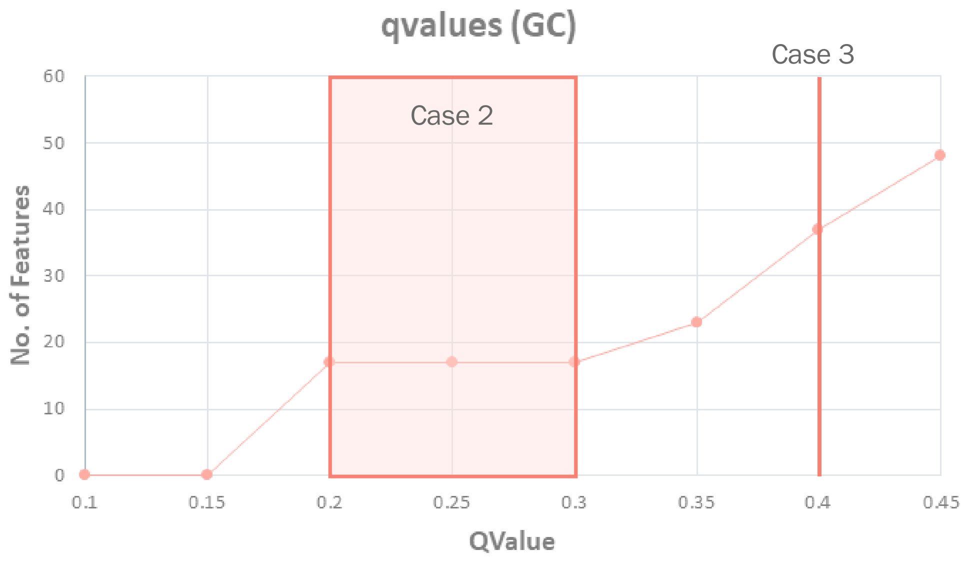


FIGURE 6-14: THRESHOLD MEASURES FOR CASE 2 AND CASE 3

Therefore, as demonstrated in Figure 6-14, the chosen thresholds explore the increase in that threshold 0.3 and 0.4, with varying feature set sizes, it provides an indication of the potential of the q-value combined with the GC method to produce interaction sets unhindered by small thresholds. Case 2 threshold, 0.3, was chosen as a lower threshold as from 0.2 to 0.3 is the first instance in which features surpass a significance threshold; within this case either 0.2 or 0.3 could have been used with neither affecting the outcome of feature set. Case 3 threshold, 0.4, was chosen given its’ increase in feature set size with consideration to the 0.5 problematic area.

6.4.2 CASE 4

Continuing from this, the thresholds chosen for the remaining method should reflect a comparable feature set size. As case 4 represents the most lenient method chosen for benchmarking, the feature set size is expected to be the largest. Table 6.4 shows the multiple testing analysis and the resulting features that qualify for each bin. Visible in the table is the large number of features indicated to have extremely significant relationship with the phenotype, Breast Cancer. This is a good example of the importance of using multiple testing while analysing large datasets as this bin is likely heavily populated with false positive results. Therefore, referring to the q values proposes more realistic significance estimates.

Figure 6-16 provides the increase in feature size set as the threshold is increased. Notable in this method is the feature set size are much larger regardless of using method q-value, evidencing the strict adjustments of GC.

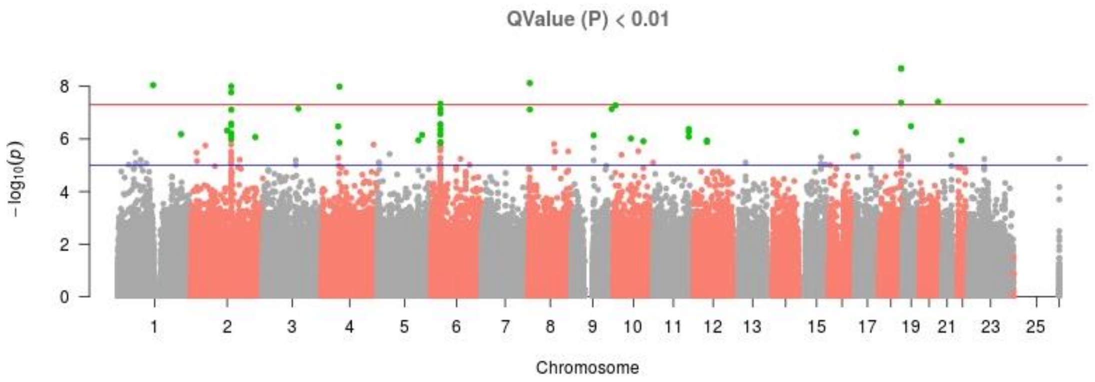


Figure 6-15: Manhattan plot showing the feature set SNPs highlighted in green when applying the threshold Q(α) < 0.01 using standard case-control approach (N =48).

Table 6.4: FDR values for p-values

	<1e-04	<0.001	<0.01	<0.025	<0.05	<0.1	<1
p-value	506	2529	12901	25004	41142	68019	404982
q-value	0	7	48	91	243	1157	405003
Local FDR	0	6	17	51	129	591	404924

Included in this figure are the number of features of case 2 and case 3, highlighted in yellow, while the threshold chosen for case 4 is represented by the pink line. This method offers an opportunity to observe the outcome interactions of the feature set given more leniency in the feature selection process. An important note at this stage is to point out that p-values which are used within q-value and adjusted for genomic control are derived from the same association analysis, therefore the features selected by case 2-4 will represent a crossover number of the same features e.g. case 3 is superset of case 2.

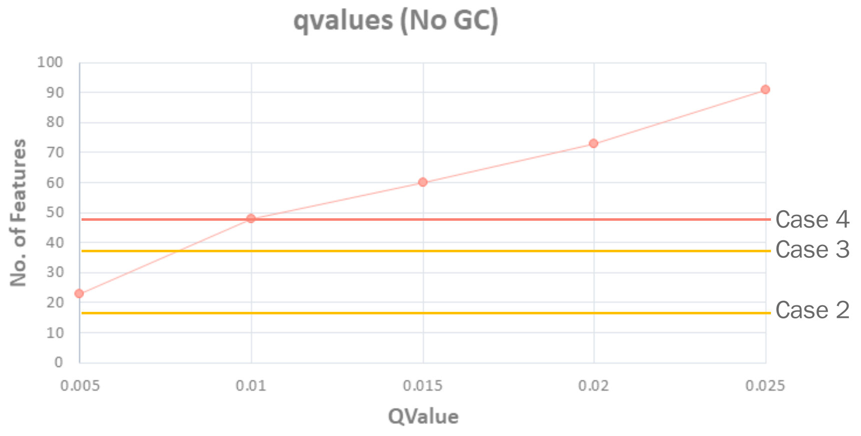


Figure 6-16: Threshold measure for case 4 in comparison to case 2 and 3

6.4.3 CASE 1

Case 1 refers to the novel methodology and is the last method to have a feature selection threshold applied. In order to decide the threshold for the RaSaR method, it was important to concentrate on filtering a comparable number of features to cases 2-4 which ideally would result in an amount within the feature sizes already outlined. The proposed methodology of this research used standard deviation and mean to produce a unique threshold that takes into consideration the fluctuation of values across random cohorts.

Figure 6-17 shows the selected feature set for case 1 in green using -$\log_{10}$ p-values adjusted by GC. Visible is the distinct difference in the -$\log_{10}$ p-values scale which indicates lower association throughout the SNP set due to the use of the mean value, however the differences visible between Figure 6-12, Figure 6-13 and Figure 6-15 to Figure 6-17 show that the overall structure of the data points has changed, indicating that many of the SNPs with higher association in Cases 2-4 had a fluctuating presence depending on the cohort being analysed.

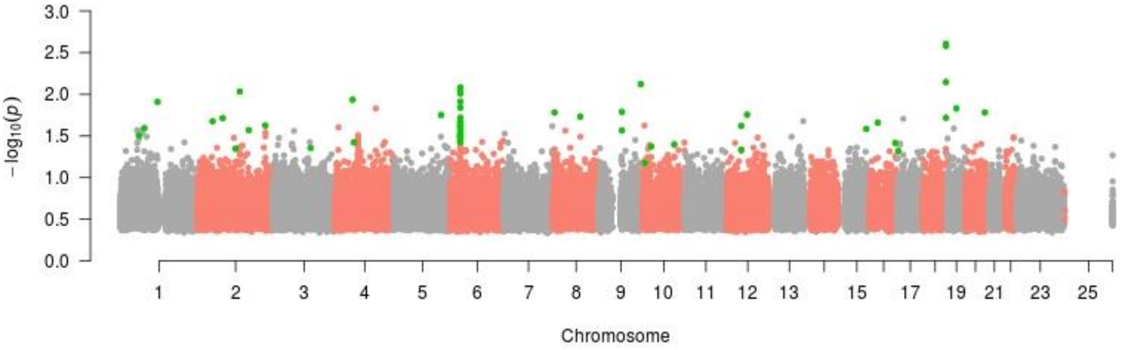


FIGURE 6-17: MANHATTAN PLOT SHOWING THE FEATURE SET SNPS HIGHLIGHTED IN GREEN WHEN APPLYING THE THRESHOLDS $\mu < 0.05$ AND $\sigma < 0.025$ USING RASAR. (N = 41)

By using the standard deviation values alongside the mean, this approach also considers any SNPs that have an inflated mean due to anomaly results will be excluded based on standard deviation value. The purpose of this feature selection method is to produce a feature set that includes SNPs that show significance but more importantly are consistently significant regardless of the subjects included in the cohort.

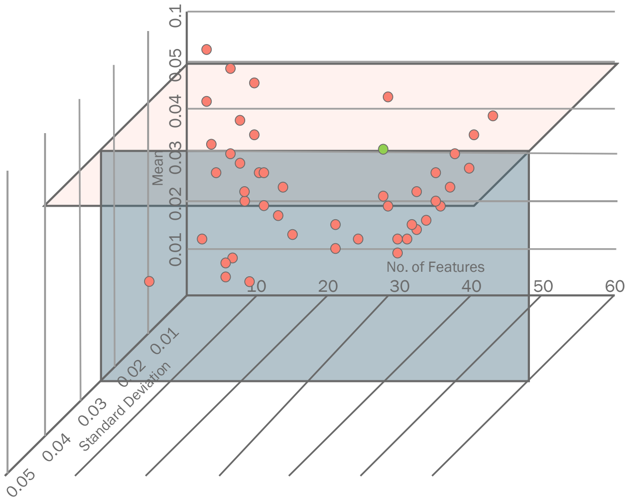


Figure 6-18: Systematic feature exploration using standard deviation and mean thresholds

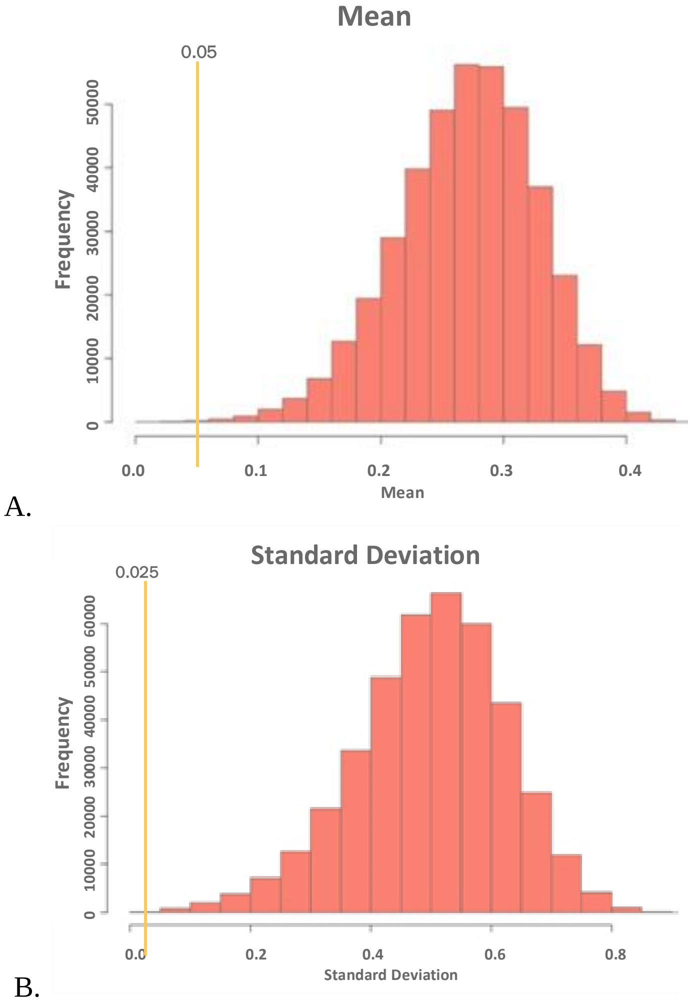


Figure 6-19: Histograms generated from association analysis using random sampling Regularisation showing A. μ and B. σ, with threshold exclusion measures.

Using a systematic approach of increasing the mean value and using a selection of standard deviation thresholds that increase by 0.05, but do not exceed the mean value for each test, the feature set sizes were explored. One main feature of the RaSaR feature selection method is to emphasise the use of standard deviation to filter SNPs that exhibit fluctuating behaviour dependent on the cohort sample that was used, therefore restricting the standard deviation is more important than restricting the mean. Figure 6-18 visualises these tests and shows the different feature set sizes dependent on the threshold explored. Highlighted in green is the feature selection threshold chosen which exists at the junction of the mean threhold 0.05 and the standard deviation threshold of 0.025. Histograms in Figure 6-19 indicate the thresholds that are used for each extracted feature. A lenient threshold of mean, μ < 0.01 and standard deviation, σ < 0.03.

Figure 6-20 provides a comparable graph in which the SNPs selected via the RaSaR technique are outlined using the GC adjusted p-values obtained from Case 2-4. In comparison with Figure 6-12, Figure 6-13 and Figure 6-15, the RaSaR method excludes SNPs whose presence among the 9 GC adjusted p-values obtained during association analysis do not adhere to the set consistency threshold, which in this case is, σ <0.025.

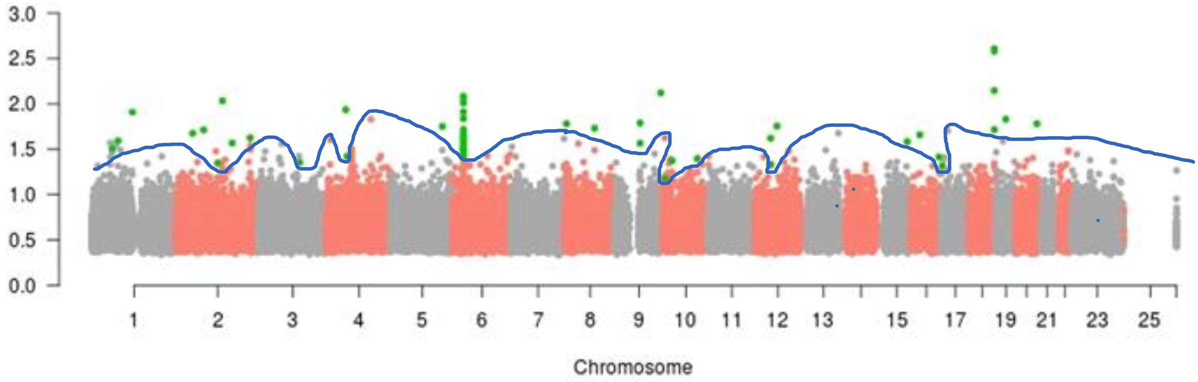


FIGURE 6-20: MANHATTAN PLOT SHOWING THE FEATURE SET SNPS HIGHLIGHTED IN GREEN WHEN APPLYING THE THRESHOLDS μ < 0.01 AND σ < 0.03 USING RASAR APPROACH WITH CASE 2-4 GC ADJUSTED P-VALUE DATA. (N = 41)

Outlined feature set numbers from Cases 1-4 are displayed in Table 6.5. Lists of all features and their significance values can be found in Appendix A . These features were extracted from the main dataset and used in the next stage of Epistasis Analysis.

TABLE 6.5: SELECTED FEATURES OF EACH THRESHOLD MEASURE

Case ID	Threshold	No. of Features
1	RaSaR	41
2	q-value (p-values) (0.01)	48
3	q-Value (GC adjusted) (0.3)	17
4	q-Value (GC adjusted) (0.4)	37

6.5 Stage 5: Epistasis Analysis

Features selected using the previous process are inputted into software LAMPlink, using a dominant model. Linkage Disequilibrium pruning is used to remove redundant features that exist as relationship as a result of high LD. Given that association analysis has already been performed with the outlined features being selected as result of significance measures during this process, the threshold used in this stage also has to be more conservative as the false positive presence is already established. Given this, a threshold, $\alpha < 0.005$ has been applied.

Table 6.6: Combinations outlined through LAMPlink.

Combination	PPP Conf.	C1	C2	C3	C4
rs4602520-rs6910087-rs7246472	0.833	O	O		
9q21.13-9q21.13	0.25	O			
rs4602520-rs4144827	0.5	O			
1q41-rs3924215	0.75		O		O
rs6911024-rs12170250	0.75		O		
rs6852865-rs4602520	1		O		
rs6852865-rs4602520-rs6910087-rs7246472	0.375		O		
rs6852865-rs6910087-rs7246472	0.667		O		
rs3924215-rs6011609	0.5			O	
1p12-rs6011609	0.5			O	
rs4602520-rs6911024-rs7246472	0.5				O
rs6911024-rs7246472	0.5				O
rs4602520-rs7246472	0.5	O	O		O
rs4602520-rs6911024	0.5				O

Table 6.6 outlines the interaction results from LAMPlink along with the case number of thresholds that detected the interaction. All outputs can be found in Appendix B . As can be seen, there are a few interactions that have been outlined by multiple cases and one interaction that shows a low PPP confidence value to evidence the suggestive effect of the PPP. Normally within this method an interaction with PPP confidence value of < 0.5 would be discarded, however in order to show the effects of a combination with low PPP confidence, combination 9q21.13-9q21.13 highlighted in red, has been evaluated in the next stage. Adjusted p-values are used to produce petal plots that show the significance value of each singular feature in relation to the overall interaction score. Table 6.7 shows the PPP results of interaction rs4602520; rs6910087; rs7246472 values for both the singular and combination outlined using case 1 and 2 methods. Figure 6-21 shows the petal plot for interaction rs4602520; rs6910087; rs7246472 where one of the singular SNPs shows a greater p-value (red) than the combined interaction p-value (orange) which reduces the confidence of the interaction.

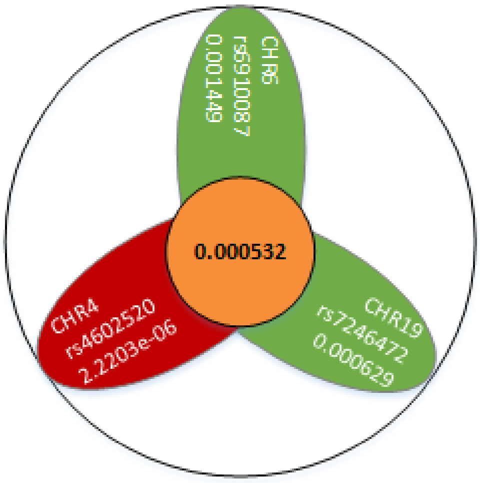


FIGURE 6-21: PETAL PLOT FOR INTERACTION RS4602520; RS6910087; RS7246472 [A]

[A] Petals in red present a lower p-value than the interaction centre, petals in green present a higher p-value than the interaction centre. Therefore, the interaction PPP confidence value is lowered as represented by Table 6.7.

TABLE 6.7: EXAMPLE OF INTERACTION OUTCOME VALUES

	Combination	Raw p-value	Adjusted p-value
Case 1	rs4602520; rs6910087; rs7246472	1.8291e-06	0.00053227
	rs4602520	7.6298e-09	2.2203e-06
	rs6910087	4.9796e-06	0.0014491
	rs7246472	2.1604e-06	0.00062868

$$\textbf{PPP Confidence} = \left(\frac{2}{3}\times 0.5\right)+\left(\frac{3-3}{3}\times 0.5\right)=0.83\dot{3}$$

Secondly, the location of the SNPs is considered to focus on interactions that occur cross-chromosome. Figure 6-21 (Refer to Appendix C for interaction values), shows the locations of the SNPs included in the interaction are dispersed across three different chromosomes that could be an indication that functionality of these SNPs interferes or cooperates with one another, leading to the focus phenotypic expression. A few of the combinations outlined in Figure 6-7 contain SNPs that are located in the same chromosome, however these are extended into the next process to investigate the outcome and implications of these interactions. All PPP Confidence workings can be found in Appendix C

6.6 Stage 6: Inference Analysis

Combinations including singular, two-way and three-way along with all possible combination states (excluding combinations that do not contain more than 10 values in one cell of case-control based contingency table) are analysed to expand and explore the relationships outlined during the previous stage.

6.6.1 Analysis of Breast Cancer training set

As LAMPlink is used to outline potential relationships between variants, further analysis is used to investigate those relationships with the main focus on identifying causal variant combinations that effect either a larger population and/or present an increased association between the interaction and the phenotype.

All interactions have been extended to include all possible relationships between outlined variant within the combination e.g.

$$B \subseteq A$$
$$rs4602520 - rs7246472 \quad \subseteq \quad rs4602520 - rs6910087 - rs7246472$$

Further to this, also considered is every state combination within the relationship that is present in the sample population with each cell of a 2x2 contingency cell contained a frequency >10, Figure 6-22 demonstrates 2 states that were analysed from combination rs4602520-rs6910087-rs7246472.

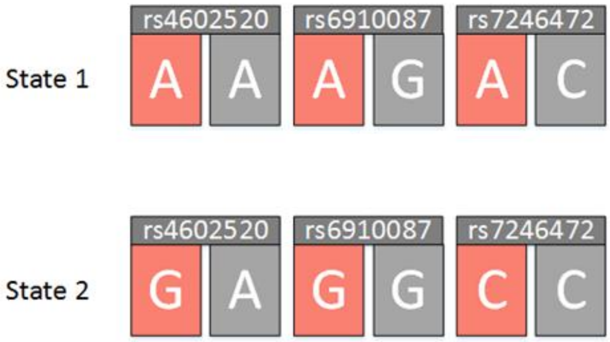


Figure 6-22: Representation of analysed allelic states 1 and 2. Allele 1 is represented in pink while allele 2 is in grey

In order to focus the results in this section, Table 6.8 to Table 6.9 provides the results from the analysis that outline any interaction combinations that produced a significant result of p-value < 0.05. All results can be found in 0with contingency table values in Appendix F . Each analysis is performed using Chi-Squared (X^2), Fisher's Exact Test (F) and Odd's Ratio (OR) risk, p-value (OR P) and upper (>CI) and lower (CI<) confidence intervals.

TABLE 6.8: STATISTICALLY SIGNIFICANT COMBINATION STATES FROM TRAINING DATA

ID	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
1	rs4602520	Dominant	0.000	0.000	1.253	1.370	1.498	0.000
	rs6910087	Dominant	0.000	0.000	1.152	1.241	1.337	0.000
	rs7246472	Dominant	0.000	0.000	1.178	1.281	1.394	0.000
	rs4602520, rs6910087	Dominant	0.000	0.000	1.296	1.531	1.809	0.000
		AAAA			1.091	1.388	1.766	0.025
		AAAG			1.036	1.125	1.221	0.019
		AAGG			0.707	0.756	0.808	0.000
		GAAG			1.273	1.525	1.828	0.000
		GAGG			1.160	1.286	1.426	0.000
	rs4602520, rs7246472	Dominant	0.001	0.001	1.257	1.544	1.896	0.001
		AAAA			1.254	1.877	2.808	0.010
		AAAC			1.081	1.185	1.298	0.002
		AACC			0.684	0.733	0.786	0.000
		GAAC			1.195	1.481	1.835	0.003
		GACC			1.187	1.310	1.446	0.000
	rs6910087, rs7246472	Dominant	0.000	0.000	1.174	1.277	1.390	0.000
		AAAC			1.138	1.978	3.438	0.042
		AGAC			1.181	1.400	1.660	0.001
		AGCC			1.041	1.131	1.229	0.015
		GGAA			1.136	1.723	2.615	0.032
		GGAC			1.048	1.154	1.271	0.015
		GGCC			0.730	0.780	0.833	0.000
	rs4602520, rs6910087, rs7246472	Dominant	0.000	0.000	2.150	3.335	5.174	0.000
		AAGGAA			1.107	1.712	2.648	0.042
		AAGGAC			1.036	1.149	1.274	0.027
		AAGGCC			0.681	0.726	0.773	0.000
		GAAGAC			1.971	3.196	5.181	0.000
		GAAGCC			1.041	1.269	1.548	0.048
		GAGGCC			1.166	1.303	1.458	0.000
2	9q21.13	Dominant	0.000	0.000	0.711	0.775	0.846	0.000
	9q21.13	Dominant	0.000	0.000	0.692	0.762	0.839	0.000
	9q21.13 9q21.13	Dominant	0.000	0.000	0.688	0.758	0.835	0.000
		AGCA			0.701	0.774	0.855	0.000
		GGAA			1.179	1.285	1.401	0.000
3	rs4144827	Dominant	0.000	0.000	1.154	1.296	1.457	0.000
	rs4144827 rs4602520	Dominant	0.000	0.000	1.443	1.801	2.248	0.000
		AAAA			0.692	0.744	0.801	0.000
		AAGA			1.151	1.268	1.397	0.000
		GAAA			1.052	1.170	1.302	0.015
		GAGA			1.502	1.906	2.419	0.000
4	1q:44	Dominant	0.000	0.000	0.710	0.774	0.844	0.000
	rs3924215	Dominant	0.000	0.000	0.724	0.779	0.838	0.000
	1q:44 rs3924215	Dominant	0.000	0.000	0.495	0.586	0.694	0.000
		AAAA			1.220	1.303	1.391	0.000
		GAGA			0.485	0.581	0.695	0.000

Table 6.9: Statistically significant combination states from Training Data cont. (1)

ID	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
5	rs6911024	Dominant	**0.000**	0.000	1.157	1.247	1.343	0.000
	rs12170250	Dominant	**0.000**	0.000	1.459	1.779	2.170	0.000
	rs6911024 rs12170250	Dominant	**0.000**	0.000	2.028	2.884	4.101	0.000
		AAGG			0.733	0.789	0.849	0.000
		GAAG			1.868	2.716	3.950	0.000
6	rs6852865	Dominant	**0.000**	0.000	1.199	1.314	1.438	0.000
	rs4602520	Dominant	**0.000**	0.000	1.253	1.370	1.497	0.000
	rs6852865 rs4602520	Dominant	**0.000**	0.000	1.230	1.357	1.496	0.000
		AGGA			1.236	1.368	0.514	0.000
		GGAA			0.689	0.749	0.815	0.000
7	rs6852865	Dominant	**0.000**	0.000	1.200	1.314	1.439	0.000
	rs6852865 rs4602520 rs6910087 rs7246472	Dominant	**0.000**	0.000	2.303	3.832	6.376	0.000
		AGGAGGCC			1.140	1.292	1.464	0.001
		GGAAGGAC			1.035	1.148	1.275	0.029
		GGAAGGCC			0.685	0.730	0.778	0.000
		GGGAGGCC			1.052	1.328	1.676	0.045
8	rs6852865 rs6910087 rs7246472	Dominant	**0.000**	**0.000**	**2.180**	**3.422**	**5.370**	**0.000**
		AGAGAC			2.268	3.938	6.839	0.000
		AGGGCC			1.097	1.230	1.378	0.003
		GGGGAC			1.031	1.143	1.267	0.033
		GGGGCC			0.699	0.744	0.793	0.000
9	rs3924215	Dominant	**0.000**	0.000	0.723	0.778	0.838	0.000
	rs6011609	Dominant	**0.000**	0.000	0.411	0.504	0.619	0.000
	rs3924215 rs6011609	Dominant	**0.000**	0.000	0.229	0.354	0.546	0.000
		AAAG			0.448	0.566	0.716	0.000
		AAGG			1.249	1.342	1.442	0.000
10	1p12	Dominant	**0.000**	0.000	0.402	0.495	0.610	0.000
	1p12 rs6011609	Dominant	**0.001**	0.001	0.034	0.115	0.396	0.004
		AAAG			0.438	0.541	0.668	0.000
		AAGG			1.690	1.963	2.281	0.000
		GAGG			0.427	0.529	0.657	0.000
11	rs6911024	Dominant	**0.000**	0.000	1.167	1.258	1.356	0.000
	rs6911024 rs7246472	Dominant	**0.000**	0.000	1.295	1.522	1.790	0.000
		AAAA			0.348	0.534	0.821	0.016
		AAAC			0.794	0.874	0.963	0.022
		AACC			1.210	1.293	1.381	0.000
		GAAC			0.592	0.703	0.834	0.001
		GACC			0.801	0.871	0.947	0.007
	rs4602520 s6911024	Dominant	**0.000**	0.000	1.311	1.550	1.834	0.000
		AAAA			1.250	1.337	1.429	0.000
		AAGA			0.805	0.875	0.951	0.008
		AAGG			0.574	0.731	0.930	0.032
		GAAA			0.702	0.778	0.862	0.000
		GAGA			0.542	0.650	0.781	0.000

6.6.1 ANALYSIS OF BREAST CANCER TESTING SET

Having conducted the inference analysis using the training data, the next process is to analyse the significant combinations outlined using a separate dataset. The purpose of this process is to analyse whether the significant combinations outlined retain significance using an unused set of data which increases confidence in a true positive association. Table 6.10 to Table 6.11 provide the results from the testing set using all significant relationships outlined in Training. Combinations with NA values were omitted as a result of a cell frequency < 10 in a 2x2 contingency table. Combination outlined in grey indicate associations that retained significance using the testing dataset and will therefore be carried forward.

Table 6.10: Statistical analysis of outlined significant variants using Testing Dataset

ID	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
1	rs4602520	Dominant	0.672	0.677	0.849	1.051	1.302	0.672
	rs6910087	Dominant	0.016	0.017	1.073	1.242	1.439	0.015
	rs7246472	Dominant	0.032	0.034	1.054	1.243	1.466	0.030
	rs4602520 rs6910087	Dominant	0.263	0.296	0.902	1.266	1.778	0.253
		AAAA			0.814	1.305	2.093	0.354
		AAAG			1.015	1.192	1.400	0.073
		AAGG			0.742	0.846	0.965	0.037
		GAAG			0.931	1.349	1.956	0.185
		GAGG			0.769	0.938	1.145	0.600
	rs4602520 rs7246472	Dominant	0.002	0.002	0.317	0.472	0.704	0.002
		AAAA			0.509	1.170	2.687	0.756
		AAAC			0.909	1.088	1.302	0.439
		AACC			0.801	0.919	1.054	0.312
		GAAC			1.283	1.925	2.887	0.008
		GACC			0.693	0.842	1.024	0.148
	rs6910087 rs7246472	Dominant	0.045	0.055	1.068	1.453	1.976	0.046
		AAAC			0.247	0.648	1.702	0.460
		AGAC			1.145	1.597	2.229	0.021
		AGCC			0.961	1.132	1.333	0.213
		GGAA			0.654	1.640	4.115	0.376
		GGAC			0.933	1.127	1.360	0.299
		GGCC			0.691	0.787	0.897	0.003
	rs4602520 rs6910087 rs7246472	Dominant	0.068	0.076	1.129	2.355	4.913	0.055
		AAGGAA			0.485	1.274	3.345	0.680
		AAGGAC			0.828	1.014	1.241	0.911
		AAGGCC			0.747	0.848	0.963	0.032
		GAAGAC			1.259	2.931	6.824	0.036
		GAAGCC			0.701	1.067	1.626	0.799
		GAGGCC			0.647	0.804	0.999	0.098
2	9q21.13	Dominant	0.438	0.453	0.776	0.920	1.091	0.423
	9q21.13	Dominant	0.731	0.740	0.792	0.956	1.155	0.697
	9q21.13 9q21.13	Dominant	0.730	0.743	0.792	0.957	1.156	0.700
		AGCA			0.797	0.969	1.179	0.793
		GGAA			0.917	1.087	1.288	0.421

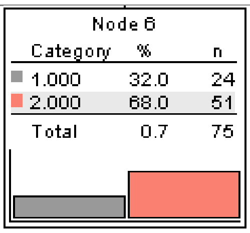
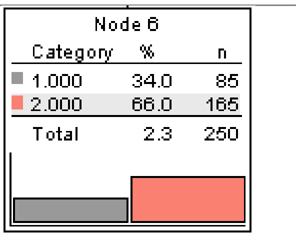
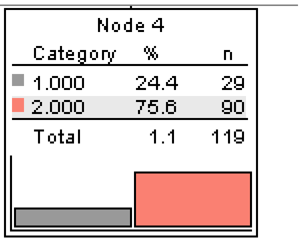
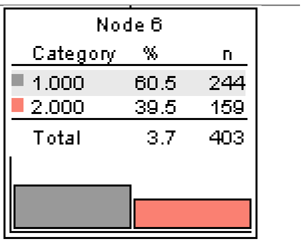
TABLE 6.11: STATISTICAL ANALYSIS OF OUTLINED SIGNIFICANT VARIANTS USING TESTING DATASET CONT. (1)

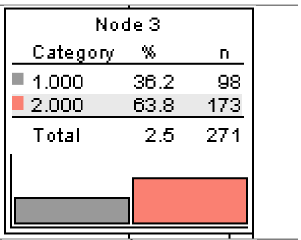
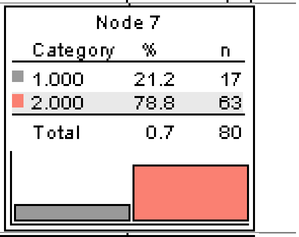
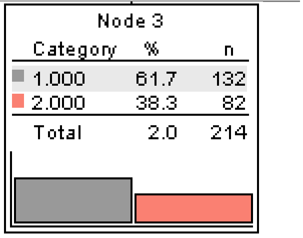
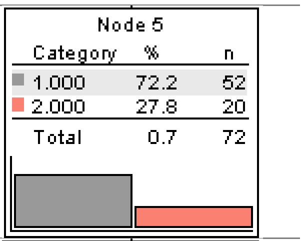
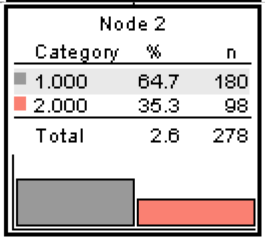
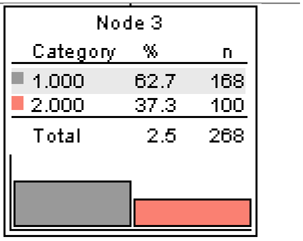
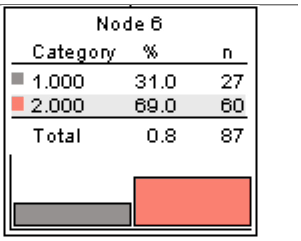
ID	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
3	rs4144827	Dominant	0.063	0.063	1.032	1.252	1.518	0.056
	rs4144827 rs4602520	Dominant	0.009	0.013	1.290	2.042	3.232	0.011
		AAAA			0.795	0.919	1.062	0.337
		AAGA			0.737	0.891	1.077	0.319
		GAAA			0.860	1.065	1.319	0.625
		GAGA			1.290	2.151	3.588	0.014
4	1q:44	Dominant	0.383	0.408	0.772	0.914	1.082	0.379
	rs3924215	Dominant	0.633	0.649	0.886	1.055	1.257	0.611
	1q:44 rs3924215	Dominant	0.897	0.990	0.675	1.038	1.597	0.886
		AAAA			0.898	1.031	1.183	0.718
		GAGA			0.534	0.840	1.321	0.526
5	rs6911024	Dominant	0.018	0.020	1.067	1.237	1.433	0.018
	rs12170250	Dominant	0.328	0.336	0.790	0.914	1.058	0.314
	rs6911024 rs12170250	Dominant	0.690	0.698	0.829	1.078	1.402	0.640
		AAGG			0.415	0.723	1.260	0.337
		GAAG			NA	NA	NA	NA
7	rs6852865	Dominant	0.550	0.581	0.893	1.068	1.278	0.544
	rs6852865 rs4602520 rs6910087 rs7246472	Dominant	0.175	0.240	0.876	1.979	4.467	0.168
		AGGAGGCC			0.644	0.820	1.042	0.173
		GGAAGGAC			0.829	1.078	1.402	0.640
		GGAAGGCC			0.746	0.847	0.961	0.031
		GGGAGGCC			0.456	0.754	1.246	0.355
8	rs6852865 rs6910087 rs7246472	Dominant	0.099	0.107	1.059	2.222	4.665	0.076
		AGAGAC			1.358	3.435	8.686	0.029
		AGGGCC			0.688	0.860	1.075	0.267
		GGGGAC			0.840	1.026	1.253	0.835
		GGGGCC			0.732	0.831	0.944	0.017
9	rs3924215	Dominant	0.633	0.649	0.886	1.055	1.257	0.611
	rs6011609	Dominant	0.232	0.243	0.495	0.735	1.091	0.200
	rs3924215 rs6011609	Dominant	0.468	0.518	0.287	0.646	1.458	0.377
		AAAG			0.489	0.767	1.204	0.333
		AAGG			0.985	1.137	1.312	0.140
10	1p12	Dominant	0.000	0.001	0.252	0.394	0.615	0.001
	1p12 rs6011609	Dominant	NA	NA	NA	NA	NA	NA
		AAAG			0.532	0.796	1.190	0.351
		AAGG			1.316	1.774	2.390	0.002
		GAGG			0.269	0.422	0.662	0.002
11	rs6911024	Dominant	0.018	0.019	1.068	1.238	1.434	0.017
	rs6911024 rs7246472	Dominant	0.068	0.070	1.047	1.422	1.930	0.058
		AAAA			0.243	0.610	1.530	0.377
		AAAC			0.736	0.888	1.073	0.302
		AACC			1.110	1.265	1.442	0.003
		GAAC			0.462	0.643	0.895	0.028
		GACC			0.749	0.883	1.040	0.212
	rs4602520 rs6911024	Dominant	0.220	0.249	0.922	1.297	1.826	0.211
		AAAA			1.032	1.177	1.342	0.041
		AAGA			0.721	0.847	0.995	0.090
		AAGG			0.478	0.767	1.230	0.355
		GAAA			0.874	1.067	1.302	0.595
		GAGA			0.495	0.719	1.046	0.148

6.6.2 Decision Trees

During the previous analysis, every combination state of SNP was analysed. However, it is important to consider the allelic expression in combinatorial states that omit 1 or more states to achieve better penetrance and/or incidence that may present more significantly than the previously outlined interaction states. Further to this, decision trees are able to outline cohort sets that appear distinct in association. Refer to section 3.3.3 and section 3.4 for more information about the decision to use CHAID. During this stage, each combination state outlined in Table 6.6 is used to produce feature sets for further analysis using a CHAID method (See 3.4 for discussion),(See for full decision tree results Appendix G). Table 6.12 provides a summarisation of the most significant nodes that were outlined from these analyses. Note: Results present the variant, allele (A1/A2) and corresponding nucleotide (A/G/C/T).

Table 6.12: CHAID analysis node results

ID	Appendix Reference	Combination	Node Data
1	Combination 1	rs4602520 (A1) (A) rs7246472 (A1) (A) rs7246472 (A2) (A)	 Node 6 Category % n ■ 1.000 32.0 24 🟥 2.000 68.0 51 Total 0.7 75
3	Combination 3	rs4144827 (A1) (G) rs4602520 (A1) (G)	 Node 6 Category % n ■ 1.000 34.0 85 🟥 2.000 66.0 165 Total 2.3 250
4	Combination 4	1q14 (A1) (G) rs3924215 (A1) (A)	 Node 4 Category % n ■ 1.000 24.4 29 🟥 2.000 75.6 90 Total 1.1 119
5	Combination 5	rs12170250 (A1) (G) rs6911024 (A1) (G)	 Node 6 Category % n ■ 1.000 60.5 244 🟥 2.000 39.5 159 Total 3.7 403

8a	Combination 8	rs6852865 (A1) (A) rs7246472 (A1) (A)	Node 3 Category % n ■ 1.000 36.2 98 🟥 2.000 63.8 173 Total 2.5 271
8b	Combination 8	rs6852865 (A1) (A) rs7246472 (A1) (A) rs6910087 (A1) (A)	Node 7 Category % n ■ 1.000 21.2 17 🟥 2.000 78.8 63 Total 0.7 80
9a	Combination 9	rs3924215 (A1) (A) rs6011609 (A1) (A)	Node 3 Category % n ■ 1.000 61.7 132 🟥 2.000 38.3 82 Total 2.0 214
9b	Combination 9	rs3924215 (A1) (G) rs6011609 (A1) (A)	Node 5 Category % n ■ 1.000 72.2 52 🟥 2.000 27.8 20 Total 0.7 72
10a	Combination 10	1p12 (A1) (G)	Node 2 Category % n ■ 1.000 64.7 180 🟥 2.000 35.3 98 Total 2.6 278
10b	Combination 10	1p12 (A1) (A) rs6011609 (A1) (A)	Node 3 Category % n ■ 1.000 62.7 168 🟥 2.000 37.3 100 Total 2.5 268
11	Combination 11	rs7246472 (A1) (A) rs7246472 (A2) (A)	Node 6 Category % n ■ 1.000 31.0 27 🟥 2.000 69.0 60 Total 0.8 87

Given that the combinations outlined above vary slightly from those outlined during the inference analysis, OR and p-values were generated based on the testing set to confirm the cohorts outlined. Full decision trees can be found in Appendix G. Using the information from the CHAID analysis, combinations outlined were analysed using the testing set. Table 6.13 outlines the results of the remaining combinations, where '?' refers to combination inputs that are not factored in and do not affect the outcome. Results outlined in grey resulted in a p-value < 0.05 using the testing dataset and are carried forward.

TABLE 6.13: CHAID OUTPUT TESTING

Interaction	Model	OR	P
1	A?AA	1.170	0.756
3	**G?G?**	**2.151**	**0.014**
4	G?A?	0.900	0.330
5	G?G?	1.080	0.632
8a	A?A?	1.099	0.127
8b	**A?A?A?**	**3.435**	**0.029**
9a	A?A?	0.768	0.333
9b	G?A?	0.554	0.254
10a	**G?**	**0.396**	**0.000**
10b	A?A?	0.796	0.351
11	AA	1.870	0.174

6.6.3 COMBINATION RELEVANCE

Using the combinations outlined from Table 6.10 to **Error! Reference source not found.** and Table 6.13, further analysis is performed to consider their context to real-world information. Using penetrance and incidence to map the extent of their effect, a threshold of >60% is used to outlined results that present more effective significance to breast cancer. Table 6.15 presents data to outline the penetrance and incidence of each combination. Results outlined in grey surpassed the threshold of 60% penetrance.

Penetrance: How many subjects have been affected by the phenotype that also carry the genomic interaction state?

$$P(Phenotype \mid Genotype) = \frac{a}{a+b}$$

Incidence: What percentage of the sample population carry this genomic interaction state?

$$P(Phenotype \mid SamplePopulation) = \frac{a+b}{n}$$

Risk: How strongly associated is the presence of the genomic interaction state with the presence of the phenotypic state?

$$\frac{odds\ of\ disease\, among\, exposed}{odds\ of\ disease\, among\, unexposed} = \frac{ad}{bc}$$

TABLE 6.14: REFERENCE OF MEASURE FOR REAL-WORLD RELEVANCE

	Case	Controls	Total
Exposed	a	b	a+b
Not Exposed	c	d	c+d
Total	a+c	b+d	n

TABLE 6.15: PENETRANCE AND INCIDENCE OF RESULT COMBINATIONS

Interaction	Model	OR	Penetrance	Incidence (%)	P
rs4144827 (A1) (G) - rs4602520 (A1) (G)	G?G?	2.151	60.0	1.90	0.0140
rs6852865 (A1) (A) - rs7246472 (A1) (A) – rs6910087 (A1) (A)	A?A?A?	3.435	56.2	0.70	0.0290
1p12 (A1) (G)	G?	0.396	60.3	2.44	0.0000
rs6910087	Dominant	1.243	56.4	25.0	0.0151
rs7246472	Dominant	1.243	54.0	18.2	0.0302
rs4602520-rs6910087	AAAG	1.192	56.0	19.4	0.0726
	AAGG	0.846	50.8	63.1	0.0369
rs4602520-rs7246472	Dominant	0.472	51.8	97.0	0.0020
	GAAC	1.925	67.5	3.00	0.0079
rs6910087-rs7246472	Dominant	1.453	61.1	5.00	0.0456
	AGAC	1.597	63.3	4.00	0.0208
	GGCC	0.787	50.0	61.5	0.0026
rs4602520, rs6910087, rs7246472	AAGGCC	0.848	50.4	51.6	0.0323
	GAAGAC	2.931	76.2	0.80	0.0364
rs4144827	Dominant	1.252	57.2	12.5	0.0555
rs4144827-rs4602520	Dominant	2.042	68.9	2.30	0.0106
	GAGA	2.151	70.0	2.00	0.0138
rs6852865-rs4602520-rs6910087-rs7246472	GGAAGGCC	0.847	50.3	50.5	0.0307
rs6852865-rs6910087-rs7246472	AGAGAC	3.435	79.0	0.70	0.0287
	GGGGCC	0.831	50.1	52.2	0.0165
1p12	Dominant	0.394	69.7	2.50	0.0006
1p12-rs6011609	AAGG	1.774	52.6	95.0	0.0016
	GAGG	0.422	47.0	2.40	0.0016
rs6911024	Dominant	1.237	56.4	25.0	0.0171
rs6911024,rs7246472	AACC	1.265	56.0	38.5	0.0030
	GAAC	0.643	52.0	96.0	0.0279
rs4602520,rs6911024	AAAA	1.177	55.0	37.0	0.0414

Table 6.16 presents the statistical characteristics of the top variants and interactions that were identified during this research.

TABLE 6.16: STATISTICAL CHARACTERISTICS OF TOP INTERACTIONS

ID	Interaction	State	OR	RR	Penetrance (%)	Incidence (%)	P
1	**1p12**	Dominant	0.394	0.578	69.7	2.50	0.0006
2	**1p12**	G?	0.396	0.397	60.3	2.44	0.0000
3	**rs4144827-rs4602520**	Dominant	2.041	1.324	68.9	2.30	0.0106
4	**rs4144827 - rs4602520**	G?G?	2.151	2.141	60.0	1.90	0.0140
5	**rs4144827-rs4602520**	GAGA	2.151	1.345	70.0	2.00	0.0138
6	**rs4602520-rs7246472**	GAAC	1.925	1.119	67.5	3.00	0.0079
7	**rs4602520- rs6910087-rs7246472**	GAAGAC	2.931	1.460	76.2	0.80	0.0364
8	**rs6910087-rs7246472**	Dominant	1.453	1.176	61.1	5.00	0.0456
9	**rs6910087-rs7246472**	AGAC	1.597	1.219	63.3	4.00	0.0208
10	**rs6852865-rs6910087-rs7246472**	AGAGAC	3.435	1.513	79.0	0.70	0.0287

Using the OR value from each of the combinations produced from the testing dataset, Figure 6-23 demonstrates the deviation from OR = 1. This provides information as to whether the interaction combination presents a risk or protective factor and how big that factor is. Additionally, the confidence intervals show the precision of the OR.

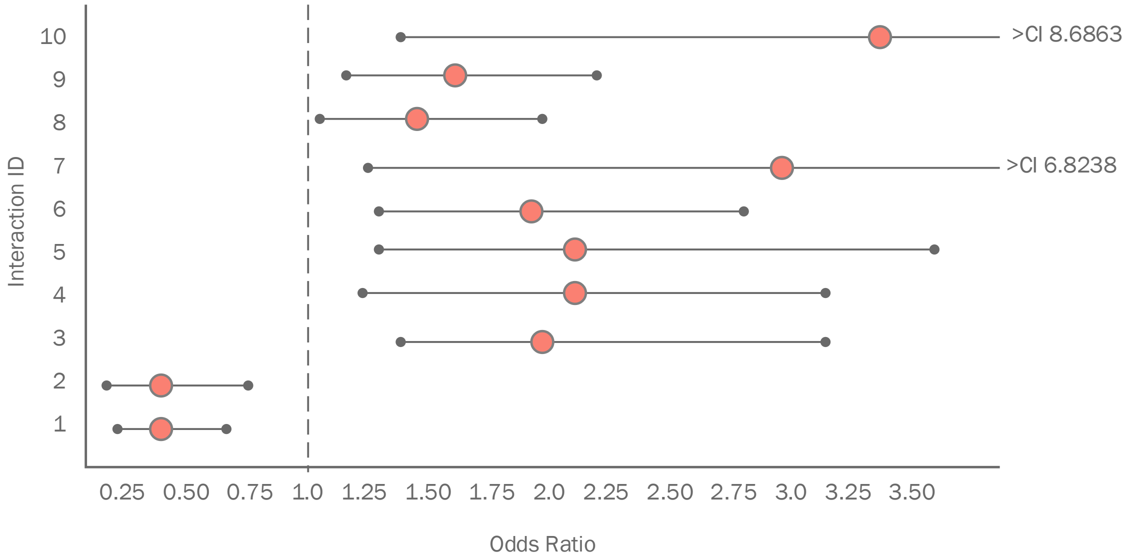


FIGURE 6-23: ODDS RATIO PLOT FOR TOP RESULTS

As presented from Figure 6-23, most interactions indicate a risk factor, with only one set indicating a protective factor (1p12 (Dominant) (G?G?)). Here it is possible to compare the statistically available risk score for each interaction outlined during the process. While it would seem the most affecting interaction would be rs6852865-rs6910087-rs7246472-AGAGAC, this interaction also presents a large confidence interval, limiting the amount of confidence in the interaction. Further to this, the confidence intervals of the interactions varies with the largest observable in risk interaction, rs6852865-rs6910087-rs7246472-AGAGAC and the smallest observable in risk interaction, rs6910087-rs7246472-AGAC.

6.7 DISCUSSION

The proposed methodology functions as a filter, reducing the feature set through the stages performed. The first stage, QC, used ~500K SNPs and ~28K subjects provided by the DRIVE project. The performed processes are further defined in section 5.1 but an overview of the process results are available in Table 6.1. This stage resulted in a dataset of 320,247 features and 13,649 observations. Using the output from QC, training and testing datasets were split 75:25. Random Cohort Sampling was performed to split the training dataset into 9 sizeable subsets of individuals to create a viable averaging sample size for later in further stages. The sample sizes were proportional in the number of cases and controls that were assigned to each subset as demonstrated in

Figure 6-4.

The first set of results that were provided during the methodology that demonstrated the significance of features within the data were obtained during the association analysis stage. During this stage, an association analysis was performed for each of the 9 outlined subsets from previous stages. A further association analysis was performed on the full data output after the QC stage for the purpose of a comparison with standard methods and was further used in the feature selection stage. The results from the association analysis showed a number of suggestive values within the standard GWAS approach, however there were no features that exceeded the genome-wide significance threshold. In Figure 6-10 a comparison was undertaken to view the difference between the values obtained from the standard GWAS and the proposed 'random sampling regularisation' method. This shows a vast difference between the values obtained by each method that indicate that either the features from the standard GWAS are inflated or that the values from the 'random sampling regularisation' method are extremely undervaluing the expression of the feature.

Using the standard GWAS and 'random sampling regularisation' method results, feature selection was performed. At this stage, benchmarking was outlined to measure the performance of the proposed methodology against standard methods while using a balanced multiple testing adjustment method to optimise false discovery rate outcomes against the proposed methodology. These methods were split into cases 1-4. Case 1 represented the proposed methodology and used the mean and standard deviation of the 9 sample measures from association analysis to control for consistency in the feature set, resulting in an output of 57 variables. Within this method, genomic control was used to control for population stratification. Case 2 was applied to the standard GWAS method and represented a lenient approach, foregoing the use of genomic control to increase the feature set size to 48. Case 3 represented the most conservative approach, using genomic control and a threshold of p<0.3 that, while very lenient for q-value, resulted in a feature set of 17. Finally, Case 4 provided a balance between the two, applying genomic control and q-value but increasing the threshold to 0.4, yielding a feature set of 37.

Feature sets extracted from feature selection for each case were input into LAMPlink which outlined combination relationships that were evidenced in the data with threshold <0.005. As many of the feature sets contained the same SNPs, it was expected that there would be overlap in the combinations detected by each method. Therefore, the combinations were combined into one set. At this stage, the first deviation within the utilised methods is visible; Table 6.6 displays the combinations alongside the case methods that detected them. It is important to note at this stage that any methods that detected combinations that encompassed others were also noted to have detected the subset e.g. rs4602520-rs7246472 is a subset relationship of rs4602520-rs6910087-rs7246472. Case 1 detected 4 combinations, case 2 detected 7 combinations, case 3 detected 2 combinations and case 4 detected 5 combinations. From the standard method cases, it appears that the number of combinations that were detected depended on the number of SNPs that were input into LAMPlink, however it should be noted at this stage that while the proposed methodology had the largest feature set it produced the lowest number of combinations. Further to this, it is also noted that one of the combinations outlined by case 1 did not present cross-chromosome interaction; during this study it has not been omitted in order to demonstrate the lack of information presented from such interactions. Therefore, it would be considered that the proposed methodology only detected 3 viable combinations.

Using the combinations outlined from LAMPlink, each interaction was expanded and explored to consider the relationships between all SNPs and genomic states. During this stage, 3 different statistical methods were applied in order to compare results using various assumptions, this also serves as an outcome control to produce results that conclusively agree upon a result (within a small deviation from one another). Odds Ratio would give a measure of the effect size of the association between the genotypic presence and phenotypic presence. The genomic states were based on additive, dominant and recessive models, additionally exhaustive allelic states were also tested. This would consider the effect of the presence of dominant and recessive alleles while taking into consideration the singular genotypic states. Allelic states were used to consider the potential effect of a singular genomic state in the population. Results indicated an abundance of combinations that showed significance with the phenotype, any and all results that expressed < 0.05 were further analysed using a fresh testing dataset.

During the testing phase, a large majority of the outlined combinations were excluded due to low significance p-value. Penetrance and incidence were computed for the remaining combinations to consider the real-world effect of the combination. Using a lenient threshold of >60%, any combinations that showed a penetrance greater than this threshold were outlined. Further to this, it was considered that although every combination state had been analysed, an exhaustive search had not been performed. By using the allelic forms of the SNP combinations, further exploration using CHAID trees was able to outline further combinations that showed significance. These combinations were generated by omitting features to consider the presence effect of one allele within a SNP. These results are outlined in Table 6.12.

At this stage, an example of the predictive power available when using ML methods is provided in order to concur the suitability of alternative classification method CHAID in scenarios where few variants are outlined for classification analysis. Figure 6-24 provides a demonstration of machine learning model, Multi-Layered Perceptron, with small sets and shows the limitations that exist when applying classification to small feature sets, particularly in cases of linearly separable problems that do not warrant autonomous processes such as machine learning.

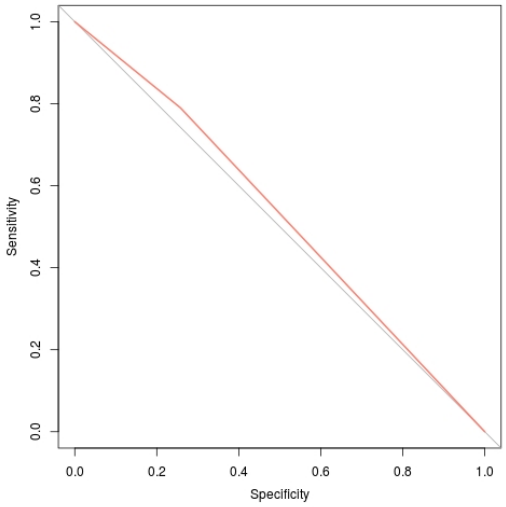


FIGURE 6-24: STANDARD MLP CLASSIFICATION USING FEATURE INTERACTION RS4602520 – RS7246472 –GAAC.

In order to confirm these finding, each combination was first compared against the previously outlined combinations to ensure that each new combination was presenting a better genomic option in either penetrance or incidence. Combinations outlined in grey in did not present a more significant option. Further to this, each combination was testing using the test dataset to confirm the finding from the CHAID analysis, however many of the combinations did not confirm a significant p-value, and those that did, showed little penetrance in the population.

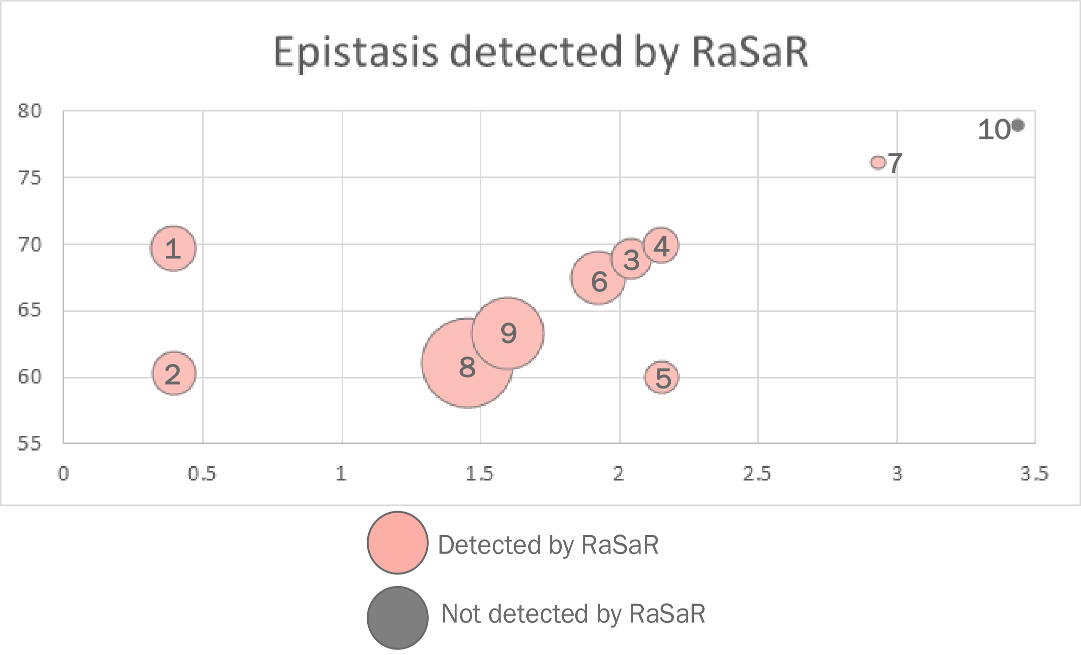


FIGURE 6-25: INTERACTIONS DETECTED USING THE PROPOSED METHODOLOGY. INTERACTIONS HIGHLIGHT IN GREY WERE NOT DETECTED.

The final set of variants as outlined in Table 6.16 showed a penetrance of >60%, significant p-value <0.05 and an OR >1 or OR <1. A focus of analyses in biomarkers requires real-world effect; this would entail reproducing the information in a clinical setting; however, this is not option at this point in time. The visualisation in Figure 6-27(a)-(d) provide a guidance for how significant these biomarkers appear to be given the information produced from this study using a separate colour for each case.. These bubble plots visualise the OR by the penetrance, while the size of the bubble indicates the incidence (population size) that carry the biomarker.

To further contextualise this information, Figure 6-26 presents a visualisation of the interactions with reference to the infamous BRCA1 gene using stats from Table 6.16 to plot by penetrance, incidence and risk association. Visible is the difference in effected size, BRCA1 is present in between 5%-15% of familial cases. The most prominent interactions found were collectively assembled using one or more of the SNPs in Table 6.17.

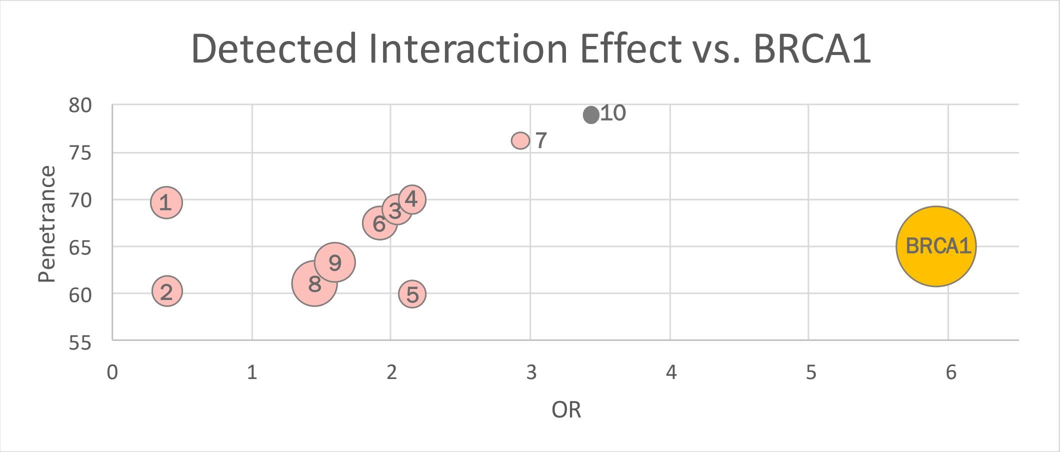


FIGURE 6-26: INTERACTIONS DETECTED IN COMPARISON WITH THE INFAMOUS BRCA1 GENE.

TABLE 6.17: GENOMIC CHARACTERISTIC OF INTERACTION VARIANTS

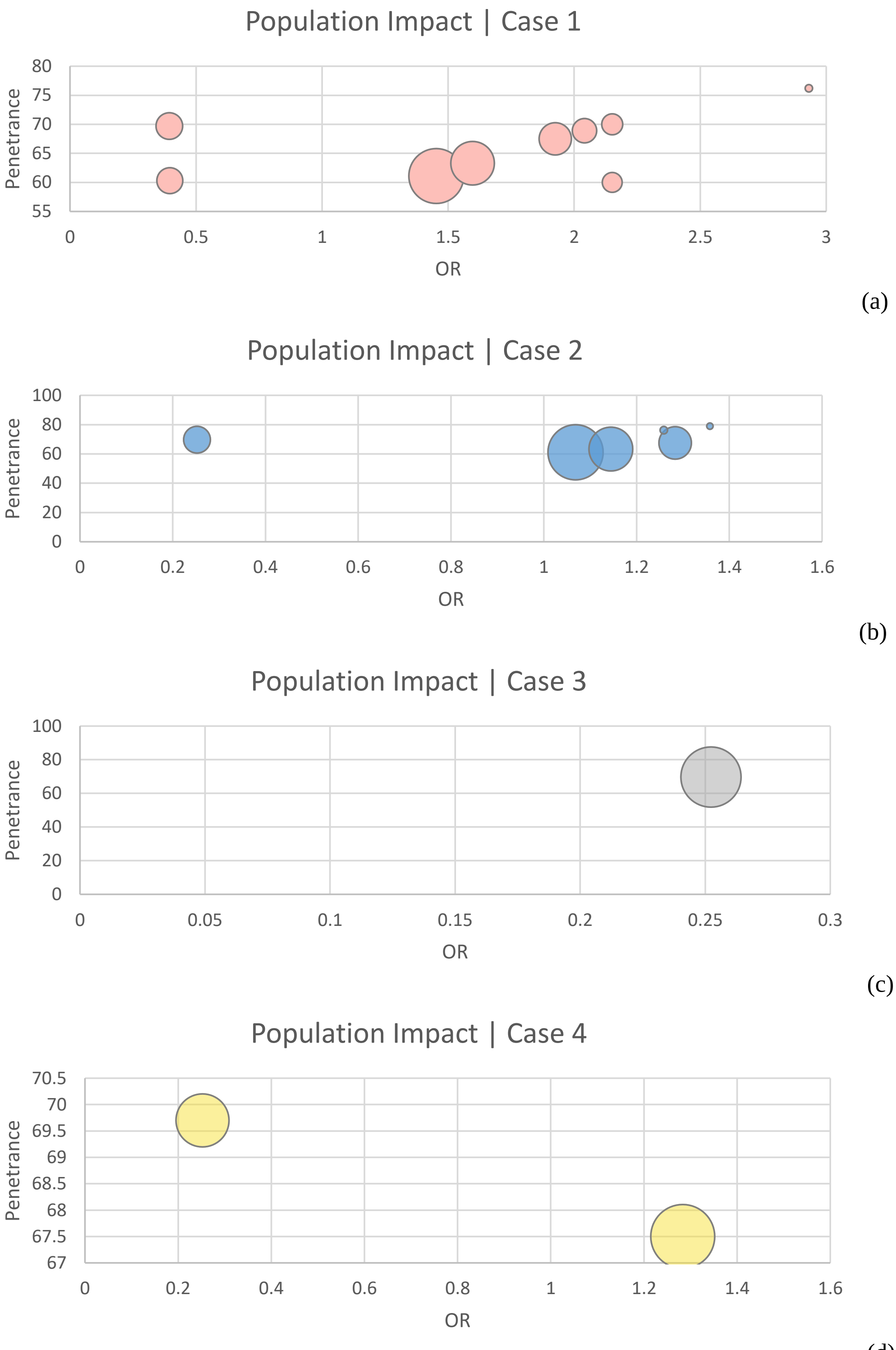
Variant	Chr	Pos (BP)	Allele	Gene/Nearest
1p12	1	120124218	C/T	HSD3BP4 (nearest(bp=9484)
rs4602520	4	61360284		AC095061.1 (nearest(bp=169383)
rs6910087	6	31377047		MICA/HCP5
rs7246472	19	29389111	C/A	AC011524.1
rs6852865	4	61251815	C/T	AC095061.1 (nearest(bp=277852)
rs4144827	2	164114775	T/C	RNU6-627P (nearest(bp=30730)

Using online tool SNPNexus [194], each SNP was explored for related publications that have highlighted or used SNPs in the context of histology and/or cancer. According to this tool, there are no publications in relation to cancer for these SNPs.

The 1p12 region of chromosome 1 appeared in a number of publication referring to its effect in breast cancer. [195] referred to its copy number imbalances, specifically in the case of losses could be used as a prognostic marker. Similarly [196] claims that alterations on 1p12 in relation to gene PHGDH are amplified in ~6% of breast cancers and 40% of melanomas.

The consequences of the outlined SNPs could lead to classification, prognostic, susceptibility or treatment guidance health systems particularly concerning the age of personalised medicine during which treatment will be catered to individual biomarkers including genetic variants. From the research presented, novel candidate variants have been outlined (See Table 6.17) as interactions as outlined in Table 6.16.

It should also be noted that the novel variants outlined during this study could be due to the optimised Illumina array whose specified 570K genotypic marker not only contain the most common 260K variants but additionally focus on the markers of interest for 5 cancer diseases.



(a)

(b)

(c)

(d)

FIGURE 6-27: DETECTED INTERACTIONS BASED ON EMPLOYED METHOD (A) CASE 1 (B) CASE 2 (C) CASE 3 (D) CASE 4.

Chapter 7: Conclusion & Future work

In this research, a novel methodology was proposed that caters for the needs of epistasis improving flexibility and inspired by random forests machine learning method. The novel methodology outlined in this research presents a statistically conservative option that outlines a number of interactions that present viable and reliable options that aim to improve reproducibility by using consistently transparent methods that are fully interpretable. To elaborate, the viability of these variants is conferred by the penetrance and risk, with initial results indicating its relevance using a variety of permutation tests in both training and testing datasets to indicate its significance of <0.05. Reliability is conferred using cascading statistical filters that aim to investigate and reduce the candidate set assuming a null hypothesis. In the context of this research, reliability is defined as the performance in relation to Odd's Ratio and Fisher's exact test, of which the novel methodology presents interaction candidates with unwavering performance of the feature regardless of the observations presented.

Reproducibility can be split into two different applications; reproducibility of the study is conferred using interpretable methods that can be replicated. Reproducibility of the variants can only be determined with a second study using an entirely different cohort set; this is harder to determine as although the evidence within this research suggest high significance in these candidate variants in relation to breast cancer, an additional dataset must be used to confirm the findings. The following sections discuss the contributions of the research with respect to the disciplines presented in this thesis.

The proposed methodology, RaSaR, outlined in this research has identified all but one of the most prominent and reliable variant interactions identified through each method. Figure 6-27 provides an overview of the interactions identified by each method as outlined in section 6.4 . Visible in Figure 6-27 is the clear distinction between the number of interactions identified by each method. The most successful method during this research next to the proposed methodology is Case 2 which used a threshold of $p<0.01$ without applying genomic control. Case 3, using a threshold of $q < 0.3$, produced no interaction results, but was still able to pick the variant that effected the largest proportion of the population. Case 4, using a threshold of $q < 0.4$, was a more lenient approach but was still only able to identify a minority of the interaction results.

This indicates that while multiple testing approaches can successfully identify the most prominent and reliable single-variant SNPs, they lack the statistical flexibility that is required when performing an Epistasis approach. Interactions that would indicate significance with a given phenotype will commonly be masked by other more dominating variants, particularly taking into consideration the long-standing issue of false positive rate. Therefore, the proposed methodology attempts to address the issue before the false detection rate becomes a problem. This method also affords the flexibility and lenience required to identify candidate variants for further analysis by applying a threshold that is lenient ($\mu<0.05$) but demanding consistency in this threshold by excluding any values whose σ sits outside of the specified threshold (σ. <0.025).

To discuss the issue that is present in the method of case 2, the abundance of interactions initially identified by this method could demonstrate and indicate the effects of not employing the genomic control method. By overlooking population stratification and omitting methods to control for population structure, the number of identified interactions is increased. While it could be argued that the number of interactions is directly related to the number of variants outlined during feature selection, it should be noted that the proposed methodology uses a much larger selection that did not produce as many interactions.

7.1 Limitations of the method

While the performance of the method has been proven significant in this research, there still remains issues that will likely effect outcomes either in a lenient or conservative fashion. One of the most prominent issues is the balance of the standard deviation threshold. While the method is adaptable to specify lenient or conservative thresholds, it is subject to the effects of anomalous data points; this occurrence would present a particular problem in cases were the majority of data points for one variant crowd in a tight cluster with one data point expressing in an anomalous range. The difficulty in addressing this point is the removal of any information could be extracting from the true representation of the variant.

Further to this, due to the nature of the method, it is accepted that an increase in False Negatives is a likely outcome of the use of this methodology. Additionally, rare alleles may be overlooked during association analysis due to lack of supporting evidence in each subset. Therefore, it is proposed that the outlined methodology would perform optimally for complex and common diseases.

While this research aims to address the issues that consistently plague the genomic research community, its viability has not been confirmed during this preliminary study. Further to this, while the exposure to type 1 errors is reduced in this method, it is still susceptible to the probability of chance, and will therefore still present type 1 errors.

7.2 Future Work

While novel contributions have been provided using the proposed methodology there are still areas that can be improved, and further research undertaken within the genomic community. The following section outlines areas for improvement pertaining to the methodology and further work that can be conducted using the proposed methodology. In its' early stages, one of the main areas for future work is related to RO2, false discovery rate, which would encourage the further development and confirmation of false positive rate reduction across a number of datasets to empirically measure the rate of improvement. This work requires further analyses to confirm the reliability of results, using a variety of datasets that range from small to large within the area of complex diseases. This will test the effects of the methodology and the efficacy of the epistasis detection capabilities given varying circumstances of the data.

Another effort of this research would be in the event of higher computational power, to benefit from the use of bootstrapping [197]. A large part of the effectiveness of this methodology lends itself to the effects of random sampling distribution; bootstrapping would provide an option to extend the k-fold sampling options to $k > 1000$, improving the standard deviation distribution metric that is heavily relied on in the feature selection method.

Further to this, the proposed methodology requires a lot of manual work to produce results, future work would focus on creating an autonomous algorithm that can perform stages 2-3, Random Cohort Sampling and Association Analysis. With its adaptability, the algorithm could be applicable in other fields outside of genomics such as social response; exploration into its transferrable aspects would be applicable for fields that use large datasets.

Concerning Breast Cancer, further research would focus on the reproducibility of these results using a separate dataset. A large sample size is required to pick up small interactions as the ones outlined in the results however with the current efforts in breast cancer, many large data samples are available via repositories such as DBGaP. The outlined future work demonstrates the intended efforts of the research to further evaluate the methodology. This discussion also outlines the potential adaptability of the methodology and areas for progression outside of the current field.

Appendix A

Selected Features Pre-Epistasis

Case 1

SNP	CHR	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	σ	μ
chr1_120124218_C_T	1	0.00182	0.000126	0.005829	0.001369	0.000856	0.001115	0.002829	0.03847	0.000913	0.005925	0.012318
chr2_121527169_A_G	2	0.1113	0.01447	0.01938	0.001959	0.002855	0.1026	0.000592	0.004446	4.08E-05	0.028627	0.044939
chr2_216887593_A_G	2	0.003812	0.06652	0.001706	0.002489	0.003678	0.03752	0.03734	0.001355	0.004537	0.017662	0.023662
chr2_47786807_A_C	2	0.01738	0.000375	0.000152	0.01725	0.0389	0.03229	0.003164	0.02648	0.06438	0.022263	0.021067
chr3_119792288_A_G	3	0.000993	0.1113	0.005661	0.09253	0.000241	0.000483	0.000303	0.000394	0.0117	0.024845	0.044109
chr6_31240692_A_G	6	0.006081	0.02646	0.118	0.01452	0.003859	0.000523	0.00456	0.003871	0.002688	0.020062	0.037596
chr6_31330066_A_G	6	0.005048	0.02416	0.004371	0.01483	0.00084	0.000487	0.000649	0.002386	0.01676	0.007726	0.008653
chr6_31336100_A_C	6	0.004838	0.02507	0.004706	0.01985	0.001114	0.00054	0.000683	0.002495	0.01945	0.00875	0.009783
chr6_31342960_A_C	6	0.01365	0.03372	0.002003	0.04623	0.003024	0.002536	0.001411	0.01107	0.1018	0.023938	0.033186
chr6_31368964_A_G	6	0.004065	0.06527	0.0187	0.04392	0.03023	0.01131	0.005611	0.000279	0.004602	0.020443	0.022047
chr9_74065947_C_T	9	0.002531	0.04396	0.000805	0.004719	0.000763	0.01025	0.0266	0.006482	0.03447	0.014509	0.016243
chr9_74076686_A_C	9	0.002709	0.03031	0.004877	0.0117	0.000573	0.001676	0.02811	0.02879	0.086	0.021638	0.027136
kgp12436430	8	0.009594	0.02186	0.000119	0.003364	0.02802	0.05406	0.02533	0.001156	0.03869	0.020244	0.018508
rs11609829	12	0.01699	0.02939	0.06681	0.02777	0.01285	0.02808	0.01992	0.02397	0.003893	0.025519	0.017558
rs11640710	16	0.07423	0.02145	0.02478	0.005283	0.0409	0.01157	0.0119	0.005686	0.01879	0.023843	0.021878
rs11876265	18	0.005667	0.002553	0.02367	0.001862	0.003959	0.002207	0.002639	0.002979	0.000346	0.005098	0.007115
rs11994	5	0.03106	0.006043	0.02204	0.000508	0.05345	0.005827	0.01599	0.000803	0.002904	0.015403	0.017705
rs1550638	15	0.0119	0.006787	0.001119	0.02548	0.008662	0.02758	0.01736	0.06301	0.07601	0.026434	0.026063
rs17330266	18	0.002136	0.000193	0.008551	0.001067	0.002476	0.000189	0.000789	0.003275	0.000726	0.002156	0.002624
rs1851736	1	0.000658	0.01329	0.01798	0.0783	0.006799	0.0136	0.04949	0.04732	0.01708	0.027169	0.025494

rs2216470	12	0.04042	0.001243	0.01474	0.01813	0.004552	0.03631	0.000568	0.07027	0.000671	0.020767	0.023921
rs2428486	6	0.003534	0.07156	0.02052	0.04497	0.02519	0.005154	0.004643	0.000206	0.006785	0.020285	0.023985
rs247930	12	0.003459	0.005506	0.000288	0.1397	0.006085	0.06417	0.000736	0.02051	0.01346	0.028213	0.046303
rs2507976	6	0.02488	0.01888	0.04445	0.0744	0.04608	0.01164	0.002606	0.005862	0.009081	0.026431	0.023931
rs2523467	6	0.005946	0.05795	0.02307	0.04159	0.04347	0.01039	0.004396	0.000203	0.004172	0.021243	0.021297
rs2596542	6	0.004849	0.05725	0.01852	0.03598	0.03433	0.0129	0.006809	0.000337	0.004564	0.019504	0.019128
rs28391573	18	0.03864	0.001486	0.01391	0.000672	0.01387	0.03901	0.03298	0.05101	0.001915	0.021499	0.01916
rs2844529	6	0.003628	0.07022	0.02098	0.04331	0.02519	0.005154	0.004643	0.000208	0.00648	0.019979	0.023431
rs2844551	6	0.004967	0.01009	0.001794	0.02926	0.02913	0.003048	0.001096	0.000453	0.02299	0.011425	0.012241
rs2974161	2	0.02873	0.02691	0.02412	0.003423	0.006656	0.06097	0.04284	0.03296	0.002816	0.025492	0.019319
rs34821683	6	0.05632	0.08188	0.006892	0.02612	0.02817	0.02521	0.00381	0.002211	0.03115	0.029085	0.026008
rs3819301	6	0.08141	0.02584	0.03642	0.001338	0.01626	0.002445	0.01476	0.006418	0.02749	0.023598	0.024741
rs3924215	9	0.004817	0.00149	0.01314	0.002765	0.01826	0.000105	0.004169	0.001316	0.01934	0.007267	0.007558
rs4144827	2	0.01268	0.009614	0.01489	0.02963	0.04619	0.08597	0.00905	0.006917	0.001235	0.02402	0.026963
rs4313504	10	0.000143	0.01838	0.003775	0.01751	0.01612	0.105	0.08949	0.001705	0.000804	0.028103	0.040074
rs4357555	1	0.001442	0.09346	0.008807	0.05822	0.001121	0.02388	0.01442	0.01112	0.008245	0.024524	0.031149
rs4602520	4	0.001682	0.06598	0.002737	0.000396	0.00037	0.1038	0.00943	0.000108	5.14E-06	0.020501	0.037823
rs4624908	6	0.01399	0.03197	0.001756	0.05327	0.003114	0.002584	0.001129	0.01149	0.1094	0.025411	0.035995
rs4959071	6	0.01352	0.01276	0.000461	0.00325	0.004586	0.001637	0.000631	0.00055	0.0243	0.006855	0.008279
rs6011609	20	5.22E-05	0.002207	0.001765	0.003905	0.000251	0.002862	0.04853	0.026	0.004507	0.010009	0.016513
rs6457402	6	0.003945	0.03393	0.003413	0.02647	0.002142	0.001092	0.00159	0.002383	0.03262	0.011954	0.014453
rs6602225	10	0.003387	0.02439	9.63E-05	0.000232	0.000402	0.000199	0.02661	0.002504	0.2036	0.029047	0.066314
rs6842825	4	0.000709	0.003448	0.01009	0.0212	0.02919	0.02815	0.009897	0.002404	0.000386	0.011719	0.011597
rs6910087	6	0.07913	0.02305	0.000698	0.003268	0.003939	0.004515	0.001973	0.00919	0.09944	0.025023	0.037386
rs6911024	6	0.07119	0.0185	0.000366	0.001538	0.002153	0.002887	0.001472	0.007005	0.08943	0.021616	0.03404
rs6932730	6	0.03496	0.09316	0.009491	0.02453	0.02483	0.01279	0.007462	0.002182	0.0447	0.028234	0.027967
rs6936035	6	0.007011	0.03223	0.001297	0.04873	0.000881	0.001464	0.000885	0.01216	0.09253	0.02191	0.031339
rs7246472	19	0.01032	0.005216	0.00598	0.000163	0.002913	0.001664	0.04812	0.01114	0.01721	0.011414	0.014765
rs7754026	6	0.01635	0.03372	0.002003	0.04348	0.003306	0.003153	0.001411	0.01146	0.09688	0.023529	0.031337

rs7775117	6	0.01477	0.03372	0.002003	0.04348	0.003306	0.002775	0.001411	0.01036	0.09688	0.023189	0.031472
rs7822226	8	0.02228	0.000107	0.04509	0.003857	0.02694	0.02529	3.84E-05	0.001079	4.07E-07	0.013854	0.016521
rs787025	10	0.03863	0.02815	0.001523	0.004796	0.001821	0.1349	0.009349	0.02958	0.005688	0.028271	0.042298
rs8066706	17	0.006088	0.001389	0.00598	0.02501	0.03331	0.1496	0.002088	0.006453	0.000165	0.025565	0.047898
rs8182119	16	0.1238	0.02538	0.01492	0.006323	0.01586	0.002915	0.00178	0.006008	0.03626	0.025916	0.03841
rs859767	2	0.02451	0.001383	4.48E-05	0.003264	0.01783	0.01585	9.73E-05	0.000343	0.00635	0.007741	0.009244
rs9263475	6	0.05646	0.007221	0.014	5.40E-05	0.01736	0.007351	0.08937	0.0111	0.01512	0.024226	0.029257
rs9958743	18	0.002309	0.0002	0.00809	0.00129	0.002492	0.000212	0.000578	0.003044	0.000658	0.002097	0.00248

CASE 2

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	chr1_120124218_C_T	1.2E+08	G	0.00884	0.01772	A	33.06	8.93E-09	0.4943
1	chr1_214950361_A_G	2.15E+08	G	0.07362	0.09216	A	24.74	6.56E-07	0.7829
2	chr2_121527169_A_G	1.22E+08	A	0.02629	0.03834	G	25.35	4.78E-07	0.6774
2	rs859767	1.35E+08	G	0.3967	0.3596	A	31.84	1.67E-08	1.171
2	rs2280219	1.35E+08	A	0.3826	0.3507	G	23.88	1.02E-06	1.147
2	rs6750788	1.35E+08	A	0.3933	0.3596	G	26.28	2.95E-07	1.154
2	rs6759065	1.35E+08	A	0.3945	0.3607	G	26.52	2.61E-07	1.155
2	rs6705916	1.35E+08	G	0.4691	0.4357	A	24.41	7.79E-07	1.144
2	rs6430538	1.36E+08	A	0.4567	0.4182	G	32.85	9.96E-09	1.17
2	rs3769027	1.36E+08	G	0.2764	0.2466	A	24.91	6.01E-07	1.167
2	rs3814354	1.36E+08	A	0.4831	0.4468	G	28.87	7.74E-08	1.157
2	chr2_216887593_A_G	2.17E+08	A	0.3378	0.3697	G	24.24	8.51E-07	0.8698
3	chr3_119792288_A_G	1.2E+08	A	0.02366	0.01371	G	29.05	7.04E-08	1.743
4	rs6842825	57125176	A	0.02007	0.01141	G	26.03	3.36E-07	1.774
4	rs6852865	61251815	A	0.08693	0.06933	G	23.31	1.38E-06	1.278
4	rs4602520	61360284	G	0.09108	0.06984	A	32.8	1.02E-08	1.334
5	rs3756765	1.38E+08	A	0.2056	0.2329	G	23.72	1.11E-06	0.8525
5	rs11994	1.5E+08	A	0.3961	0.4292	G	24.62	6.98E-07	0.8724
6	chr6_31240692_A_G	31240692	A	0.423	0.3899	G	24.67	6.82E-07	1.147
6	chr6_31330066_A_G	31330066	G	0.2644	0.2328	A	28.97	7.36E-08	1.184
6	rs6457402	31334864	A	0.2507	0.2211	C	26.37	2.82E-07	1.178
6	chr6_31336100_A_C	31336100	A	0.2643	0.2331	C	28.26	1.06E-07	1.182
6	rs6936035	31341156	G	0.2294	0.2017	A	24.69	6.74E-07	1.178
6	rs2844551	31342781	G	0.32	0.288	A	26.4	2.77E-07	1.164
6	rs4959071	31346436	G	0.1722	0.1451	A	29.88	4.59E-08	1.226
6	rs2844529	31353593	A	0.354	0.323	G	23.32	1.37E-06	1.148
6	rs2428486	31354104	G	0.354	0.323	A	23.28	1.40E-06	1.148
6	rs6911024	31368451	G	0.1394	0.1164	A	25.52	4.37E-07	1.229
6	rs6910087	31377047	A	0.1403	0.1183	G	23.35	1.35E-06	1.216
8	rs11992223	4089132	A	0.4394	0.4034	C	28.92	7.56E-08	1.159
8	rs7822226	4091132	G	0.4864	0.4472	A	33.4	7.50E-09	1.171
9	chr9_74065947_C_T	74065947	A	0.07308	0.09148	G	24.55	7.23E-07	0.783
9	rs3924215	1.35E+08	G	0.1193	0.1439	A	28.99	7.27E-08	0.8058
10	rs6602225	7124442	A	0.1873	0.1593	C	29.61	5.28E-08	1.216
10	rs9703900	58994564	G	0.3853	0.3533	A	24	9.65E-07	1.148
10	rs4313504	1.01E+08	A	0.1491	0.1732	G	23.55	1.22E-06	0.8362
11	rs10736499	1.19E+08	G	0.3019	0.2709	A	25.61	4.19E-07	1.164
11	rs10790316	1.19E+08	A	0.3596	0.3279	G	24.3	8.26E-07	1.151
11	rs7950231	1.19E+08	G	0.3632	0.3309	A	25.12	5.38E-07	1.154
12	rs2216470	46101115	A	0.4675	0.4347	G	23.68	1.14E-06	1.142
12	rs247930	46239012	G	0.5094	0.4766	A	23.43	1.30E-06	1.14

17	rs8066706	1692140	G	0.03181	0.02091	A	25.02	5.69E-07	1.539
18	rs17330266	73695934	A	0.269	0.2338	G	35.84	2.14E-09	1.206
18	rs9958743	73697477	A	0.2691	0.2338	G	35.92	2.06E-09	1.206
18	rs11876265	73706661	G	0.2803	0.2476	A	30.05	4.20E-08	1.184
19	rs7246472	29389111	A	0.09928	0.0795	C	26.08	3.27E-07	1.276
20	rs6011609	61727307	A	0.009231	0.01781	G	30.2	3.90E-08	0.5139
22	rs12170250	29538730	A	0.01823	0.01036	G	23.67	1.15E-06	1.774

CASE 3

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	chr1_120124218_C_T	1.2E+08	G	0.00884	0.01772	A	33.06	8.93E-09	0.4943
2	rs859767	1.35E+08	G	0.3967	0.3596	A	31.84	1.67E-08	1.171
2	rs6430538	1.36E+08	A	0.4567	0.4182	G	32.85	9.96E-09	1.17
2	rs3814354	1.36E+08	A	0.4831	0.4468	G	28.87	7.74E-08	1.157
3	chr3_119792288_A_G	1.2E+08	A	0.02366	0.01371	G	29.05	7.04E-08	1.743
4	rs4602520	61360284	G	0.09108	0.06984	A	32.8	1.02E-08	1.334
6	chr6_31330066_A_G	31330066	G	0.2644	0.2328	A	28.97	7.36E-08	1.184
6	chr6_31336100_A_C	31336100	A	0.2643	0.2331	C	28.26	1.06E-07	1.182
6	rs4959071	31346436	G	0.1722	0.1451	A	29.88	4.59E-08	1.226
8	rs11992223	4089132	A	0.4394	0.4034	C	28.92	7.56E-08	1.159
8	rs7822226	4091132	G	0.4864	0.4472	A	33.4	7.50E-09	1.171
9	rs3924215	1.35E+08	G	0.1193	0.1439	A	28.99	7.27E-08	0.8058
10	rs6602225	7124442	A	0.1873	0.1593	C	29.61	5.28E-08	1.216
18	rs17330266	73695934	A	0.269	0.2338	G	35.84	2.14E-09	1.206
18	rs9958743	73697477	A	0.2691	0.2338	G	35.92	2.06E-09	1.206
18	rs11876265	73706661	G	0.2803	0.2476	A	30.05	4.20E-08	1.184
20	rs6011609	61727307	A	0.009231	0.01781	G	30.2	3.90E-08	0.5139

CASE 4

CHR	SNP	BP	A1	F_A	F_U	A2	CHISQ	P	OR
1	chr1_120124218_C_T	1.2E+08	G	0.00884	0.01772	A	33.06	8.93E-09	0.4943
1	chr1_214950361_A_G	2.15E+08	G	0.07362	0.09216	A	24.74	6.56E-07	0.7829
2	chr2_121527169_A_G	1.22E+08	A	0.02629	0.03834	G	25.35	4.78E-07	0.6774
2	rs859767	1.35E+08	G	0.3967	0.3596	A	31.84	1.67E-08	1.171
2	rs6750788	1.35E+08	A	0.3933	0.3596	G	26.28	2.95E-07	1.154
2	rs6759065	1.35E+08	A	0.3945	0.3607	G	26.52	2.61E-07	1.155
2	rs6705916	1.35E+08	G	0.4691	0.4357	A	24.41	7.79E-07	1.144
2	rs6430538	1.36E+08	A	0.4567	0.4182	G	32.85	9.96E-09	1.17
2	rs3769027	1.36E+08	G	0.2764	0.2466	A	24.91	6.01E-07	1.167
2	rs3814354	1.36E+08	A	0.4831	0.4468	G	28.87	7.74E-08	1.157
2	chr2_216887593_A_G	2.17E+08	A	0.3378	0.3697	G	24.24	8.51E-07	0.8698

3	chr3_119792288_A_G	1.2E+08	A	0.02366	0.01371	G	29.05	7.04E-08	1.743
4	rs6842825	57125176	A	0.02007	0.01141	G	26.03	3.36E-07	1.774
4	rs4602520	61360284	G	0.09108	0.06984	A	32.8	1.02E-08	1.334
5	rs11994	1.5E+08	A	0.3961	0.4292	G	24.62	6.98E-07	0.8724
6	chr6_31240692_A_G	31240692	A	0.423	0.3899	G	24.67	6.82E-07	1.147
6	chr6_31330066_A_G	31330066	G	0.2644	0.2328	A	28.97	7.36E-08	1.184
6	rs6457402	31334864	A	0.2507	0.2211	C	26.37	2.82E-07	1.178
6	chr6_31336100_A_C	31336100	A	0.2643	0.2331	C	28.26	1.06E-07	1.182
6	rs6936035	31341156	G	0.2294	0.2017	A	24.69	6.74E-07	1.178
6	rs2844551	31342781	G	0.32	0.288	A	26.4	2.77E-07	1.164
6	rs4959071	31346436	G	0.1722	0.1451	A	29.88	4.59E-08	1.226
6	rs6911024	31368451	G	0.1394	0.1164	A	25.52	4.37E-07	1.229
8	rs11992223	4089132	A	0.4394	0.4034	C	28.92	7.56E-08	1.159
8	rs7822226	4091132	G	0.4864	0.4472	A	33.4	7.50E-09	1.171
9	chr9_74065947_C_T	74065947	A	0.07308	0.09148	G	24.55	7.23E-07	0.783
9	rs3924215	1.35E+08	G	0.1193	0.1439	A	28.99	7.27E-08	0.8058
10	rs6602225	7124442	A	0.1873	0.1593	C	29.61	5.28E-08	1.216
11	rs10736499	1.19E+08	G	0.3019	0.2709	A	25.61	4.19E-07	1.164
11	rs10790316	1.19E+08	A	0.3596	0.3279	G	24.3	8.26E-07	1.151
11	rs7950231	1.19E+08	G	0.3632	0.3309	A	25.12	5.38E-07	1.154
17	rs8066706	1692140	G	0.03181	0.02091	A	25.02	5.69E-07	1.539
18	rs17330266	73695934	A	0.269	0.2338	G	35.84	2.14E-09	1.206
18	rs9958743	73697477	A	0.2691	0.2338	G	35.92	2.06E-09	1.206
18	rs11876265	73706661	G	0.2803	0.2476	A	30.05	4.20E-08	1.184
19	rs7246472	29389111	A	0.09928	0.0795	C	26.08	3.27E-07	1.276
20	rs6011609	61727307	A	0.009231	0.01781	G	30.2	3.90E-08	0.5139

APPENDIX B

LAMPlink combinations output based on features selected for each case.

CASE 1

COMBID	RAW_P	Adjusted_P	COMB
1	4.7615e-09	1.3856e-06	chr1_120124218_C_T
2	7.6298e-09	2.2203e-06	rs4602520
3	1.8594e-08	5.4108e-06	rs6011609
4	2.0857e-08	6.0693e-06	rs3924215
5	1.2047e-07	3.5057e-05	chr3_119792288_A_G
6	4.9655e-07	0.0001445	rs6842825
7	5.1652e-07	0.00015031	chr2_121527169_A_G
8	6.1626e-07	0.00017933	rs6911024
10	9.666e-07	0.00028128	rs8066706
11	1.4024e-06	0.00040811	chr9_74065947_C_T
12	1.4654e-06	0.00042642	chr2_47786807_A_C
13	1.8291e-06	0.00053227	rs4602520, rs6910087,rs7246472
14	2.1604e-06	0.00062868	rs7246472
15	2.4134e-06	0.00070229	chr9_74065947_C_T, chr9_74076686_A_C
17	2.8572e-06	0.00083143	rs787025
18	3.2627e-06	0.00094946	chr9_74076686_A_C
19	4.9796e-06	0.0014491	rs6910087
20	1.0127e-05	0.0029469	rs4144827
21	1.0708e-05	0.0031159	rs4144827, rs4602520

CASE 2

COMBID	RAW_P	Adjusted_P	COMB
1	4.7615e-09	1.2856e-06	chr1_120124218_C_T
2	7.6298e-09	2.06e-06	rs4602520
3	1.8594e-08	5.0204e-06	rs6011609
4	2.0857e-08	5.6313e-06	rs3924215
5	1.2047e-07	3.2527e-05	chr3_119792288_A_G
6	1.4659e-07	3.958e-05	chr1_214950361_A_G, rs3924215
7	1.9913e-07	5.3764e-05	rs6911024, rs12170250
9	3.3143e-07	8.9486e-05	rs6852865, rs4602520
10	4.9655e-07	0.00013407	rs6842825
11	5.1652e-07	0.00013946	chr2_121527169_A_G
12	5.3133e-07	0.00014346	rs6852865
13	6.1626e-07	0.00016639	rs6911024
15	9.666e-07	0.00026098	rs8066706
16	1.0708e-06	0.00028911	chr1_214950361_A_G
17	1.0939e-06	0.00029536	rs12170250
18	1.4024e-06	0.00037866	chr9_74065947_C_T
19	1.8291e-06	0.00049385	rs4602520, rs6910087,rs7246472
20	2.1604e-06	0.00058331	rs7246472
22	3.1746e-06	0.00085714	rs6852865, rs4602520,rs6910087,rs7246472
23	3.9059e-06	0.0010546	rs6852865, rs6910087,rs7246472
25	4.9796e-06	0.0013445	rs6910087

CASE 3

COMBID	RAW_P	Adjusted_P	COMB
1	7.0224e-12	1.1938e-10	chr1_120124218_C_T
2	1.6752e-08	2.8478e-07	rs6011609
3	3.4216e-08	5.8167e-07	rs3924215
4	8.0663e-08	1.3713e-06	rs4602520
5	1.4145e-07	2.4046e-06	chr3_119792288_A_G
6	5.473e-05	0.00093042	rs3924215, rs6011609
7	8.6376e-05	0.0014684	chr1_120124218_C_T, rs6011609

CASE 4

COMBID	RAW_P	Adjusted_P	COMB
1	7.0224e-12	1.1657e-09	chr1_120124218_C_T
2	1.6752e-08	2.7808e-06	rs6011609
3	3.4216e-08	5.6799e-06	rs3924215
4	3.7644e-08	6.2488e-06	rs6911024
5	8.0663e-08	1.339e-05	rs4602520
6	1.4145e-07	2.348e-05	chr3_119792288_A_G
7	2.6106e-07	4.3336e-05	rs7246472
8	2.9374e-07	4.8761e-05	chr1_214950361_A_G, rs3924215
9	4.1186e-07	6.8369e-05	rs4602520, rs6911024,rs7246472
10	1.0484e-06	0.00017404	rs6842825
11	1.5974e-06	0.00026516	rs8066706
12	2.4179e-06	0.00040137	chr1_214950361_A_G
13	3.1453e-06	0.00052212	chr9_74065947_C_T
14	3.8726e-06	0.00064285	rs6911024, rs7246472
15	5.1619e-06	0.00085688	rs4602520, rs7246472
16	8.0748e-06	0.0013404	rs4602520, rs6911024

APPENDIX C

PPP Confidence Scores

Combination	C1	C2	C3	C4
rs4602520; rs6910087; rs7246472	O	O		
9q21.13; 9q21.13	O			
rs4602520; rs4144827	O			
1q41; rs3924215		O		O
rs6911024; rs12170250		O		
rs6852865; rs4602520		O		
rs6852865; rs4602520; rs6910087; rs7246472		O		
rs6852865; rs6910087; rs7246472		O		
rs3924215; rs6011609			O	
1p12; rs6011609			O	
rs4602520; rs6911024; rs7246472				O
rs6911024; rs7246472				O
rs4602520; rs7246472	O	O		O
rs4602520; rs6911024				O

Variant/s	Adjusted p-value	CHR
rs4602520; rs6910087; rs7246472	0.00053227	
rs4602520	2.2203e-06	4
rs6910087	0.0014491	6
rs7246472	0.00062868	19
Petals	2	0

PPPConf = (2/3(.5))+((3-0)/3(.5)) = 0.83333

Variant/s	Adjusted p-value	CHR
chr9_74065947_C_T,chr9_74076686_A_C	0.00070229	
chr9_74065947_C_T	0.00040811	9
chr9_74076686_A_C	0.00094946	9
Petals	1	2

PPPConf = (1/2(.5))+((2-2)/2(.5)) = 0.25

Variant/s	Adjusted p-value	CHR
rs4602520; rs4144827	0.0031159	
chr9_74065947_C_T	2.2203e-06	4
chr9_74076686_A_C	0.0029469	2
Petals	0	0

PPPConf = (0/2(.5))+((2-2)/2(.5)) = 0.75

Variant/s	Adjusted p-value	CHR
chr1_214950361_A_G,rs3924215	3.958e-05	
chr1_214950361_A_G	0.00028911	1
rs3924215	5.6313e-06	9
Petals	1	0

PPPConf = (1/2(.5))+((2-0)/2(.5)) = 0.75

Variant/s	Adjusted p-value	CHR
rs6911024,rs12170250	5.3764e-05	
rs6911024	0.00016639	6
rs12170250	0.00029536	22
Petals	1	0

PPPConf = (2/2(.5))+((2-0)/2(.5)) = 1

Variant/s	Adjusted p-value	CHR
rs6852865; rs4602520; rs6910087; rs7246472	0.00085714	
rs6852865	0.00014346	4
rs4602520	2.06e-06	4
rs6910087	0.0013445	6
rs7246472	0.00058331	19
Petals	1	2

PPPConf = (1/4(.5))+((4-2)/4(.5)) = 0.375

Variant/s	Adjusted p-value	CHR
rs6852865; rs4602520; rs6910087; rs7246472	0.0010546	
rs6852865	0.00014346	4
rs6910087	0.0013445	6
rs7246472	0.00058331	19
Petals	1	0

PPPConf = (1/3(.5))+((3-3)/3(.5)) = 0.667

Variant/s	Adjusted p-value	CHR
rs3924215,rs6011609	0.00093042	
rs3924215	5.8167e-07	9
rs6011609	2.8478e-07	20
Petals	0	0

PPPConf = (0/2(.5))+((2-0)/2(.5)) = 0.5

Variant/s	Adjusted p-value	CHR
chr1_120124218_C_T,rs6011609	0.0014684	
chr1_120124218_C_T	1.1938e-10	1
rs6011609	2.8478e-07	20
Petals	0	0

PPPConf = (0/2(.5))+((2-0)/2(.5)) = 0.5

Variant/s	Adjusted p-value	CHR
rs4602520,rs6911024,rs7246472	6.8369e-05	
rs4602520	1.339e-05	4
rs6911024	6.2488e-06	6
rs7246472	4.3336e-05	19
Petals	0	0

PPPConf = (0/3(.5))+((3-0)/3(.5)) = 0.5

Variant/s	Adjusted p-value	CHR
rs6911024,rs7246472	0.00064285	
rs6911024	6.2488e-06	6
rs7246472	4.3336e-05	19
Petals	0	0

PPPConf = (0/2(.5))+((2-0)/2(.5)) = 0.5

Variant/s	Adjusted p-value	CHR
rs4602520,rs7246472	0.00085688	
rs4602520	1.339e-05	4
rs7246472	4.3336e-05	19
Petals	0	0

PPPConf = (0/2(.5))+((2-0)/2(.5)) = 0.5

Variant/s	Adjusted p-value	CHR
rs4602520,rs6911024,rs7246472	0.0013404	
rs4602520	1.339e-05	4
rs6911024	6.2488e-06	6
Petals	0	0

PPPConf = (0/2(.5))+((2-0)/2(.5)) = 0.5

APPENDIX D

Interaction ID combinations

ID	Combination
1	rs4602520, rs6910087, rs7246472
2	9q21.13, 9q21.13
3	rs4144827, rs4602520
4	1q44, rs3924215
5	rs6911024, rs12170250
6	rs6852865, rs4602520
7	rs6852865, rs4602520, rs6910087, rs7246472
8	rs6852865, rs6910087, rs7246472
9	rs3924215, rs6011609
10	1p12, rs6011609
11	rs4602520, rs6911024, rs7246472

APPENDIX E

Inference Analysis extended training results

EXTENDED VERSION OF TABLE 6.8

Combination	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
1	rs4602520	Dominant	5.51e-09	6.32e-09	1.253399	1.370042	1.497540	5.889e-09
		Additive	3.63e-08	4.017e-08				
		Recessive	0.264	0.2768	0.9099653	1.276314	1.7901530	0.2356
	rs6910087	Dominant	1.895e-06	2.14e-06	1.151754	1.241021	1.337206	1.955e-06
		Additive	5.189e-06	5.407e-06				
		Recessive	0.01464	0.01573	1.117309	1.393698	1.738458	0.0135
	rs7246472	Dominant	1.317e-06	1.503e-06	1.177445	1.281104	1.393890	1.371e-06
		Additive	6.017e-07	6.844e-07				
		Recessive	0.001678	0.002388	1.396208	2.046763	3.000439	0.002069
	rs4602520, rs6910087	Dominant	2.216e-05	2.815e-05	1.296391	1.531307	1.808791	2.568e-05
		Recessive	NA	NA	NA	NA	NA	NA
		AAAA			1.090952	1.388031	1.766008	0.02513
		AAAG			1.035472	1.124566	1.221325	0.01931
		AAGG			0.7074017	0.756109	0.8081700	4.984e-12
		GAAG			1.272643	1.52535	1.828236	0.0001259
		GAAA			0.7790973	1.375466	2.4283307	0.3563
		GGAG			0.8988065	1.667903	3.0951043	0.1735
		GGGG			0.7310389	1.106232	1.6739878	0.6885
		GAGG			1.159623	1.285759	1.425616	6.228e-05
	rs4602520, rs7246472	Dominant	0.0005296	0.0005853	1.256756	1.543559	1.895812	0.0005136
		Recessive	NA	NA	NA	NA	NA	NA
		AAAA			1.254384	1.87664	2.807575	0.01016
		AAAC			1.080824	1.184612	1.298368	0.002373
		AACC			0.6836437	0.7328811	0.7856647	1.981e-13
		GAAC			1.195404	1.48089	1.834557	0.002564
		GACC			1.187408	1.310497	1.446347	6.502e-06
		GGCC			0.850018	1.223292	1.760484	0.3625
	rs6910087, rs7246472	Dominant	2.033e-06	2.15e-06	1.173655	1.27729	1.390076	1.961e-06
		Recessive	NA	NA	NA	NA	NA	NA
		AAAC			1.137773	1.977866	3.438256	0.04248
		AACC			1.029704	1.311648	1.670791	0.06521
		AGAC			1.180675	1.399776	1.659537	0.001155
		AGCC			1.040629	1.130713	1.228596	0.01494
		GGAA			1.135758	1.723343	2.614916	0.03179
		GGAC			1.047497	1.153669	1.270603	0.01487
		GGCC			0.7298925	0.7795632	0.8326141	4.923e-10
	rs4602520, rs6910087, rs7246472	Dominant	1.227e-06	2.977e-06	2.149330	3.334774	5.174038	6.479e-06
		Recessive	NA	NA	NA	NA	NA	NA
		AAGGAA			1.107386	1.712342	2.647782	0.04238
		AAGGAC			1.036098	1.148786	1.273730	0.02712
		AAGGCC			0.6808518	0.7256118	0.7733145	2.22e-16
		GAAGAC			1.971355	3.195814	5.180815	7.632e-05
		GAAGCC			1.041139	1.269444	1.547812	0.04778
		GAGGCC			1.165563	1.303446	1.457641	9.671e-05
2	9q21.13	Dominant	1.543e-06	1.566e-06	0.7106870	0.7751917	0.8455511	1.427e-06
		Additive	5.081e-06	5.19e-06				
		Recessive	0.05874	0.06384	0.4455113	0.6450179	0.9338667	0.0513
	9q21.13	Dominant	3.322e-06	3.69e-06	0.6918528	0.7617881	0.8387926	3.358e-06
		Additive	1.346e-05	1.392e-05				
		Recessive	0.1094	0.1386	0.3790459	0.6186729	1.0097885	0.1069
	9q21.13, 9q21.13	Dominant	2.546e-06	2.736e-06	0.6880050	0.7579426	0.8349895	2.492e-06
		Recessive	0.1094	0.1386	0.3790459	0.6186729	1.0097885	0.1069
		AACA			0.2783766	0.5365995	1.0343506	0.1187
		AACC			0.3790459	0.6186729	1.0097885	0.1069
		AGAA			0.7374653	0.8762802	1.0412245	0.2078
		AGCA			0.7008870	0.7739858	0.8547084	2.158e-05
		GGAA			1.17869	1.285239	1.40142	1.846e-06
3	rs4144827	Dominant	9.028e-06	1.04e-05	1.153841	1.296139	1.456627	9.028e-06
		Additive	3.068e-05	3.353e-05				
		Recessive	0.1103	0.1149	1.017672	1.673539	2.752098	0.08861
	rs4144827, rs4602520	Dominant	8.961e-06	1.328e-05	1.442503	1.800931	2.248420	1.298e-05
		Recessive	NA	NA	NA	NA	NA	NA
		AAAA			0.6917603	0.7442101	0.8006367	2.949e-11
		AAGA			1.150461	1.267872	1.397265	5.887e-05
		AAGG			0.9329632	1.376581	2.0311349	0.1766
		GAAA			1.052304	1.170382	1.301711	0.01496
		GAGG			0.438461	0.9155752	1.911864	0.8438
		GGAA			0.9034239	1.57224	2.7361880	0.1792

		GAGA			1.501587	1.906013	2.419364	8.643e-06
4	1q:44	Dominant	1.19e-06	1.217e-06	0.7102	0.7743	0.8441	1.11e-06
		Additive	3.822e-06	3.946e-06				
		Recessive	0.05692	0.07004	0.4223340	0.6290329	0.9368946	0.05562
	rs3924215	Dominant	2.007e-08	2.073e-08	0.7235	0.7785	0.8377	1.893e-08
		Additive	1.33e-07	1.322e-07				
		Recessive	0.2272	0.2558	0.6492069	0.8326492	1.0679257	0.2261
	1q:44, rs3924215	Dominant	1.513e-07	1.92e-07	0.4953	0.5863	0.6942	1.973e-07
		Recessive	NA	NA	NA	NA	NA	NA
		AAAA			1.2201	1.3027	1.3908	3.11e-11
		AAGA			0.7841644	0.849142	0.9195039	0.000728
		AAGG			0.676286	0.8879441	1.165845	0.4728
		GAAA			0.8010246	0.8838137	0.9751593	0.03887
		GAGG			0.3462324	0.6442479	1.1987765	0.2442
		GGAA			0.3875557	0.6175804	0.9841312	0.08888
		GAGA			0.4854	0.5807	0.6949	6.284e-07

EXTENDED VERSION OF TABLE 6.9

Combination	Variant/s	Model/ Comb	Results					
			F	X^2	CI <	OR	> CI	OR P
8		Dominant	1.355e-06	3.148e-06	2.180334	3.421732	5.369932	7.129e-06
		Recessive	NA	NA	NA	NA	NA	NA
		AAGGCC			0.4823589	0.823883	1.4072159	0.5517
		AGAACC			0.6963769	1.244028	2.2223662	0.5359
		AGAGAC			2.268098	3.938433	6.838879	4.392e-05
		AGAGCC			1.004063	1.22784	1.501492	0.09335
		AGGGAC			0.8959171	1.150594	1.4776663	0.3564
		AGGGCC			1.096949	1.229665	1.378439	0.002906
		GGAACC			1.003834	1.309822	1.709081	0.09521
		GGAGAC			1.000959	1.201448	1.442094	0.09823
		GGAGCC			1.007230	1.100851	1.203175	0.07537
		GGGGAA			0.9725669	1.491969	2.2887600	0.1241
		GGGGAC			1.031120	1.142969	1.266952	0.03282
		GGGGCC			0.6984741	0.7444199	0.7933880	2.531e-14
9	rs3924215	Dominant	2.357e-08	2.458e-08	0.7227262	0.7780847	0.8376835	2.243e-08
		Additive	1.532e-07	1.562e-07				
		Recessive	0.2236	0.2302	0.6412022	0.8236716	1.0580669	0.2026
	rs6011609	Dominant	2.643e-08	3.415e-08	0.4104905	0.5041824	0.6192589	2.643e-08
		Additive	3.201e-08	1.607e-07				
		Recessive	NA	NA	NA	NA	NA	NA
	rs3924215, rs6011609	Dominant	4.233e-05	6.43e-05	0.2292273	0.3537798	0.5460090	8.201e-05
		Recessive	NA	NA	NA	NA	NA	NA
		AAAG			0.4477369	0.5661758	0.7159451	6.7e-05
		GAAG			0.2376908	0.3721416	0.5826450	0.000287
		GAGG			0.7455804	0.8046655	0.8684329	2.768e-06
		GGGG			0.6672929	0.8606741	1.1100971	0.3322
		AAGG			1.248754	1.341981	1.442168	1.82e-11
10	1p12	Dominant	1.697e-08	2.451e-08	0.4018188	0.4952112	0.6103102	3.178e-08
		Additive	2.428e-08	1.052e-07				
		Recessive	NA	NA	NA	NA	NA	NA
	1p12, rs6011609	Dominant	0.0005024	0.001227	0.0335983	0.1154139	0.3964597	0.004002
		Recessive	NA	NA	NA	NA	NA	NA
		AAAG			0.4378091	0.5407107	0.6677981	1.66e-06
		AAGG			1.689720	1.963191	2.280922	1.397e-13
		GAGG			0.4266089	0.5292943	0.6566962	1.222e-06
11	rs6911024	Dominant	4.66e-07	5.331e-07	1.167226	1.258262	1.356397	4.865e-07
		Additive	1.688e-06	1.751e-06				
		Recessive	0.01747	0.0205	1.103029	1.376314	1.717308	0.01762
	rs6911024, rs7246472	Dominant	1.6e-05	2.151e-05	1.294671	1.5222	1.789717	1.967e-05
		Recessive	NA	NA	NA	NA	NA	NA
		AAAA			0.3480120	0.5344931	0.8208995	0.01633
		AAAC			0.7941352	0.8744423	0.9628705	0.02197
		AACC			1.210169	1.292758	1.380983	1.577e-10
		GAAC			0.5920245	0.7028669	0.8344619	0.0007267
		GGAC			0.3022718	0.5271586	0.9193585	0.05829
		GGCC			0.6016892	0.7664424	0.9763079	0.07064
		GACC			0.8012848	0.8711706	0.9471516	0.00667
		Dominant	1.541e-05	1.918e-05	1.310632	1.550424	1.834087	1.762e-05

	rs4602520, rs6911024	Recessive	NA	NA	NA	NA	NA	NA
		AAAA			1.250170	1.336547	1.428893	9.204e-13
		AAGA			0.8054270	0.8752244	0.9510704	0.008346
		AAGG			0.5739463	0.7305342	0.9298433	0.03229
		GAAA			0.7015318	0.7778493	0.8624692	6.293e-05
		GAGG			0.4139599	0.7308316	1.2902573	0.3642
		GGAA			0.6005097	0.9087141	1.3751007	0.7039
		GGGA			0.2894733	0.5477948	1.0366385	0.1206
		GAGA			0.5417441	0.6503037	0.7806174	0.0001065
5	rs6911024	Dominant	1.286e-06	1.39e-06	1.1567	1.2465	1.3434	1.269e-06
		Additive	4.521e-06	4.614e-06				
		Recessive	0.02489	0.02596	1.089573	1.35857	1.693978	0.02235
	rs12170250	Dominant	1.245e-06	1.806e-06	1.4585	1.7789	2.1697	1.839e-06
		Additive	2.122e-06	8.532e-06				
		Recessive	NA	NA	NA	NA	NA	NA
	rs6911024, rs12170250	Dominant	1.595e-07	3.728e-07	2.0276	2.8836	4.1010	7.567e-07
		Recessive	NA	NA	NA	NA	NA	NA
		AAAG			1.046805	1.335801	1.704580	0.05076
		AAGG			0.7334	0.7889	0.8487	9.29e-08
		GAGG			1.076066	1.164065	1.259261	0.001478
		GGGG			1.002669	1.257392	1.576825	0.09607
		GAAG			1.8680	2.7163	3.9500	1.135e-05
6	rs6852865	Dominant	7.401e-07	8.642e-07	1.1994	1.3135	1.4383	7.897e-07
		Additive	3.848e-06	4.365e-06				
		Recessive	0.641	0.662	0.7740845	1.138723	1.6751274	0.5799
	rs6852865, rs4602520	Dominant	2.782e-07	3.072e-07	1.2304	1.3567	1.4959	2.826e-07
		Recessive	0.3825	0.4377	0.827657	1.256559	1.907722	0.3683
		AAGG			0.827657	1.256559	1.907722	0.3683
		AGAA			0.8449817	1.048816	1.3018211	0.7168
		AGGA			1.2360	1.3677	0.5135	3.638e-07
		AGGG			0.6530297	1.198081	2.1980581	0.6242
		GGGA			1.076322	1.298829	1.567334	0.0221
		GGAA			0.6890	0.7494	0.8151	1.653e-08
7		Dominant	1.753e-06	5.453e-06	2.303162	3.832153	6.376189	1.424e-05
		Recessive	NA	NA	NA	NA	NA	NA
		AAGGGGCC			0.4855043	0.861772	1.5296485	0.6698
		AGAAAGCC			0.7311888	1.222341	2.0434073	0.5204
		AGAAGGAC			0.6097411	1.127805	2.0860405	0.7477
		AGAAGGCC			0.6838781	0.9020966	1.1899465	0.5406
		AGGAAACC			0.7020486	1.375066	2.6932718	0.4358
		AGGAAGCC			0.9834412	1.226686	1.5300950	0.1284
		AGGAGGAC			0.8923749	1.176241	1.5504048	0.3337
		AGGAGGCC			1.140013	1.292075	1.464419	0.0007618
		GGAAAACC			1.021367	1.336666	1.749299	0.07604
		GGAAAGAC			0.977733	1.178044	1.419393	0.1481
		GGAAAGCC			0.9918097	1.085478	1.1879919	0.1349
		GGAAGGAA			1.019739	1.586202	2.467334	0.08586
		GGAAGGAC			1.034578	1.14845	1.274856	0.02923
		GGAAGGCC			0.6849178	0.7298533	0.7777368	4.441e-16
		GGGAAGCC			0.8969771	1.396439	2.1740138	0.2147
		GGGAGGAC			0.5555386	0.9698373	1.6931037	0.928
		GGGAGGCC			1.051960	1.328002	1.676479	0.04524

APPENDIX F

CONTINGENCY TABLE VALUES FOR EXTENDED TABLE 6.8

ID	Variant/s	Model/ Comb	Results			
			Control|Exposed	Case|Exposed	Control|NotExposed	Case|NotExposed
1	rs4602520	Dominant	4500	4684	683	974
		Recessive	41	57	5142	5601
	rs6910087	Dominant	4055	4206	1128	1452
		Recessive	94	142	5089	5516
	rs7246472	Dominant	4384	4587	799	1071
		Recessive	27	60	5156	5598
	rs4602520 rs6910087	Dominant	5020	5390	163	268
		Recessive	5182	5656	1	2
		AAAA	79	119	5104	5539
		AAAG	886	1065	4297	4593
		AAGG	3535	3500	1648	2158
		GAAG	137	225	5046	5433
		GAAA	14	21	5169	5637
		GGAG	11	20	5172	5638
		GGGG	29	35	5154	5623
		GAGG	491	671	4692	4987
	rs4602520 rs7246472	Dominant	5078	5483	105	175
		Recessive	0	0	5183	5658
		AAAA	25	51	5158	5607
		AAAC	669	845	4514	4813
		AACC	3806	3788	1377	1870
		GAAC	98	157	5085	5501
		GACC	542	751	4641	4907
		GGCC	36	48	5147	5610
	rs6910087 rs7246472	Dominant	4389	4596	794	1062
		Recessive	1	0	5182	565813
		AAAC	13	28	5170	5630
		AACC	80	114	5103	5544
		AGAC	161	243	5022	5415
		AGCC	871	1052	4312	4606
		GGAA	24	45	5159	5613
		GGAC	598	740	4585	4918
		GGCC	3433	3421	1750	2237
	rs4602520 rs691008 rs7246472	Dominant	18	64	5165	5593
		Recessive	0	0	5183	5658
		AAAAAC	10	22	5173	5636
		AAAACC	68	97	5115	5561
		AAAGAC	146	189	5037	5469
		AAAGCC	738	866	4445	4792
		AAGGAA	22	41	5161	5617
		AAGGAC	513	634	4670	5024
		AAGGCC	3000	2825	2183	2833
		GAAACC	12	16	5171	5642
		GAAGAC	15	52	5168	5606
		GAAGCC	122	168	5061	5490
		GAGGAC	81	100	5102	5558
		GGAGCC	11	18	5172	5640
		GGGGCC	25	29	5158	5629
		GAGGCC	408	567	4775	5091
2	9q21.13	Dominant	4307	4906	906	800
		Recessive	48	34	5165	5672
	9q21.13	Dominant	4495	5087	718	619
		Recessive	28	19	5185	5687
	9q21.13 9q21.13	Dominant	4502	5096	711	610
		Recessive	28	19	5185	5687
		AACA	17	10	5196	5696
		AACC	28	19	5185	5687
		AGAA	192	185	5021	5521
		AGCA	666	581	4547	5125
		GGAA	4300	4897	913	809
3	rs4144827	Dominant	4608	4872	575	788
		Recessive	17	31	5166	5629
	rs4144827 rs4602520	Dominant	5098	5495	85	165
		Recessive	1	2	5182	5658
		AAAA	4010	4063	1173	1597
		AAGA	568	764	4615	4896
		AAGG	30	45	5153	5615
		GAAA	476	599	4707	5061
		GAGG	10	10	5173	5650
		GGAA	14	24	5169	5636
		GAGA	72	148	5111	5512

4	1q:44	Dominant	4293	4893	919	811
		Recessive	42	29	5170	5675
	rs3924215	Dominant	1407	1275	3805	4429
		Recessive	93	85	5119	5619
	1q:44 rs3924215	Dominant	4963	5541	249	163
		Recessive	1	0	5211	5704
		AAAA	3135	3781	2077	1923
		AAGA	1083	1039	4129	4665
		AAGG	75	73	5137	5631
		GAAA	639	627	4573	5077
		GAGG	17	12	5195	5692
		GGAA	31	21	5181	5683
		GAGA	221	143	4991	5561

CONTINGENCY TABLE VALUES FOR EXTENDED TABLE 6.9

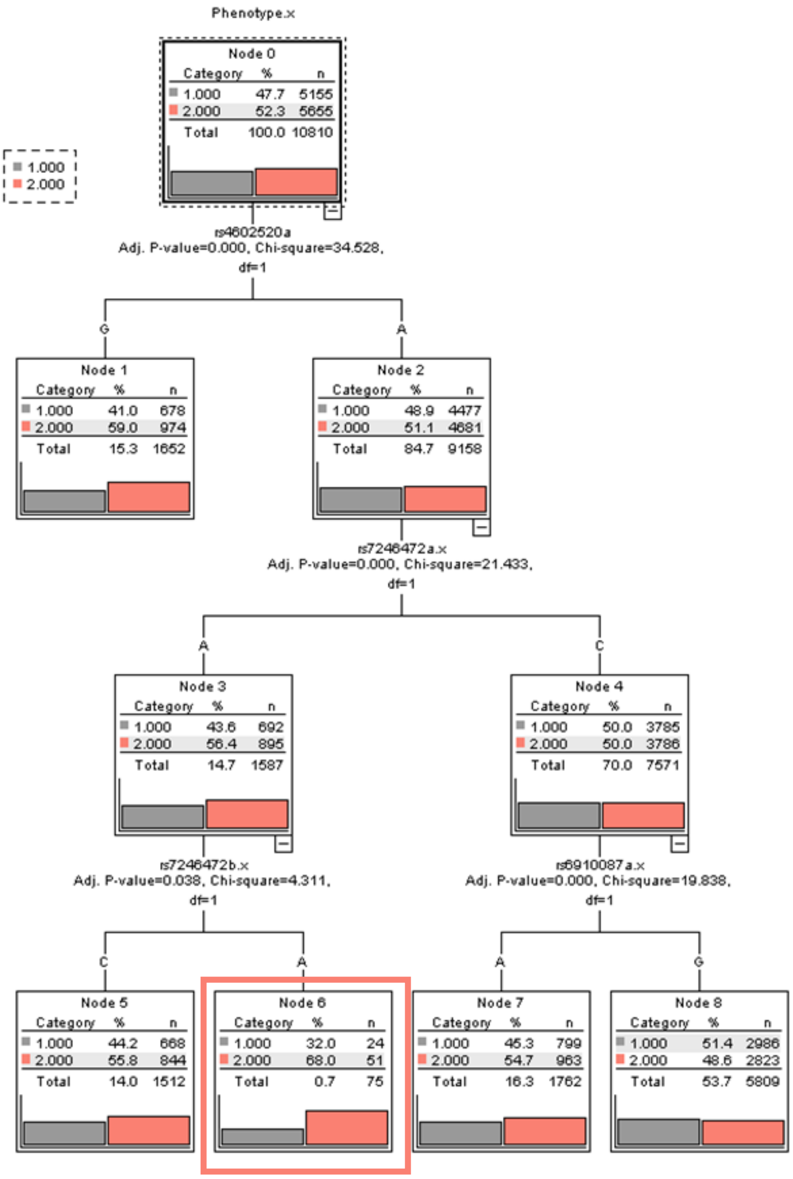
Combination	Variant/s	Model/ Comb	Results			
			Control|Exposed	Case|Exposed	Control|NotExposed	Case|NotExposed
5	rs6911024	Dominant	4075	4256	1113	1449
		Recessive	95	141	5093	5564
	rs12170250	Dominant	5081	5499	107	206
		Recessive	1	2	5187	5703
	rs6911024 rs12170250	Dominant	5159	5614	29	91
		Recessive	0	0	5188	5705
		AAAG	78	114	5110	5591
		AAGG	3997	4141	1191	1564
		GAGG	991	1230	4197	4475
		GGGG	93	128	5095	5577
		GAAG	26	77	5162	5628
6	rs6852865	Dominant	4524	4751	659	909
		Recessive	33	41	5150	5619
	rs6852865 rs4602520	Dominant	4636	4879	547	781
		Recessive	27	37	5156	5623
		AAGG	27	37	5156	5623
		AGAA	111	127	5072	5533
		AGGA	502	724	4681	4936
		AGGG	13	17	5170	5643
		GGGA	135	190	5048	5470
		GGAA	4388	4558	795	1102
7	rs6852865 rs4602520 rs6910087 rs7246472	Dominant	5170	5604	13	54
		Recessive	0	0	5183	5658
		AAGGGGCC	17	16	5166	5642
		AGAAAGCC	18	24	5165	5634
		AGAAGGAC	13	16	5170	5642
		AGAAGGCC	72	71	5111	5587
		AGGAAACC	10	15	5173	5643
		AGGAAGCC	99	132	5084	5526
		AGGAGGAC	64	82	5119	5576
		AGGAGGCC	319	442	4864	5216
		GGAAAACC	64	93	5119	5565
		GGAAAGAC	143	183	5040	5475
		GGAAAGCC	719	842	4464	4816
		GGAAGGAA	22	38	5161	5620
		GGAAGGAC	500	618	4683	5040
		GGAAGGCC	2928	2753	2255	2905
		GGGAAGCC	23	35	5160	5623
		GGGAGGAC	17	18	5166	5640
		GGGAGGCC	86	124	5097	5534
8		Dominant	5166	5595	17	63
		Recessive	0	0	5183	5658
		AAGGCC	20	18	5163	5640
		AGAACC	14	19	5169	5639
		AGAGAC	11	47	5172	5611
		AGAGCC	120	160	5063	5498
		AGGGAC	79	99	5104	5559
		AGGGCC	398	525	4785	5133
		GGAACC	66	94	5117	5564
		GGAGAC	148	193	5035	5465
		GGAGCC	742	879	4441	4779
		GGGGAA	24	39	5159	5619
		GGGGAC	517	636	4666	5022
		GGGGCC	3015	2878	2168	2780
9	rs3924215	Dominant	3773	4335	1396	1248
		Recessive	93	83	5076	5500
	rs6011609	Dominant	184	102	4985	5481
		Recessive	1	1	5168	5582

	rs3924215 rs6011609	Dominant	52	20	5117	5563
		Recessive	0	0	5169	5583
		AAAG	131	81	5038	5502
		GAAG	47	19	5122	5564
		GAGG	1256	1146	3913	4437
		GGGG	88	82	5081	5501
		AAGG	3641	4253	1528	1330
10	1p12	Dominant	180	98	4989	5485
		Recessive	2	2	5167	5581
	1p12 rs6011609	Dominant	16	2	5153	5581
		Recessive	0	0	5169	5583
		AAAG	167	99	5002	5484
		AAGG	4821	5385	348	198
		GAGG	162	94	5007	5489
11	rs6911024	Dominant	4055	4217	1102	1442
		Recessive	94	141	5063	5518
	rs6911024 rs7246472	Dominant	173	284	4984	5375
		Recessive	1	0	5156	5659
		AAAA	22	45	5135	5614
		AAAC	602	743	4555	4916
		AACC	1726	2230	3431	3429
		GAAC	157	242	5000	5417
		GGAC	13	27	5144	5632
		GGCC	80	114	5077	5545
		GACC	849	1044	4308	4615
12	rs4602520 rs6911024	Dominant	4998	5393	159	266
		Recessive	1	2	5156	5657
		AAAA	3536	3509	1621	2150
		AAGA	864	1058	4293	4601
		AAGG	79	118	5078	5541
		GAAA	490	673	4667	490
		GAGG	14	21	5143	5638
		GGAA	29	35	5128	5624
		GGGA	10	20	5147	5639
		GAGA	134	223	5023	5436

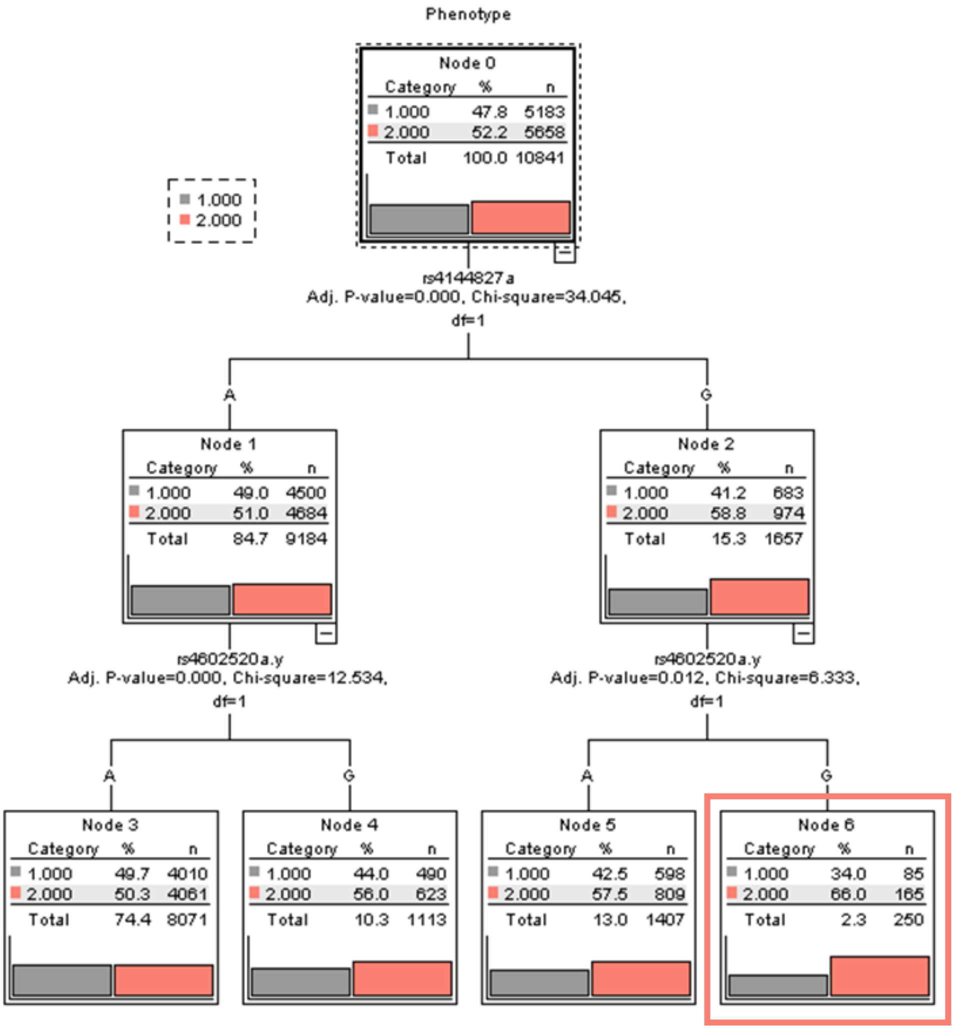
APPENDIX G

Decision Trees Full Extension for all combinations that yielded significant results.

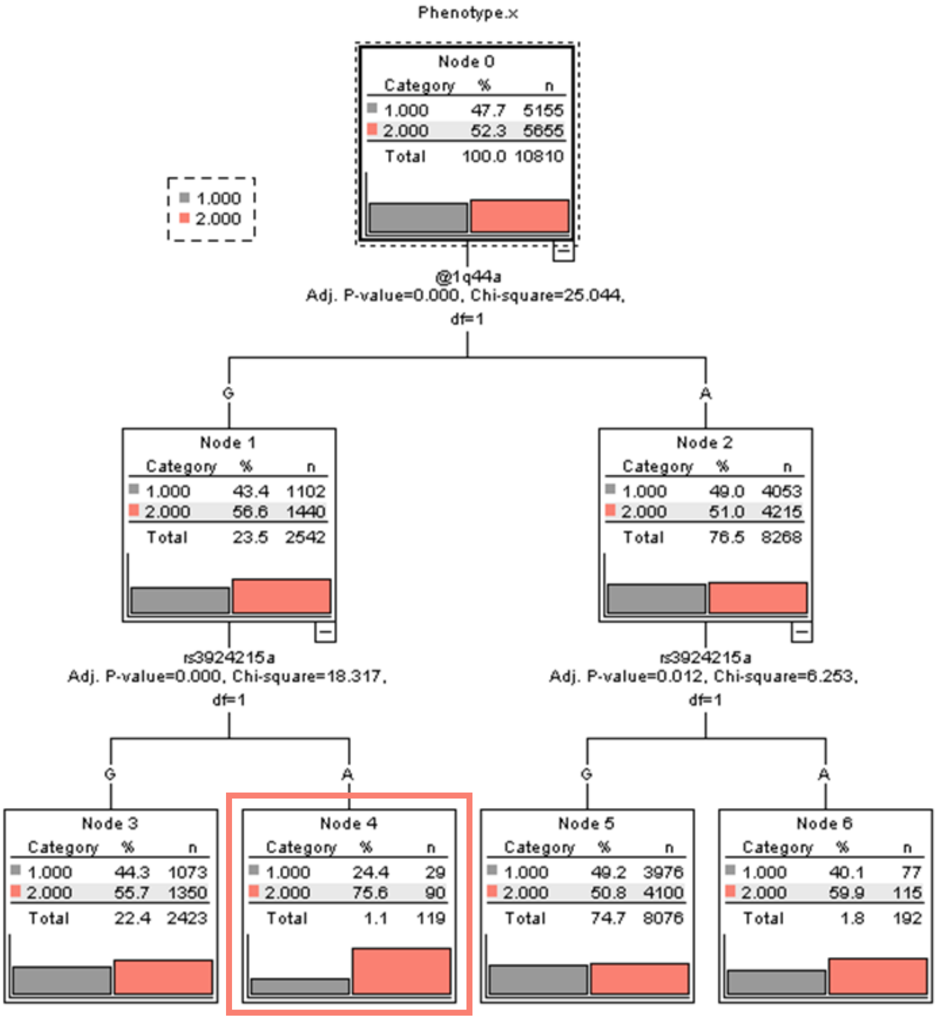
COMBINATION 1



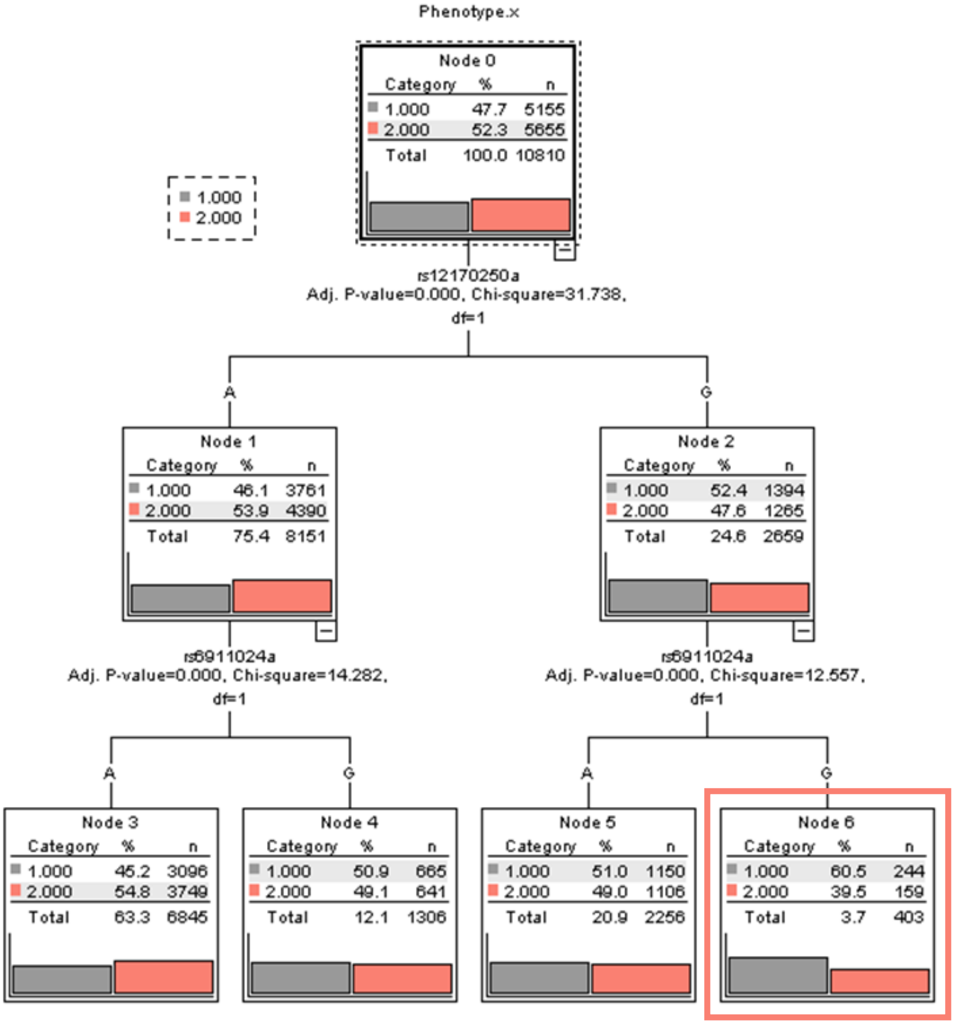
COMBINATION 3



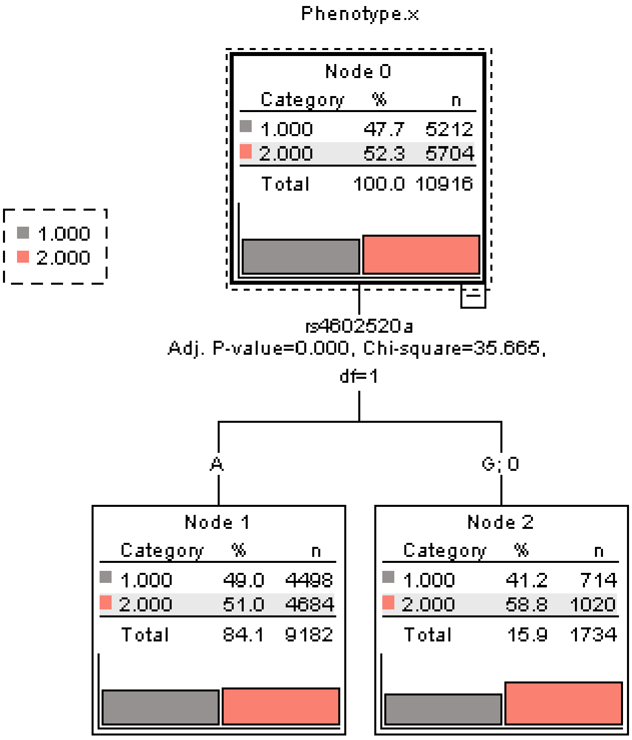
COMBINATION 4



COMBINATION 5



COMBINATION 6



Node 0

Category	%	n
■ 1.000	47.7	5212
🟥 2.000	52.3	5704
Total	100.0	10916

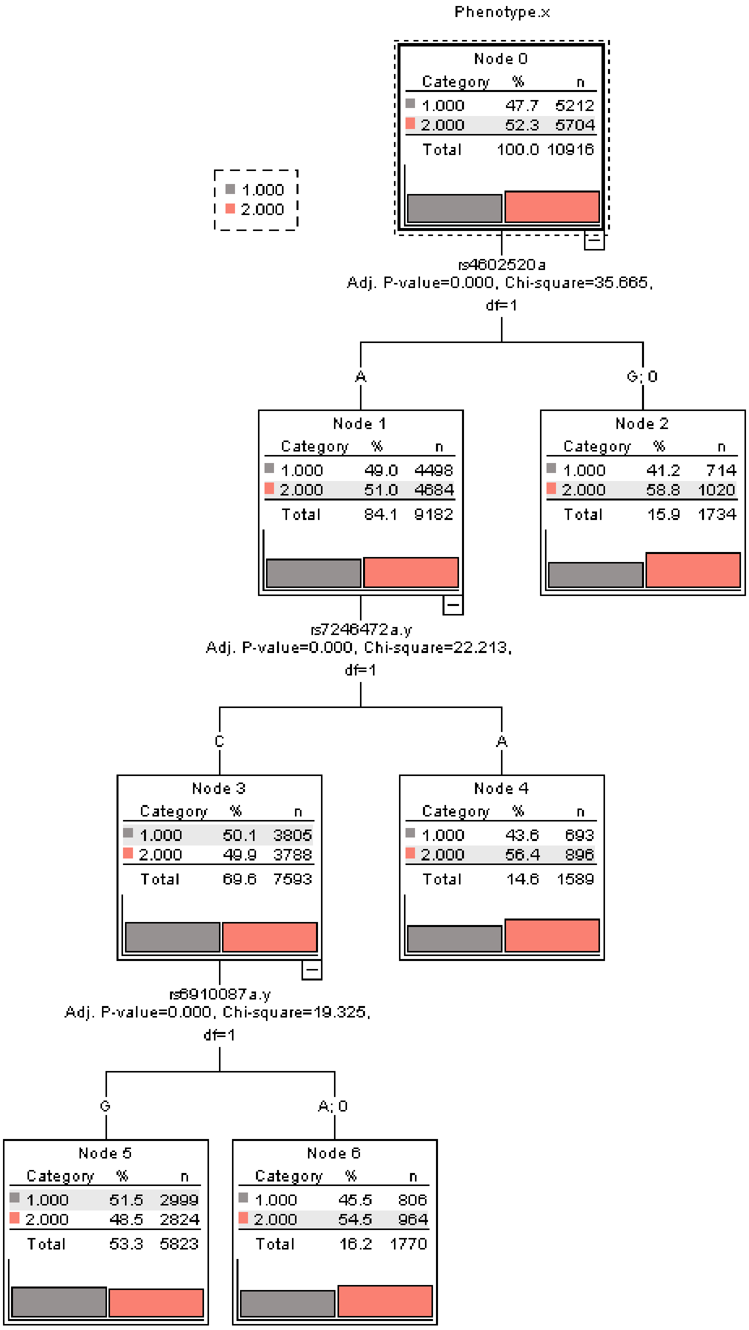
Node 1

Category	%	n
■ 1.000	49.0	4498
🟥 2.000	51.0	4684
Total	84.1	9182

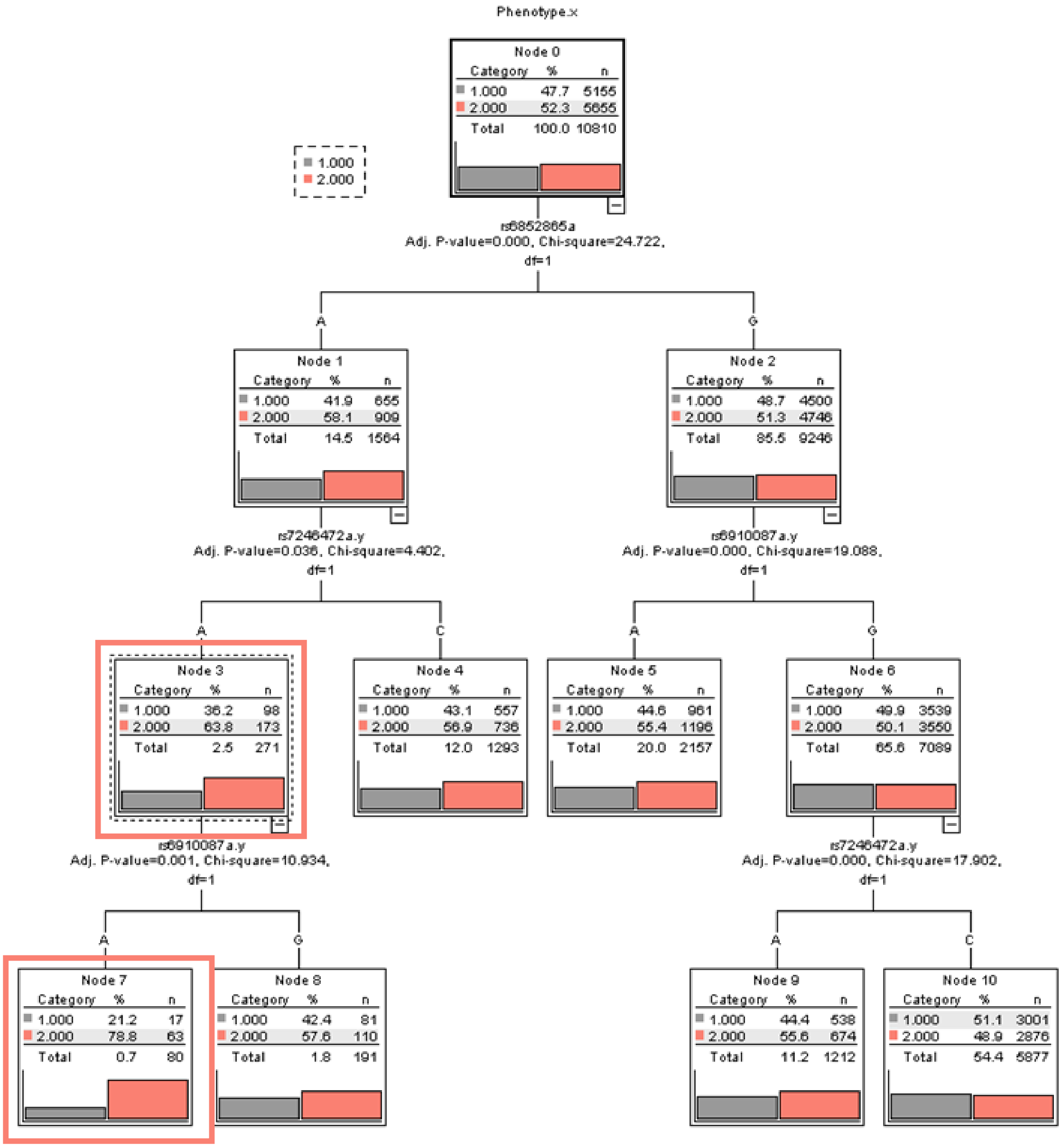
Node 2

Category	%	n
■ 1.000	41.2	714
🟥 2.000	58.8	1020
Total	15.9	1734

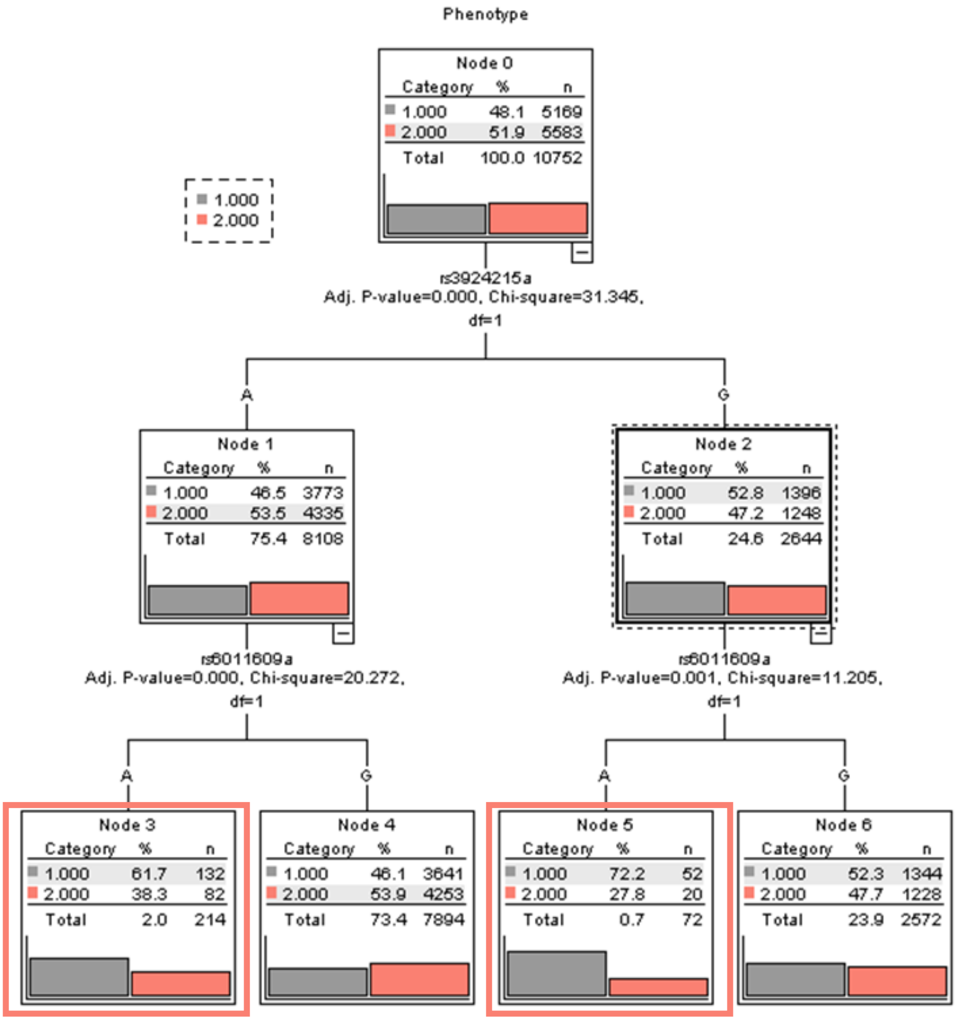
COMBINATION 7



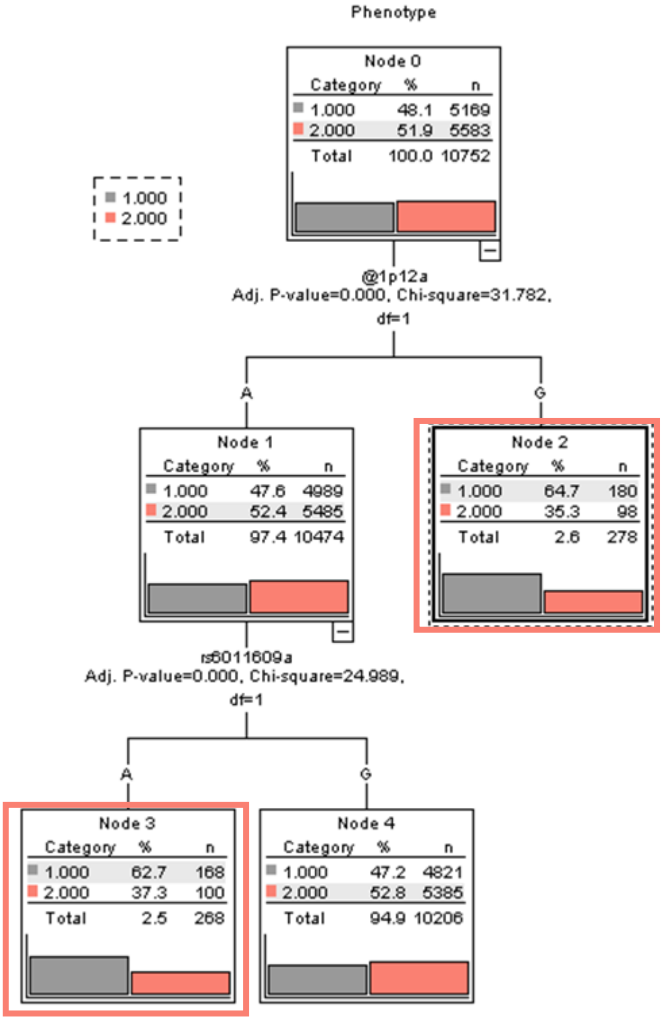
COMBINATION 8



COMBINATION 9



COMBINATION 10


Phenotype
▨ 1.000
🟥 2.000
@1p12a
Adj. P-value=0.000, Chi-square=31.782, df=1
A
G
rs6011609a
Adj. P-value=0.000, Chi-square=24.989, df=1
A
G

Node 0

Category	%	n
▨ 1.000	48.1	5169
🟥 2.000	51.9	5583
Total	100.0	10752

Node 1

Category	%	n
▨ 1.000	47.6	4989
🟥 2.000	52.4	5485
Total	97.4	10474

Node 2

Category	%	n
▨ 1.000	64.7	180
🟥 2.000	35.3	98
Total	2.6	278

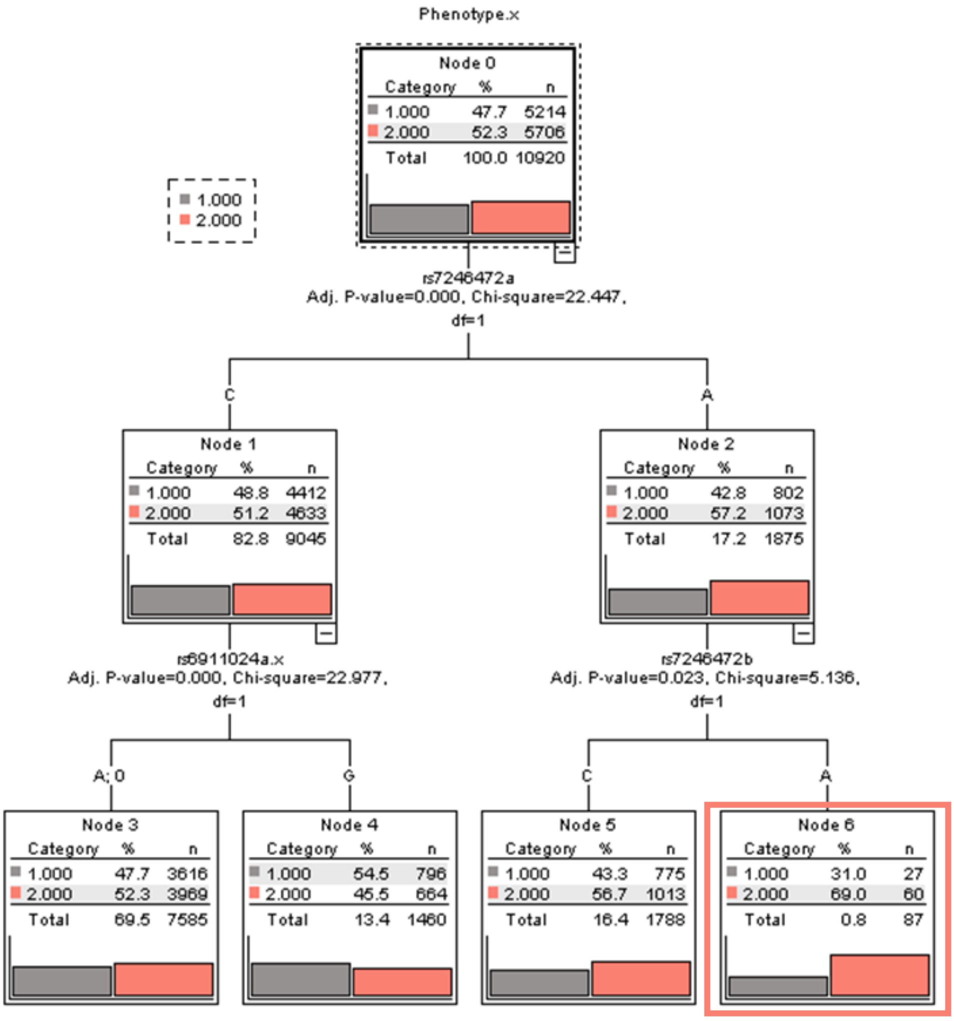
Node 3

Category	%	n
▨ 1.000	62.7	168
🟥 2.000	37.3	100
Total	2.5	268

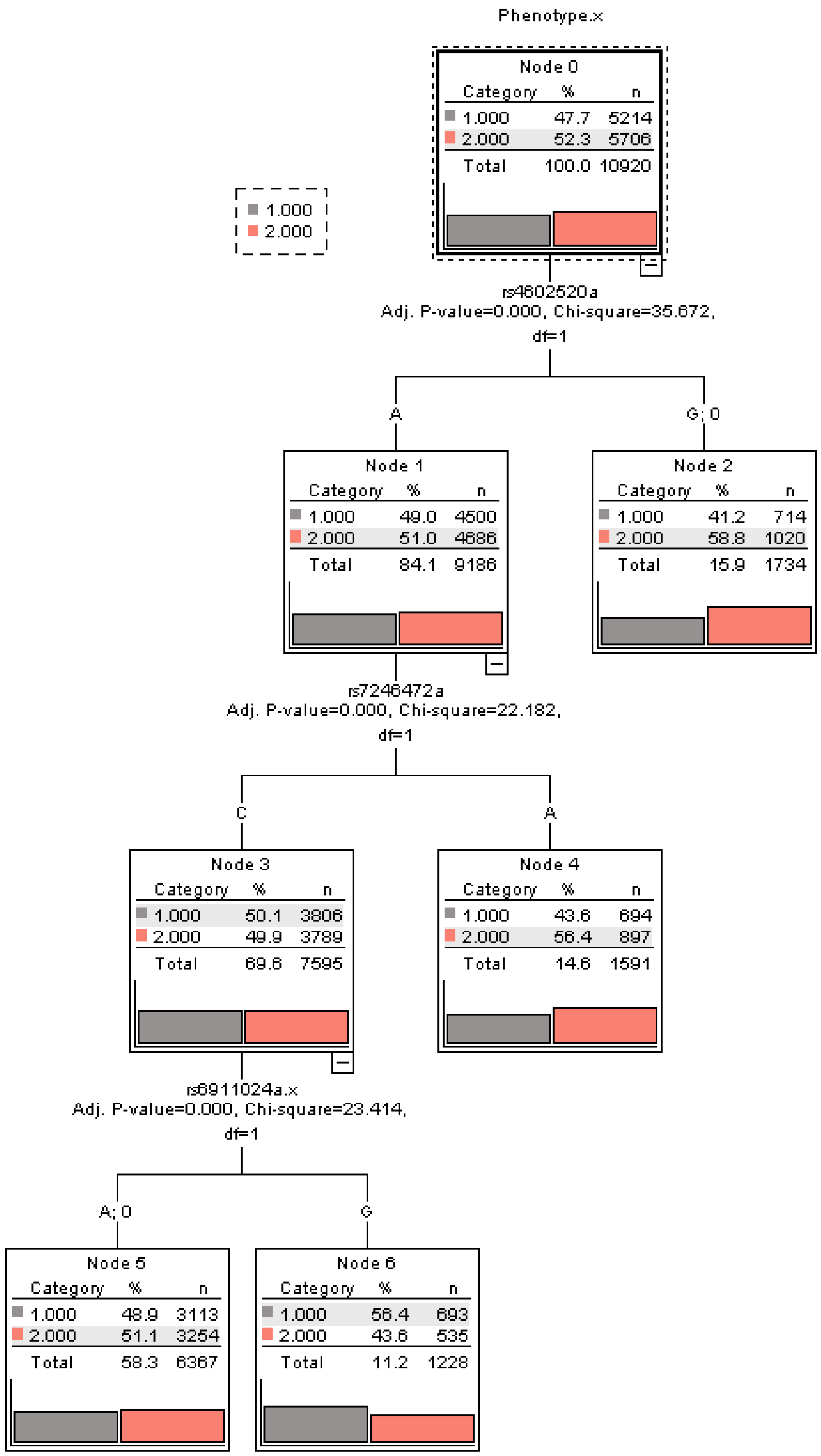
Node 4

Category	%	n
▨ 1.000	47.2	4821
🟥 2.000	52.8	5385
Total	94.9	10206

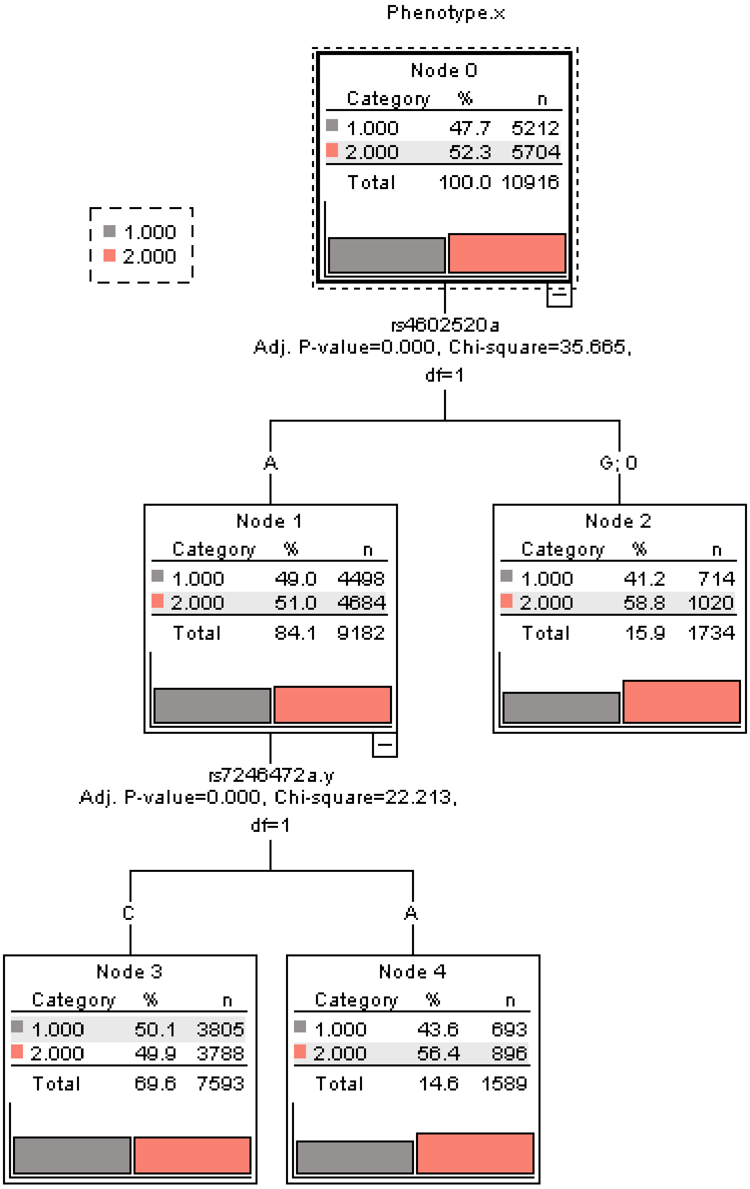
COMBINATION 11



COMBINATION 12



COMBINATION 13


Phenotype.x
■ 1.000
■ 2.000
rs4602520a
Adj. P-value=0.000, Chi-square=35.665,
df=1
A
G; 0
rs7246472a.y
Adj. P-value=0.000, Chi-square=22.213,
df=1
C
A

Node 0

Category	%	n
■ 1.000	47.7	5212
■ 2.000	52.3	5704
Total	100.0	10916

Node 1

Category	%	n
■ 1.000	49.0	4498
■ 2.000	51.0	4684
Total	84.1	9182

Node 2

Category	%	n
■ 1.000	41.2	714
■ 2.000	58.8	1020
Total	15.9	1734

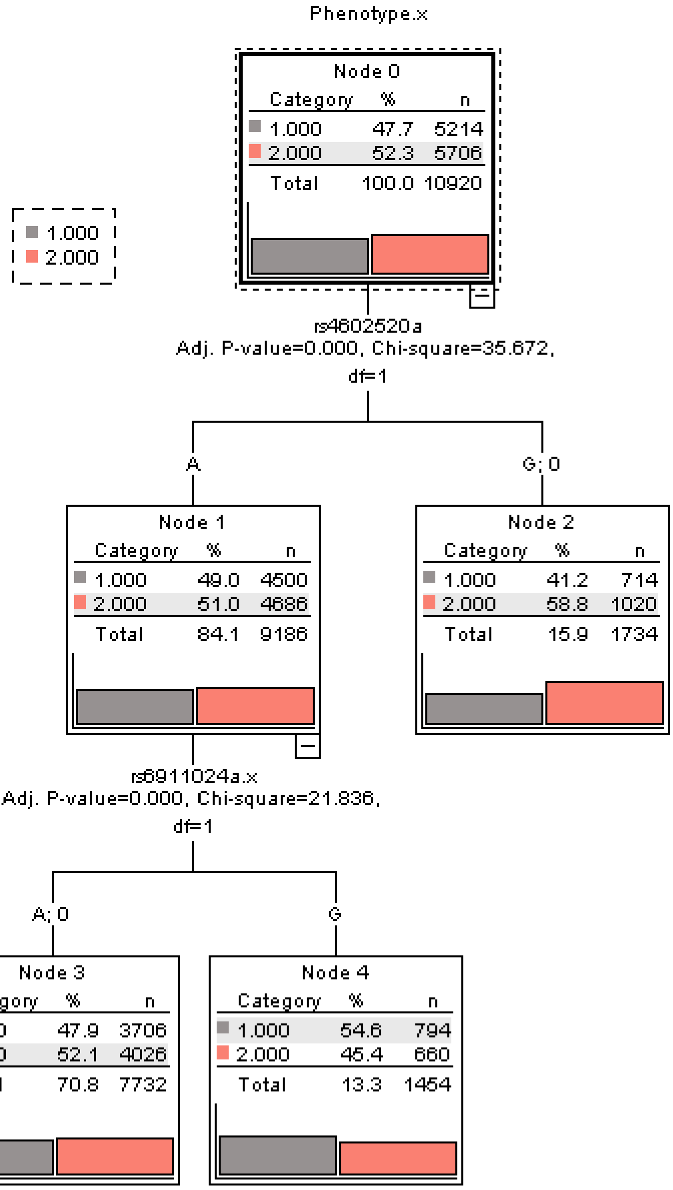
Node 3

Category	%	n
■ 1.000	50.1	3805
■ 2.000	49.9	3788
Total	69.6	7593

Node 4

Category	%	n
■ 1.000	43.6	693
■ 2.000	56.4	896
Total	14.6	1589

COMBINATION 14



Publication List

J. Hind, A. Hussain, D. Al-jumeily, B. Abdulaimma, C. A. C. Montañez and P. Lisboa, "A robust method for the interpretation of genomic data," *2017 International Joint Conference on Neural Networks (IJCNN)*, Anchorage, AK, 2017, pp. 3385-3390.

J. Hind, A. Hussain, D. Al-jumeily, B. Abdulaimma, C. A. C. Montañez, C. Chalmers and P. Lisboa, "Robust Interpretation of Genomic Data in Chronic Obstructive Pulmonary Disease (COPD)," 2018 International Conference on Developments in eSystems Engineering (DeSE), Cambridge, UK, 2018. *Accepted awaiting publication*

J. Hind, A. Hussain, D. Al-jumeily, B. Abdulaimma, C. A. C. Montañez, C. Chalmers and P. Lisboa, "Random Sampling Regularisation: A Proposed Methodology for Detection of Epistasis Interactions in Genomic Studies, " Computational Intelligence Methods for Bioinformatics and Biostatistics - 15th International Meeting, CIBB 2018, Caparica, Portugal, September 6-8, 2018. *Accepted awaiting publication*

Bibliography

[1] K. Hemminki *et al.*, "The 'Common Disease-Common Variant' Hypothesis and Familial Risks," *PLoS One*, vol. 3, no. 6, p. e2504, Jun. 2008.

[2] K. N. Nathanson *et al.*, "Breast cancer genetics: What we know and what we need," *Nat. Med.*, vol. 7, no. 5, pp. 552–556, May 2001.

[3] UKTRIALOFEARLYDETECTIONOFBRE, "FIRST RESULTS ON MORTALITY REDUCTION IN THE UK TRIAL OF EARLY DETECTION OF BREAST CANCER," *Lancet*, vol. 332, no. 8608, Aug. 1988.

[4] J. D. Wulfkuhle *et al.*, "Proteomic applications for the early detection of cancer," *Nat. Rev. Cancer*, vol. 3, no. 4, pp. 267–275, Apr. 2003.

[5] S. Shiovitz *et al.*, "Genetics of breast cancer: A topic in evolution," *Ann. Oncol.*, vol. 26, no. 7, pp. 1291–1299, 2015.

[6] S. E. Jackson *et al.*, "Personalised cancer medicine," *Int. J. Cancer*, vol. 137, no. 2, pp. 262–266, 2015.

[7] E. Graham, "Improving Outcomes Through Personalised Medicine," *NHS Engl.*, pp. 6–10, 2016.

[8] K. JL, "A review of the epidemiology of human breast cancer.," *Epidemiol. Rev.*, vol. 1, no. July, pp. 74–109, 1979.

[9] J. Czernin *et al.*, "Breast cancer.," *Methods Mol. Biol.*, vol. 727, no. October, pp. 141–170, 2011.

[10] "Breast cancer statistics," *Cancer Research, UK.*, 2018. [Online]. Available: http://www.cancerresearchuk.org/health-professiona. [Accessed: 01-May-2018].

[11] Ncbi.nlm.nih.gov., "dbGaP | phs001265.v1.p1 | Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE) - OncoArray Genotypes," 2018. [Online]. Available: https://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs001265.v1.p1. [Accessed: 23-Nov-2018].

[12] G. Montana, "Statistical methods in genetics," *Brief. Bioinform.*, vol. 7, no. 3, pp. 297–308, 2006.

[13] C. Niel *et al.*, "A survey about methods dedicated to epistasis detection," vol. 6, no. September, 2015.

[14] J. D. Storey *et al.*, "Statistical significance for genomewide studies," *Proc. Natl. Acad. Sci.*, vol. 100, no. 16,

pp. 9440–9445, 2003.

[15] D. J. Balding, “A tutorial on statistical methods for population association studies,” *Nat. Rev. Genet.*, vol. 7, no. 10, pp. 781–791, Oct. 2006.

[16] C. A. Rietveld *et al.*, “Replicability and Robustness of Genome-Wide-Association Studies for Behavioral Traits,” *Psychol. Sci.*, vol. 25, no. 11, pp. 1975–1986, Nov. 2014.

[17] R. Heller *et al.*, “Replicability analysis for genome-wide association studies,” *Ann. Appl. Stat.*, vol. 8, no. 1, pp. 481–498, Mar. 2014.

[18] P. Giusti-Rodríguez *et al.*, “The genomics of schizophrenia: update and implications,” *J. Clin. Invest.*, vol. 123, no. 11, pp. 4557–4563, Nov. 2013.

[19] M. Kanai *et al.*, “Empirical estimation of genome-wide significance thresholds based on the 1000 Genomes Project data set,” *J. Hum. Genet.*, vol. 61, no. 10, pp. 861–866, 2016.

[20] A. L. Price *et al.*, “New approaches to population stratification in genome-wide association studies,” *Nat. Rev. Genet.*, vol. 11, no. 7, pp. 459–463, Jul. 2010.

[21] M. Slatkin, “Linkage disequilibrium — understanding the evolutionary past and mapping the medical future,” *Nat. Rev. Genet.*, vol. 9, no. 6, pp. 477–485, Jun. 2008.

[22] U. M. Marigorta *et al.*, “Replicability and Prediction: Lessons and Challenges from GWAS,” *Trends Genet.*, vol. 34, no. 7, pp. 504–517, Jul. 2018.

[23] L. Cheng *et al.*, “Compositional epistasis detection using a few prototype disease models,” *PLoS One*, vol. 14, no. 3, p. e0213236, Mar. 2019.

[24] A. Korte *et al.*, “The advantages and limitations of trait analysis awith GWAS: a review,” *Plant Methods*, vol. 9, no. 29, pp. 1–9, 2013.

[25] E. Duncan *et al.*, “Unlocking the Genetics of Complex Diseases : THE GWAS and Beyond,” *Bioinformatics*, vol. 5, no. AUG, pp. 1–14, 2014.

[26] P. Phillips, “Epistasis—the essential role of gene interactions in the structure and evolution of genetic systems,” *Nat. Rev. Genet.*, vol. 9, no. 11, pp. 855–867, 2008.

[27] B. Goudey *et al.*, “GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS,” *BMC Genomics*, vol. 14, no. Suppl 3, p. S10, 2013.

[28] W. H. Wei *et al.*, “Detecting epistasis in human complex traits,” *Nat. Rev. Genet.*, vol. 15, no. 11, pp. 722–733, 2014.

[29] S.-H. Lo *et al.*, “Discovering interactions among BRCA1 and other candidate genes associated with sporadic breast cancer.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 105, no. 34, pp. 12387–92, 2008.

[30] C. Cybulski *et al.*, “Epistatic relationship between the cancer susceptibility genes CHEK2 and p27,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 16, no. 3, pp. 572–576, 2007.

[31] O. Anunciação *et al.*, “Using Information Interaction to Discover Epistatic Effects in Complex Diseases,” *PLoS One*, vol. 8, no. 10, 2013.

[32] Z. Q. Tao *et al.*, “Breast Cancer: Epidemiology and Etiology,” *Cell Biochem. Biophys.*, vol. 72, no. 2, pp. 333–338, 2015.

[33] P. C. Winter *et al.*, “Basic Mendelian Genetics,” in *Genetics*, 1998, pp. 119–126.

[34] P. C. Winter *et al.*, “Molecular Genetics,” in *Genetics*, BIOS Scientific Publishers Ltd, 1998, pp. 1–56.

[35] H. Lodish *et al.*, “Nucleic Acids, the Genetic Code, and the Synthesis of Macromolecules,” in *Molecular Cell Biology*, 4th ed., S. Tenney, Ed. W.H. Freeman and Company, 1999, pp. 100–137.

[36] H. Lodish *et al.*, “Protein Sorting Organelle Biogenesis and Protein Secretion,” in *Molecular Cell Biology*, 4th ed., W.H. Freeman and Company, 1999, pp. 675–750.

[37] H. Lodish *et al.*, “Genetic Analysis in Cell Biology,” in *Molecular Cell Biology*, 4th ed., S. Tenney, Ed. W.H. Freeman and Company, 1999, pp. 254–293.

[38] N. Saitou, “Introduction to Evolutionary Genomics,” vol. 17, 2013.

[39] P. C. Winter *et al.*, “DNA Mutation,” in *Genetics*, BIOS Scientific Publishers Ltd, 1998, pp. 86–91.

[40] P. Y. Kwok *et al.*, “Detection of single nucleotide polymorphisms.,” *Curr Issues Mol Biol*, vol. 5, no. 2, pp. 43–60, 2003.

[41] T. Robinson, *Genetics For Dummies*, 2nd Editio. John Wiley & Sons Ltd, 2010.

[42] Q. Najeeb *et al.*, “Personalized Medicine versus era of " Trial and Error " Abstract : Introduction : Personalized medicine : Pros and Cons,” vol. 19, no. 19, pp. 1–5, 2012.

[43] M. E. Lynch *et al.*, “‘One Size Fits All’ Doesn’t Fit When It Comes to Long-Term Opioid Use for People with Chronic Pain,” *Can. J. Pain*, vol. 1, no. 1, pp. 2–7, 2017.

[44] M. Cloitre, “The ‘“one size fits all”’ approach to trauma treatment: Should we be satisfied?,” *Eur. J. Psychotraumatol.*, vol. 6, no. May, 2015.

[45] G. Kang *et al.*, “Gene-based Genomewide Association Analysis: A Comparison Study,” *Curr. Genomics*, vol. 14, no. 4, pp. 250–255, 2013.

[46] J. Yang *et al.*, “Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits,” *Nat. Genet.*, vol. 44, no. 4, pp. 369–375, 2012.

[47] X. Qu *et al.*, “Association between two CHRNA3 variants and susceptibility of lung cancer: a meta-analysis.,” *Sci. Rep.*, vol. 6, p. 20149, 2016.

[48] G. M. Clarke *et al.*, “Basic statistical analysis in genetic case-control studies,” *Nat. Protoc.*, vol. 6, no. 2, pp. 121–133, Feb. 2011.

[49] P. M. Visscher *et al.*, “10 Years of GWAS Discovery: Biology, Function, and Translation,” *Am. J. Hum. Genet.*, vol. 101, no. 1, pp. 5–22, 2017.

[50] T. A. Pearson, “How to Interpret a Genome-wide Association Study,” *JAMA*, vol. 299, no. 11, p. 1335, Mar. 2008.

[51] C. C. A. Spencer *et al.*, “Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip,” *PLoS Genet.*, vol. 5, no. 5, p. e1000477, May 2009.

[52] B. Verstockt *et al.*, “Genome-wide association studies in Crohn’s disease: Past, present and future,” *Clin. Transl. Immunol.*, vol. 7, no. 1, p. e1001, 2018.

[53] G. Orozco *et al.*, “Genetics of rheumatoid arthritis: GWAS and beyond,” *Open Access Rheumatol. Res. Rev.*, p. 31, Jun. 2011.

[54] V. Kumar *et al.*, “From genome-wide association studies to disease mechanisms : celiac disease as a model for autoimmune diseases,” pp. 567–580, 2012.

[55] P. Kraft *et al.*, “Genetic Risk Prediction — Are We There Yet?,” *N. Engl. J. Med.*, vol. 360, no. 17, pp. 1701–1703, Apr. 2009.

[56] J. Kwon *et al.*, “The candidate gene approach.,” *Alcohol Res Heal.*, vol. 24, no. 3, pp. 164–168, 2000.

[57] D. Rosmarin *et al.*, “A candidate gene study of capecitabine-related toxicity in colorectal cancer identifies new toxicity variants at DPYD and a putative role for ENOSF1 rather than TYMS,” *Gut*, vol. 64, no. 1, pp. 111–120, 2015.

[58] X. Zhao *et al.*, “Loci and candidate gene identification for resistance to Sclerotinia sclerotiorum in soybean (Glycine max L. Merr.) via association and linkage maps,” *Plant J.*, vol. 82, no. 2, pp. 245–255, 2015.

[59] A. Long, “Identification of Candidate Genes,” *Encyclopedia of Genetics*. 2001.

[60] A. G. Cardno *et al.*, “Twin studies of schizophrenia, from Bow and Arrow concordances to star wars. Mx and functional genomics.,” *AM J med Genet.*, vol. 90, no. 1, pp. 12–17, 2000.

[61] I. Lobo, “Epistasis: Gene interaction and the phenotypic expression of complex diseases like Alzheimer’s,” *Nat. Educ.*, vol. 1, no. 1, p. 180, 2008.

[62] S. Tripp *et al.*, “Economic Impact of the Human Genome Project,” *Battelle Meml. Inst.*, p. 58, 2011.

[63] D. M. Altshuler *et al.*, “Integrating common and rare genetic variation in diverse human populations,” *Nature*, vol. 467, no. 7311, pp. 52–58, 2010.

[64] A. Auton *et al.*, “A global reference for human genetic variation,” *Nature*, vol. 526, no. 7571, pp. 68–74, 2015.

[65] N. J. Schork *et al.*, “Common vs. rare allele hypotheses for complex diseases,” *Curr. Opin. Genet. Dev.*, vol. 19, no. 3, pp. 212–219, Jun. 2009.

[66] H. J. Cordell, “Epistasis: what it means, what it doesn’t mean, and statistical methods to detect it in humans,” *Hum. Mol. Genet.*, vol. 11, no. 20, pp. 2463–2468, 2002.

[67] R. L. Siegel *et al.*, “Cancer Statistics , 2016,” vol. 66, no. 1, pp. 7–30, 2016.

[68] K. D. Miller *et al.*, “Cancer treatment and survivorship statistics, 2016,” *CA. Cancer J. Clin.*, vol. 66, no. 4, pp. 271–289, 2016.

[69] L. Galluzzi *et al.*, “Pathophysiology of Cancer Cell Death,” no. January, 2013.

[70] D. J. Mcconkeyt *et al.*, “Apoptosis, cancer and’ cancer therapy,” vol. 6, no. 3, pp. 133–142, 1998.

[71] J. Ferlay *et al.*, “Cancer incidence and mortality worldwide : Sources , methods and major patterns in GLOBOCAN 2012.”

[72] “Breast cancer symptoms | Cancer Research UK,” *Cancer Research, UK.*, 2018. [Online]. Available: http://www.cancerresearchuk.org/about-cancer/breast-cancer/symptoms. [Accessed: 01-May-2018].

[73] C. Harding *et al.*, “Breast cancer screening, incidence, and mortality across US counties,” *JAMA Intern. Med.*, vol. 175, no. 9, pp. 1483–1489, 2015.

[74] Public Health England, “Clinical guidance for breast cancer screening assessment (NHSBSP Publication No 49),” *Gov.Uk*, no. 4, pp. 1–36, 2016.

[75] C. E. Desantis *et al.*, “Breast Cancer Statistics , 2015 : Convergence of Incidence Rates Between Black and White Women,” vol. 66, no. 1, pp. 31–42, 2016.

[76] S. Mj *et al.*, “Early-Onset Familial Breast Cancer,” vol. 250, pp. 17–22.

[77] K. Mcpherson *et al.*, “Breast cancer — epidemiology , risk factors , and genetics Risk factors for breast cancer,” vol. 321, no. September, 2000.

[78] K. N. Anderson *et al.*, “Reproductive risk factors and breast cancer subtypes: A review of the literature,” *Breast Cancer Res. Treat.*, vol. 144, no. 1, pp. 1–10, 2014.

[79] J. Wang *et al.*, “Value of Breast Cancer Molecular Subtypes and Ki67 Expression for the Prediction of Efficacy and Prognosis of Neoadjuvant Chemotherapy in a Chinese Population,” *Medicine (Baltimore).*, vol. 95, no. 18, p. e3518, May 2016.

[80] B. Ardou *et al.*, “Novel Indications for Brca 1 Screening Using Individual Clinical,” vol. 267, no. November 1998, pp. 263–267, 1999.

[81] E. Gabai-Kapara *et al.*, “Population-based screening for breast and ovarian cancer risk due to *BRCA1* and *BRCA2*,” *Proc. Natl. Acad. Sci.*, vol. 111, no. 39, pp. 14205–14210, 2014.

[82] Wooster *et al.*, “Localization of a Breast Cancer Susceptibility Gene , BRCA2 , to Chromosome 13q1 2-13,” *Science*, vol. 265, no. September, pp. 2088–2090, 1994.

[83] S. Anjum *et al.*, “A BRCA1-mutation associated DNA methylation signature in blood cells predicts sporadic breast cancer incidence and survival,” *Genome Med.*, vol. 6, no. 6, p. 47, 2014.

[84] D. J. Hunter *et al.*, “A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cance,” *NIH Public Access*, vol. 39, no. 7, pp. 870–874, 2012.

[85] D. F. Easton *et al.*, “Europe PMC Funders Group Genome-wide association study identifies novel breast cancer susceptibility loci,” vol. 447, no. 7148, pp. 1087–1093, 2009.

[86] M. Garcia-Closas *et al.*, “Heterogeneity of breast cancer associations with five susceptibility loci by clinical and pathological characteristics,” *PLoS Genet.*, vol. 4, no. 4, 2008.

[87] S. N. Stacey *et al.*, “Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer,” *Nat. Genet.*, vol. 39, no. 7, pp. 865–869, 2007.

[88] P. E. A. Huijts *et al.*, “Clinical correlates of low-risk variants in FGFR2, TNRC9, MAP3K1, LSP1 and 8q24 in a Dutch cohort of incident breast cancer cases,” *Breast Cancer Res.*, vol. 9, no. 6, pp. 1–9, 2007.

[89] G. Thomas *et al.*, “A multi-stage genome-wide association in breast cancer identifies two novel risk alleles at 1p11.2 and 14q24.1 (RAD51L1),” *Nature*, vol. 41, no. 5, pp. 579–584, 2010.

[90] W. Wang *et al.*, “Pathway-based discovery of genetic interactions in breast cancer,” *PLoS Genet.*, vol. 13, no. 9, pp. 1–29, 2017.

[91] Q. Milne, R. L., Herranz, J., Michailidou, K., Dennis, J., Tyrer, J. P., Zamora, M. P., ... & Wang, “A large-scale assessment of two-way SNP interactions in breast cancer susceptibility using 46 450 cases and 42 461 controls from the breast cancer association,” *Hum. Mol. Genet.*, vol. 23, no. 7, pp. 1934–1946, 2013.

[92] Y. Sapkota *et al.*, “Assessing SNP-SNP Interactions among DNA Repair, Modification and Metabolism Related Pathway Genes in Breast Cancer Susceptibility,” *PLoS One*, vol. 8, no. 6, pp. 4–9, 2013.

[93] C. A. Anderson *et al.*, “Data quality control in genetic case-control association studies,” *Nat. Protoc.*, vol. 5, no. 9, pp. 1564–1573, Sep. 2010.

[94] N. A. Al-Tassan *et al.*, “A new GWAS and meta-analysis with 1000Genomes imputation identifies novel risk variants for colorectal cancer,” *Sci. Rep.*, vol. 5, no. 1, p. 10442, Sep. 2015.

[95] H. Wang *et al.*, “Trans-ethnic genome-wide association study of colorectal cancer identifies a new susceptibility locus in VTI1A,” *Nat. Commun.*, vol. 5, no. 1, p. 4613, Dec. 2014.

[96] R. A. Eeles *et al.*, “Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array,” *Nat. Genet.*, vol. 45, no. 4, pp. 385–391, Apr. 2013.

[97] A. T. Marees *et al.*, “A tutorial on conducting genome-wide association studies: Quality control and statistical analysis,” *Int. J. Methods Psychiatr. Res.*, vol. 27, no. 2, p. e1608, Jun. 2018.

[98] H. Wang *et al.*, “Novel colon cancer susceptibility variants identified from a genome-wide association study in African Americans,” *Int. J. Cancer*, vol. 140, no. 12, pp. 2728–2733, Jun. 2017.

[99] A. Amin Al Olama *et al.*, “Multiple novel prostate cancer susceptibility signals identified by fine-mapping of known risk loci among Europeans,” *Hum. Mol. Genet.*, vol. 24, no. 19, pp. 5589–5602, Oct. 2015.

[100] M. Kardos *et al.*, “Genomics advances the study of inbreeding depression in the wild,” no. August, pp. 1205–1218, 2016.

[101] S. Turner *et al.*, “Quality Control Procedures for Genome-Wide Association Studies,” *Curr. Protoc. Hum. Genet.*, vol. 68, no. 1, pp. 1.19.1-1.19.18, Jan. 2011.

[102] S. Purcell *et al.*, “PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses,” *Am. J. Hum. Genet.*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[103] A. R. Martin *et al.*, “Human Demographic History Impacts Genetic Risk Prediction across Diverse Populations,” *Am. J. Hum. Genet.*, vol. 100, no. 4, pp. 635–649, Apr. 2017.

[104] G. Baharian *et al.*, “Genetic Ancestry and Natural Selection Drive Population Differences in Immune Responses to Article Genetic Ancestry and Natural Selection Drive Population Differences,” pp. 657–669, 2016.

[105] H. M. Kang *et al.*, “Variance component model to account for sample structure in genome-wide association studies,” *Nat. Genet.*, vol. 42, no. 4, pp. 348–354, Apr. 2010.

[106] J. H. Zhao, “2LD, GENECOUNTING and HAP: computer programs for linkage disequilibrium analysis,” *Bioinformatics*, vol. 20, no. 8, pp. 1325–1326, May 2004.

[107] Y. Li *et al.*, “Genotype Imputation,” *Annu. Rev. Genomics Hum. Genet.*, vol. 10, no. 1, pp. 387–406, Sep. 2009.

[108] J. Marchini *et al.*, “Genotype imputation for genome-wide association studies,” *Nat. Rev. Genet.*, vol. 11, no. 7, pp. 499–511, Jul. 2010.

[109] UK Biobank, “Genotype imputation and genetic association studies of UK Biobank Interim Data Release , May 2015,” no. May, pp. 1–14, 2015.

[110] Our 2 SNPs...®., “To Impute, or not to Impute,” 2015. [Online]. Available: http://blog.goldenhelix.com/goldenadmin/to-impute-or-not-to-impute/. [Accessed: 04-Mar-2017].

[111] Y. Li *et al.*, “MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes,”

Genet. Epidemiol., vol. 34, no. 8, pp. 816–834, Dec. 2010.

[112] J. Marchini *et al.*, “A new multipoint method for genome-wide association studies by imputation of genotypes,” *Nat. Genet.*, vol. 39, no. 7, pp. 906–913, Jul. 2007.

[113] S. R. Browning *et al.*, “Rapid and Accurate Haplotype Phasing and Missing-Data Inference for Whole-Genome Association Studies By Use of Localized Haplotype Clustering,” *Am. J. Hum. Genet.*, vol. 81, no. 5, pp. 1084–1097, Nov. 2007.

[114] F. Brøndum *et al.*, “Comparison of different methods for imputing genome-wide marker genotypes in Swedish and Finnish red Cattle,” pp. 4666–4677, 2013.

[115] P. Zeng *et al.*, “Statistical analysis for genome-wide association study,” *J. Biomed. Res.*, vol. 29, no. November 2014, pp. 285–297, Jul. 2015.

[116] N. H. G. R. I. (NHGRI), “Genome-Wide Association Studies,” 2015. [Online]. Available: https://www.genome.gov/20019523/genomewide-association-studies-fact-sheet/. [Accessed: 15-Sep-2018].

[117] M. D. Fortune *et al.*, “simGWAS: a fast method for simulation of large scale case-control GWAS summary statistics,” *bioRxiv*, p. 313023, 2018.

[118] O. François *et al.*, “Naturalgwas: An R package for evaluating genomewide association methods with empirical data,” *Mol. Ecol. Resour.*, vol. 18, no. 4, pp. 789–797, 2018.

[119] D. Ray *et al.*, “Methods for meta-analysis of multiple traits using GWAS summary statistics,” *Genet. Epidemiol.*, vol. 42, no. 2, pp. 134–145, Mar. 2018.

[120] Y. H. Lee, “Meta-Analysis of Genetic Association Studies,” *Ann. Lab. Med.*, vol. 35, no. 3, p. 283, 2015.

[121] N. Horita *et al.*, “Genetic model selection for a case–control study and a meta-analysis,” *Meta Gene*, vol. 5, pp. 1–8, Sep. 2015.

[122] F. Zhao *et al.*, “Genetic model,” *J. Cell. Mol. Med.*, vol. 20, no. 4, p. 765, 2016.

[123] P. Zeng *et al.*, “Statistical analysis for genome-wide association study,” *J. Biomed. Res.*, vol. 29, no. September 2014, pp. 285–297, 2015.

[124] J. C. Barrett *et al.*, “Genome-wide association defines more than thirty distinct susceptibility loci for Crohn’s disease,” *Nat Genet.*, vol. 40, no. 8, pp. 955–962, 2009.

[125] O. A. Panagiotou *et al.*, “What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations,” *Int. J. Epidemiol.*, vol. 41, no. 1, pp. 273–286, Feb. 2012.

[126] J. Fadista *et al.*, “The (in)famous GWAS P-value threshold revisited and updated for low-frequency variants,” *Eur. J. Hum. Genet.*, vol. 24, no. 8, pp. 1202–1205, 2016.

[127] R. Irizarry, “Lowering the GWAS threshold would save millions of dollars,” *Simply Statistics*, 2017. [Online]. Available: https://simplystatistics.org/2017/06/20/lowering-the-gwas-threshold-would-save-millions-of-dollars/. [Accessed: 25-Sep-2018].

[128] D. Chavalarias *et al.*, “Evolution of reporting P values in the biomedical literature, 1990-2015,” *JAMA - J. Am. Med. Assoc.*, vol. 315, no. 11, pp. 1141–1148, 2016.

[129] S. Greenland *et al.*, “Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations,” *Eur. J. Epidemiol.*, vol. 31, no. 4, pp. 337–350, 2016.

[130] C. Yang *et al.*, “Identifying main effects and epistatic interactions from large-scale SNP data via adaptive group Lasso,” *BMC Bioinformatics*, vol. 11, no. Suppl 1, p. S18, 2010.

[131] X. Wan *et al.*, “BOOST: A Fast Approach to Detecting Gene-Gene Interactions in Genome-wide Case-Control Studies,” *Am. J. Hum. Genet.*, vol. 87, no. 3, pp. 325–340, Sep. 2010.

[132] D. Gilbert-Diamond *et al.*, “Analysis of Gene-Gene Interactions,” in *Current Protocols in Human Genetics*, vol. 6, no. 9, Hoboken, NJ, USA: John Wiley & Sons, Inc., 2011, pp. 2166–2171.

[133] M. L. Calle *et al.*, “mbmdr: an R package for exploring gene–gene interactions associated with binary or quantitative traits,” *Bioinformatics*, vol. 26, no. 17, pp. 2198–2199, Sep. 2010.

[134] J. Gui *et al.*, “A Robust Multifactor Dimensionality Reduction Method for Detecting Gene-Gene Interactions

with Application to the Genetic Analysis of Bladder Cancer Susceptibility," *Ann. Hum. Genet.*, vol. 75, no. 1, pp. 20–28, Jan. 2011.

[135] J. H. Moore *et al.*, "A flexible computational framework for detecting, characterizing, and interpreting statistical patterns of epistasis in genetic studies of human disease susceptibility," *J. Theor. Biol.*, vol. 241, no. 2, pp. 252–261, Jul. 2006.

[136] J. Namkung *et al.*, "Identification of gene-gene interactions in the presence of missing data using the multifactor dimensionality reduction method," *Genet. Epidemiol.*, vol. 33, no. 7, pp. 646–656, Nov. 2009.

[137] T. Ban *et al.*, "A study on association rule mining of darknet big data," in *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, pp. 1–7.

[138] Z. Yu *et al.*, "Automated detection of unusual soil moisture probe response patterns with association rule learning," *Environ. Model. Softw.*, vol. 105, pp. 257–269, 2018.

[139] Q. Zhang *et al.*, "AprioriGWAS, a New Pattern Mining Strategy for Detecting Genetic Variants Associated with Disease through Interaction Effects," *PLoS Comput. Biol.*, vol. 10, no. 6, p. e1003627, Jun. 2014.

[140] A. Terada *et al.*, "LAMPLINK: detection of statistically significant SNP combinations from GWAS data," *Bioinformatics*, vol. 32, no. July, p. btw418, Jul. 2016.

[141] P. Fournier-Viger *et al.*, "A survey of itemset mining," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 7, no. 4, p. e1207, Jul. 2017.

[142] M. G. G'Sell *et al.*, "Sequential selection procedures and false discovery rate control," *J. R. Stat. Soc. Ser. B Stat. Methodol.*, vol. 78, no. 2, pp. 423–444, 2016.

[143] E. L. Lehmann *et al.*, "Generalizations of the Familywise Error Rate," in *Selected Works of E. L. Lehmann*, J. Rojo, Ed. Boston, MA: Springer US, 2012, pp. 719–735.

[144] A. Miceli *et al.*, "Teoria Statistica Delle Classi e Calcolo Delle Probabilità," *Encycl. Res. Des.*, pp. 1493–1494, 2018.

[145] B. Efron, "Size, power and false discovery rates," *Ann. Stat.*, vol. 35, no. 4, pp. 1351–1377, Aug. 2007.

[146] S. R. Narum, "Beyond Bonferroni: Less conservative analyses for conservation genetics," *Conserv. Genet.*, vol. 7, no. 5, pp. 783–787, Sep. 2006.

[147] P. H. Westfall *et al.*, "Multiple Tests for Genetic Effects in Association Studies," in *Biostatistical Methods*, New Jersey: Humana Press, pp. 143–168.

[148] J. Hind *et al.*, "A robust method for the interpretation of genomic data," in *Proceedings of the International Joint Conference on Neural Networks*, 2017, vol. 2017-May.

[149] L. Excoffier *et al.*, "Robust Demographic Inference from Genomic and SNP Data," *PLoS Genet.*, vol. 9, no. 10, p. e1003905, Oct. 2013.

[150] R. Nielsen *et al.*, "SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation from New-Generation Sequencing Data," *PLoS One*, vol. 7, no. 7, p. e37558, Jul. 2012.

[151] E. A. Stahl *et al.*, "Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis," *Nat. Genet.*, vol. 44, no. 5, pp. 483–489, May 2012.

[152] C. Provost *et al.*, "Perinatal Clinical Decision Support System: A Documentation Tool for Patient Safety," *Nurs. Womens. Health*, vol. 11, no. 4, pp. 407–410, 2007.

[153] B. Malmir *et al.*, "A medical decision support system for disease diagnosis under uncertainty," *Expert Syst. Appl.*, vol. 88, pp. 95–108, 2017.

[154] E. Alickovic *et al.*, "Medical Decision Support System for Diagnosis of Heart Arrhythmia using DWT and Random Forests Classifier," *J. Med. Syst.*, vol. 40, no. 4, pp. 1–12, 2016.

[155] L. M. Connelly, "Fisher's Exact Test," vol. 25, no. 1, p. 2016, 2016.

[156] E. W. Corty, *Using and Interpreting Statistics*, 3rd ed. Worth, 2016.

[157] R. M. Cantor *et al.*, "Prioritizing GWAS Results: A Review of Statistical Methods and Recommendations for Their Application," *Am. J. Hum. Genet.*, vol. 86, no. 1, pp. 6–22, 2010.

[158] R. S. Society, “Contingency Tables Involving Small Numbers and the χ2 Test Author (s): F . Yates Source : Supplement to the Journal of the Royal Statistical Society , Vol . 1 , No . 2 (1934), pp . Published by : Wiley for the Royal Statistical Society,” vol. 1, no. 2, pp. 217–235, 2014.

[159] G. Zheng *et al.*, “On estimation of the variance in Cochran-Armitage trend tests for genetic association using case-control studies,” *Stat. Med.*, vol. 25, no. 18, pp. 3150–3159, 2006.

[160] S. Wellek *et al.*, “Cochran-Armitage Test versus Logistic Regression in the Analysis of Genetic Association Studies,” *Hum. Hered.*, vol. 73, no. 1, pp. 14–17, 2012.

[161] P. Sedgwick *et al.*, “Odds ratios,” *BMJ*, vol. 341, no. aug18 1, pp. c4414–c4414, Aug. 2010.

[162] J. Zaragoza Cortes *et al.*, “Odds ratio between sociocultural factors, body dissatisfaction, and body mass index in university students of Hidalgo, Mexico,” *Arch. Latinoam. Nutr.*, vol. 61, no. 1, pp. 20–7, 2011.

[163] C. O. Schmidt *et al.*, “When to use the odds ratio or the relative risk?,” *Int. J. Public Health*, vol. 53, no. 3, pp. 165–167, Jun. 2008.

[164] J. Zhang *et al.*, “What’s the relative risk?: A method of correcting the odds ratio in cohort studies of common outcomes,” *Jama*, vol. 280, no. 19, pp. 1690–1691, 1998.

[165] C. A. C. Montanez *et al.*, “Machine learning approaches for the prediction of obesity using publicly available genetic profiles,” in *Proceedings of the International Joint Conference on Neural Networks*, 2017, vol. 2017-May.

[166] A. Mayr *et al.*, “The Evolution of Boosting Algorithms,” *Methods Inf. Med.*, vol. 53, no. 06, pp. 419–427, Jan. 2014.

[167] I. Salian, “SuperVize Me: What’s the Difference Between Supervised, Unsupervised, Semi-Supervised and Reinforcement Learning?,” *NVIDIA*, Aug-2018.

[168] D. Kirsch *et al.*, “Approaches to Machine Learning,” in *Machine Learning for Dummies*, John Wiley & Sons, Inc., 2018, pp. 14–16.

[169] S. Shalev-Shwartz *et al.*, *Understanding machine learning: From theory to algorithms*, vol. 9781107057. 2013.

[170] M. W. Libbrecht *et al.*, “Machine learning applications in genetics and genomics,” *Nat. Rev. Genet.*, vol. 16, no. 6, pp. 321–332, Jun. 2015.

[171] L. Breiman, “Random Forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[172] M. Milanović *et al.*, “CHAID Decision Tree: Methodological Frame and Application,” *Econ. Themes*, vol. 54, no. 4, pp. 563–586, 2016.

[173] D. Steinberg, *CART: classification and regression trees*. 2009.

[174] J. Maria *et al.*, “Stacked Autoencoders Using Low-Power Accelerated Architectures for Object Recognition in Autonomous Systems,” *Neural Process. Lett.*, vol. 43, no. 2, pp. 445–458, 2016.

[175] P. Fergus *et al.*, “Utilising Deep Learning and Genome Wide Association Studies for Epistatic-Driven Preterm Birth Classification in African-American Women,” *IEEE/ACM Trans. Comput. Biol. Bioinforma.*, pp. 1–1, 2018.

[176] S. Suthaharan, *Support vector machine*, Machine le. Boston, MA: Springer, 2016.

[177] Breogan, “Breast Oncology Galicia Network (BREOGAN),” 2018. [Online]. Available: http://proyectobreogan.es/. [Accessed: 03-Jun-2018].

[178] Cancer.org, “Cancer Prevention Study II (CPS II) l American Cancer Society,” 2018. [Online]. Available: https://www.cancer.org/research/we-conduct-cancer-research/epidemiology/cancer-prevention-study-2.html. [Accessed: 15-Sep-2018].

[179] M. B. Mortensen *et al.*, “The high-density lipoprotein-adjusted SCORE model worsens SCORE-based risk classification in a contemporary population of 30 824 Europeans: The Copenhagen General Population Study,” *Eur. Heart J.*, vol. 36, no. 36, pp. 2446–2453, 2015.

[180] R. L. Milne *et al.*, “Cohort Profile: The Melbourne Collaborative Cohort Study (Health 2020),” *Int. J. Epidemiol.*, vol. 46, no. 6, pp. 1757-1757i, Dec. 2017.

[181] Uhcancercenter.org, “The Multiethnic Cohort Study.” [Online]. Available: http://www.uhcancercenter.org/research/the-multiethnic-cohort-study-mec. [Accessed: 20-Sep-2018].

[182] M.-R. Han *et al.*, “Evaluating 17 breast cancer susceptibility loci in the Nashville breast health study,” *Breast Cancer*, vol. 22, no. 5, pp. 544–551, Sep. 2015.

[183] Nurseshealthstudy.org, “History | Nurses’ Health Study.” [Online]. Available: http://www.nurseshealthstudy.org/about-nhs/history. [Accessed: 20-Sep-2018].

[184] M. M. Gaudet *et al.*, “Genetic variation in SIPA1 in relation to breast cancer risk and survival after breast cancer diagnosis,” *Int. J. Cancer*, vol. 124, no. 7, pp. 1716–1720, Apr. 2009.

[185] C. A. McCarty *et al.*, “Alcohol, genetics and risk of breast cancer in the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Trial,” *Breast Cancer Res. Treat.*, vol. 133, no. 2, pp. 785–792, Jun. 2012.

[186] G. C. Barnett *et al.*, “Risk factors for the incidence of breast cancer: Do they affect survival from the disease?,” *J. Clin. Oncol.*, vol. 26, no. 20, pp. 3310–3316, 2008.

[187] R. Suzuki *et al.*, “Body weight and postmenopausal breast cancer risk defined by estrogen and progesterone receptor status among Swedish women: A prospective cohort study,” *Int. J. Cancer*, vol. 119, no. 7, pp. 1683–1689, Oct. 2006.

[188] E. Riboli *et al.*, “European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection,” *Public Health Nutr.*, vol. 5, no. 6b, pp. 1113–1124, Dec. 2002.

[189] D. P. Sandler *et al.*, “The Sister Study Cohort: Baseline Methods and Participant Characteristics,” *Environ. Health Perspect.*, vol. 125, no. 12, p. 127003, Dec. 2017.

[190] R. T. Chlebowski *et al.*, “Estrogen Plus Progestin and Breast Cancer Incidence and Mortality in the Women’s Health Initiative Observational Study,” *JNCI J. Natl. Cancer Inst.*, vol. 105, no. 8, pp. 526–535, Apr. 2013.

[191] WHO, “International Statistical Classification of Diseases and Related Health Problems,” *Dermatology*, 2015.

[192] Cancervic.org.au, “Melbourne Collaborative Cohort Study - Cancer Council Victoria,” 2018. .

[193] C. I. Amos *et al.*, “The oncoarray consortium: A network for understanding the genetic architecture of common cancers,” *Cancer Epidemiol. Biomarkers Prev.*, vol. 26, no. 1, pp. 126–135, 2017.

[194] A. Z. Dayem Ullah *et al.*, “A practical guide for the functional annotation of genetic variations using SNPnexus,” *Brief. Bioinform.*, vol. 14, no. 4, pp. 437–447, Jul. 2013.

[195] P. A. Thompson *et al.*, “Selective Genomic Copy Number Imbalances and Probability of Recurrence in Early-Stage Breast Cancer,” *PLoS One*, vol. 6, no. 8, p. e23543, Aug. 2011.

[196] R. J. DeBerardinis, “Serine Metabolism: Some Tumors Take the Road Less Traveled,” *Cell Metab.*, vol. 14, no. 3, pp. 285–286, Sep. 2011.

[197] J. Bennewitz *et al.*, “Improved confidence intervals in quantitative trait loci mapping by permutation bootstrapping,” *Genetics*, vol. 160, no. 4, pp. 1673–1686, 2002.

[198] Wikimedia Commons, “File:DNA simple2.svg,” *Commons.wikimedia.org.*, 2008. .

[199] Flickr., “DNA vs RNA.,” 2018. .

[200] En.wikipedia.org., “Protein CTNND1 PDB 3L6X,” 2011. .